PROCEEDINGS
International Conference on Education
Faculty of Tarbiyah | UIN Syekh Wasil Kediri
E-ISSN: 3025-6828

Volume 4, 2026

Detecting Culture-Religious Hallucination of AI-Generated Feedback on EFL Students' Argumentative Essay

Sari Dewi Noviyanti[1*]; Awwalia Fitratin Izza[1]
[1]Universitas Islam Negeri (UIN) Walisongo Semarang
Correspondence e-mail: sari.dewi@walisongo.ac.id

Abstract:
The implications of using large language models (LLMs) in educational technology around generative feedback have positive and negative implications particularly when its content related to religion and culture. This study examines the patterns of hallucination in AI feedback towards Islamic content in EFL student essays in Islamic learning context. A systematic evaluation was conducted of five leading LLMs (ChatGPT, Claude, Grok, DeepSeek, and Gemini) using a newly created hallucination multi-type classification framework. The LLMs were prompt to provide feedback on EFL student essays containing Islamic content; then an evaluation was conducted to the responses using 6 types of hallucination indicators namely source fabrication, attribution errors, content misrepresentation, citation errors, contextual misapplication, and theological misstatement. This study found clear patterns of hallucination across all models, which ChatGPT generated the most errors and Claude Sonnet 4 produces the least errors in his feedback. While LLMs responded to the prompts with hallucination in some forms, the patterns and implications of hallucination were distinctly different across all examples. This information is valuable for educators and developers to consider regarding the legitimacy of AI feedback in religious contexts.

Keywords: AI Hallucination; EFL Essay Assessment; Generative Feedback; Large Language Models; Islamic Content

Introduction

One of the characteristics of Islamic education in Indonesia is exposing students to various landscape of references that include the Quran, Hadith, case law, theology, and modern academic publications. It forces both teachers and students to adapt and contextualize the Islamic content to teaching and learning process (Ghufron et al, 2024; Syuhda et al., 2024). Representing these contents not only must be accurate, but also include the understanding of complex intertextuality, historical contexts, and interpretive modes (Abbas et al, 2023; Addzaky et al, 2025). This factor then makes teaching English as a foreign language in Islamic Higher Education become more challenging. While in public education institution EFL students are being exposed through to multilingual and multicultural contexts, both to native and non-native content (Deiniatur & Hasanah, 2024; Kimsesiz, 2024), In Islamic education, religious content is highly explored. It makes the teacher approach in teaching, should provide not only linguistic structures but also cultural and religious contexts (Fathia Baresh, 2024; Ilyas & Afzal, 2021).

One of the significant issues facing the application of LLMs in educational contexts is the phenomenon of AI hallucination (defined as the generation of plausible but always inaccurate, unsupported, or fabricated information) (Sun, Sheng, Zhou, & Wu, 2024; Zhang, 2025). The idea of hallucination has been extensively explored in literature addressing general AI contexts, a study showed that chatbots can hallucinate and contain factual errors of texts produced (Bobula, 2024). This led to an idea that information verification then becomes a significant way to keep the accuracy (Lee et al, 2025). However, the idea of hallucination in an educational context, such as a religiously sensitive feedback model, is still largely underexplored. This is particularly significant in light of the unique epistemological constructions, citation conventions and values, and scholarly traditions associated with Islamic academic content (Achruh et al, 2024). When AI systems provide feedback about essays with Islamic content, they should engage with this complicated work of scholarship and at the same time respect academic rigor and cultural consideration, but recent literature has established that LLMs have a hard time even with highly contextual terms from specialized sub-disciplines of Islamic law.

Existing research on AI hallucination has primarily focused on the nuances of general knowledge, while the challenge with religious and cultural content remains less explored (Barnett et al, 2024; Berglund et al., 2024). Research studies that assess the capabilities of LLMs typically focus on factual accuracy in non-sectarian academic contexts, and therefore do not typically engage with the particular needs and nuances of religious scholarship. There are better, secular approaches developing for the study of misinformation and hallucinations as it relates to syntactical, artificial and contextual nature of learning in LLMs, and there is growing recognition that hallucinations occur in AI systems due to quality and the scope of training data, which is particularly problematic in community college contexts where disciplinary content can be so narrowly focused in terms of its specialization and contextualization (Casal-Otero et al., 2023). This is especially problematic in view of the growing presence of AI tools across a variety of educational contexts and cultures around the world, where generating reliable representations of culturally and religiously sensitive content into learning materials which were or are not knowingly disrespectful representations of 'truth' or of people who hold this knowledge as sacred, is important for educational integrity and cultural respect (Mozelius et al, 2024).

To respond to this need, this study conducted a comprehensive comparative analysis of AI hallucination patterns across 5 major LLM platforms (ChatGPT, Claude,

Grok, DeepSeek, and Gemini) generating feedback of Islamic content in EFL conducted in Islamic learning contexts at Faculty of Ushuluddin and Humanities, State Islamic University of Walisongo Semarang. The study implemented a classification framework specifically for Islamic content hallucination and applies the framework to assess five of the leading AI systems in education that educators can find in everyday use. The results add to the theoretical understanding of hallucination, as well as identify the important factors for educators to consider when using AI for assessments in cultural contexts involving Islamic education.

Methods

Hallucination patterns occurring in five major Large Language Model platforms namely Claude, Grok, Deep Seek, Gemini and ChatGPT. The data collection was conducted through content analysis and while data analysis was conducted through three qualitative levels of analysis namely (1) quantitative level of hallucination frequency and distribution across platforms, (2) qualitative level of hallucination error types and associated severity and relevance, and (3) comparative level of platform credibility with respect to Islamic content. This approach was necessary because of the multi-faceted nature of Islamic scholarly traditions, and the different manifestations of hallucination within religious content. The study's design incorporated established frameworks of AI evaluation research as well as the unique requirements for religious content evaluation. The Islamic content evaluation is particularly novel as there are not existing standardized protocols for AI hallucination evaluation. Therefore, this research also developed new evaluation protocols tailored specifically to Islamic scholarly appropriateness and cultural sensitivity.

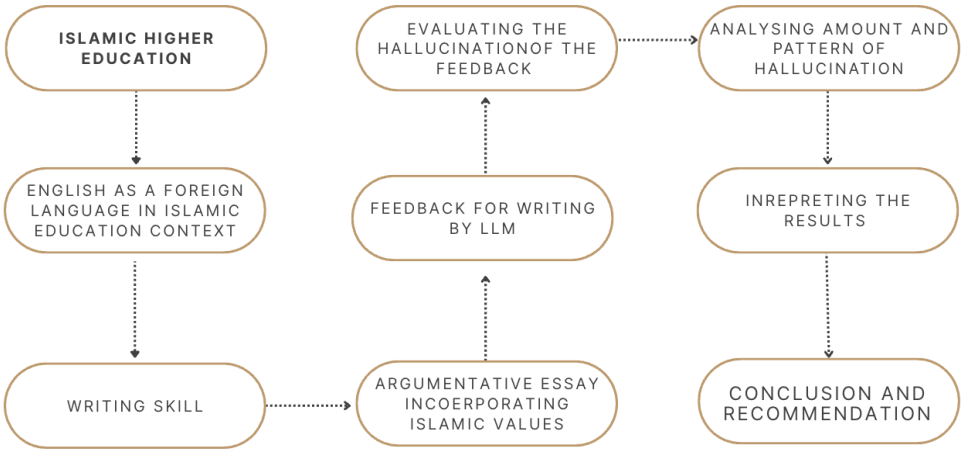


Figure 1. Conceptual framework of the study

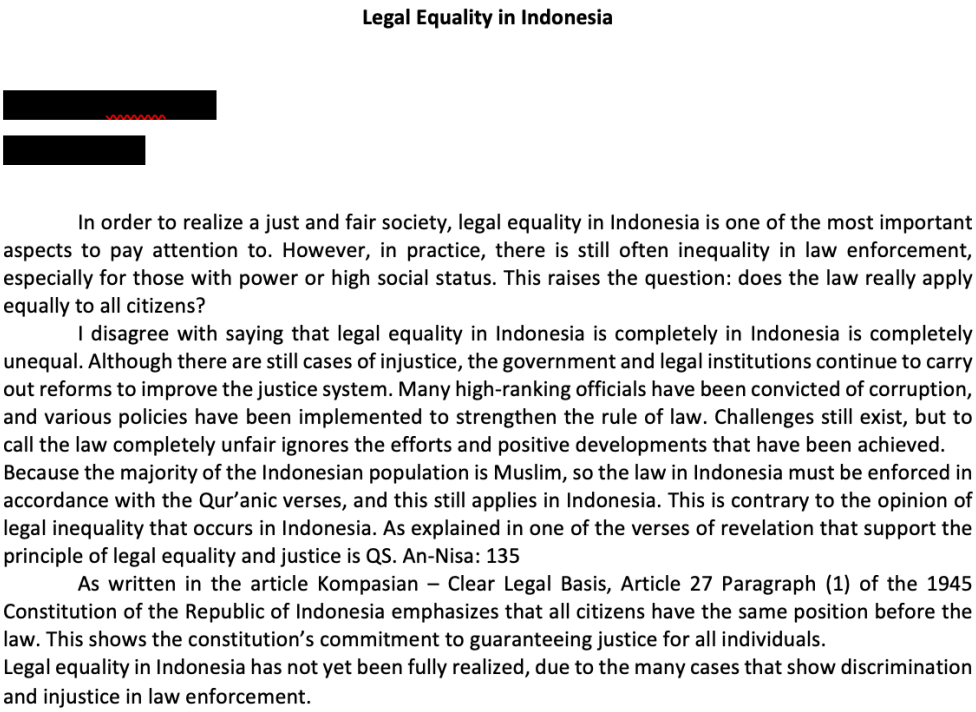
The initial stage of this study is students were assigned to compose an argumentative essay consist of Islamic content. This stage was conducted at English course held Faculty if Ushuluddin and Humanities, State Islamic University of Semarang. A student's essay was randomly chosen to be evaluated by the LLMs. The essay then evaluated by those five LLM models by implementing a prompt. Each prompt was pilot-tested with each platform to ensure they would produce substantive answers that could be analyzed for hallucination. Then the feedback was analyzed by three experts of Qur'an and Tafseer lecturers from three Islamic Universities in Indonesia whose all of them have minimum ten years teaching experiences and hold doctoral degree. The through six error identifications namely: source fabrication, attribution errors, content misrepresentation,

citation errors, contextual misapplication and theological misstatement. A number of validation methods were used to increase reliability and validity of identification and categorization of hallucinations in the essays. A group of subject-matter experts in Islamic studies, language structure, and educational technology provided confirming evidence of the hallucinations and groupings, and also considered ramifications of the groups of hallucinations in the context of educational usage.

Results and Discussion

Findings

The initial stage of this study is students were assigned to compose an argumentative essay consist of Islamic content. This stage was conducted at English course held Faculty if Ushuluddin and Humanities, State Islamic University of Semarang. A student's essay was randomly chosen to be evaluated by the LLMs. The essay then evaluated by those five LLM models by implementing a prompt. Each prompt was pilot-tested with each platform to ensure they would produce substantive answers that could be analyzed for hallucination. Then the initial stage of this study is 25 students at Qur'an and Tafseer Department at Faculty of Ushuluddin and Humanities State Islamic University of Semarang were assigned to compose an argumentative essay. The students were asked to compose an essay by adapting any issue based on their interest. The essay should be written by providing both secular and Islamic perspectives as their arguments. The Islamic perspectives should be cited by Qur'an, hadith and published book and or article journals only. The essay chosen is entitled "Legal Equity in Indonesia".



Legal Equality in Indonesia

[redacted]

In order to realize a just and fair society, legal equality in Indonesia is one of the most important aspects to pay attention to. However, in practice, there is still often inequality in law enforcement, especially for those with power or high social status. This raises the question: does the law really apply equally to all citizens?

I disagree with saying that legal equality in Indonesia is completely in Indonesia is completely unequal. Although there are still cases of injustice, the government and legal institutions continue to carry out reforms to improve the justice system. Many high-ranking officials have been convicted of corruption, and various policies have been implemented to strengthen the rule of law. Challenges still exist, but to call the law completely unfair ignores the efforts and positive developments that have been achieved.

Because the majority of the Indonesian population is Muslim, so the law in Indonesia must be enforced in accordance with the Qur'anic verses, and this still applies in Indonesia. This is contrary to the opinion of legal inequality that occurs in Indonesia. As explained in one of the verses of revelation that support the principle of legal equality and justice is QS. An-Nisa: 135

As written in the article Kompasian – Clear Legal Basis, Article 27 Paragraph (1) of the 1945 Constitution of the Republic of Indonesia emphasizes that all citizens have the same position before the law. This shows the constitution's commitment to guaranteeing justice for all individuals.

Legal equality in Indonesia has not yet been fully realized, due to the many cases that show discrimination and injustice in law enforcement.

Figure 2. Student's essay

The next phase was generating feedback by evaluating the essay through five LLM models. This evaluation conducted by implementing same essay with same prompt. After all feedback were generated, the feedback then analyzed through the hallucination classification schema for Islamic content. The development of a comprehensive hallucination classification schema was a conducted by adapting several previous schemas of analysis of Islamic deception in various media (Ahmed Mohamed, Madihah Sheikh Abdul Aziz, & Murni Mahmud, 2023; Mahmud & Abubakar, 2015). However sine

the existing classifications were unable to sufficiently capture the types of errors occurring in the generation of Islamic content of AI hallucination, the new schema is emerged to be deployed. After a comprehensive initial analysis and discussions with a number of scholars in Islamic studies, this study finally arrived at a six-category classification model to include the full range of possible hallucination types in Islamic content.

The identification errors were conducted through six hallucinations namely: type 1 is source fabrication. This error involves the complete creation of an entirely fictitious source, for example a fictitious book, article, author, or scholarly work. This is the most serious hallucination given it includes the complete creation of fictitious academic material that could mislead students about the existence of Islamic scholarship sources.

The type 2 is attribution errors. It involves placing real quotes, ideas or scholarly perspectives in the palm of the wrong authors or identifying authentic work with the wrong source. This can be particularly serious in Islamic contexts where accurate attribution is arguably critical to understanding the diversity of Islamic thought.

Type 3 is content misrepresentation. It includes inaccurate descriptions of real source content, and misrepresenting arguments contained in real sources. This category covers instances when there are legitimate sources but the AI system misrepresents the content, arguments and/or position of the source.

Type 4 is citation errors considers incorrect bibliographic detail for real sources including wrong publication date, publisher, page number, or references related details. Although it will happen less than the complete fabrication of content, and less serious, citation errors can hinder students' ability to locate or verify sources. Type 5 is contextual misapplication. These errors presented real and legitimate sources, but they have been deployed in an inappropriate context. It refers to real sources being used in inappropriate contexts or sources being cited to support arguments they do not actually support.

Type 5 is especially pertinent in Islamic contexts, wherein sources may be cited correctly but in contexts that are theologically or legally inappropriate. Lastly, the type 6 is theological misstatement (misrepresentation/interpretation of an Islamic perspective). It refers to inaccurate statements, representations or interpretations of Islamic beliefs, teachings or scholarly positions. This represents a distinct challenge in terms of ensuring theological accuracy within AI-generated writings on Islam.

Evaluation of Hallucination Trends

The systematic duration of evaluation of the AI-produced feedback across the five predominant LLM platforms effectively identified complex and multi-faceted hallucination trends that varied by frequency, type, and severity. The broad evaluation of 12 unique hallucination examples from 5 platforms have been a rare glimpse at the reliability issues associated with the evaluation of AI-assisted Islamic content in educational environments.

Table 1. Hallucination Frequency on Feedback

Model	Errors	Percentage
Claude	1	8.3%
Grok	2	16.7%
Deep Seek	2	16.7%
Gemini	3	25%
ChatGPT	4	33.3%

Source: Authors' analysis

Distributions of hallucinations presented in Table 1 showed uneven across platforms, as totals of errors ranged from 1-4 errors per platform, indicating that not only the architectures, but methods used in training and underling data will all impact the eventual performance of LLMs in producing Islamic content. The overall hallucination count of 2.4 errors per platform demonstrates that while we will see hallucinations exist in all systems, these errors will not be uniformly distributed, and it may be possible to locate and utilize platforms that are more accurate in educational contexts.

Based on the table 1, the errors distribution yields no perfect compliance across platforms, but they variations in perceived performance indicate systematic considerations in how these models produce and construct content related to Islamic scholarship. The three-fold difference between the best (Claude) and worst (ChatGPT) performer indicates a significant potential for improvement and careful platform selection in educational contexts.

Table 2. Hallucination Types on Feedback

Model	Fab.	Attr.	Content	Citation	Context	Theo.	Total
Claude	–	–	–	–	1	–	1
Grok	–	–	–	2	–	–	2
Deep Seek	–	–	–	–	2	–	2
Gemini	–	–	–	1	2	–	3
ChatGPT	–	–	–	4	–	–	4

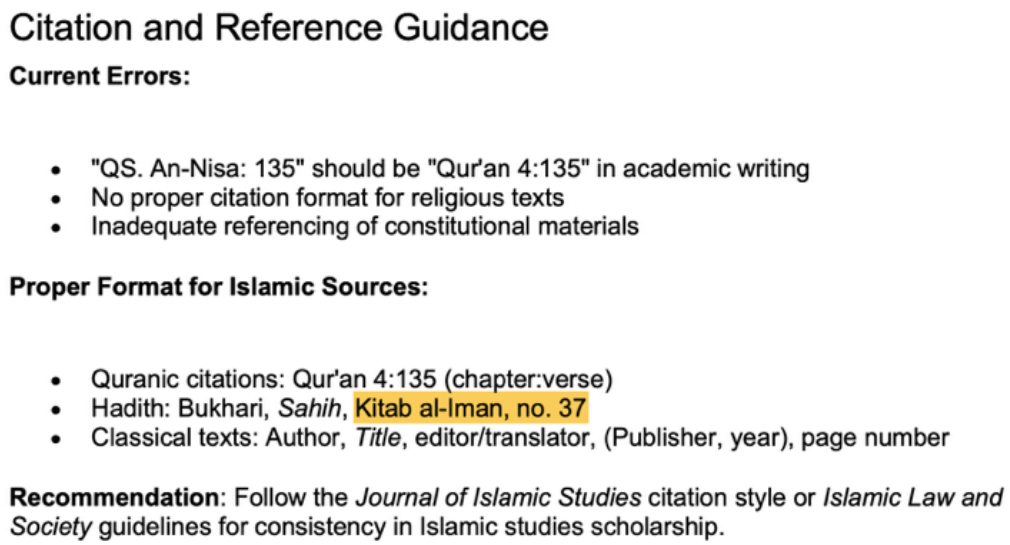
Source: Authors' analysis

Note. Fab. = source fabrication; Attr. = attribution errors; Content = content misrepresentation; Citation = citation errors; Context = contextual misapplication; Theo. = theological misstatement.

The distributions of hallucination categories exhibit meaningful patterns that demonstrate the specific difficulties experienced by AI systems when producing feedback on Islamic content. Citation Error (Type 4) was the most prevalent hallucination, accounting for 7 of the 12 (58.3%) occurrences. This prevalence indicates that while AI systems reasonably handle Islamic concepts and sources, they substantially fall short in bibliographic representation. Contextual Misapplication (Type 5) was the second highest frequency with 4 occurrences (33.3%), which demonstrates serious struggle when attempting the connect Islamic sources with contextually appropriate content. Content Misrepresentation (Type 3) had one occurrence (8.3%), while Source Fabrication (Type 1), Attribution Errors (Type 2), and Theological Misstatements (Type 6) were entirely absent.

Platform-Specific Performance

Claude showed excellent capabilities, it has been docked as the most reliable for generative feedback on Islamic content, with only one occurrence of contextual misapplication due partially to the citation of a Hadith.



Citation and Reference Guidance

Current Errors:

- "QS. An-Nisa: 135" should be "Qur'an 4:135" in academic writing
- No proper citation format for religious texts
- Inadequate referencing of constitutional materials

Proper Format for Islamic Sources:

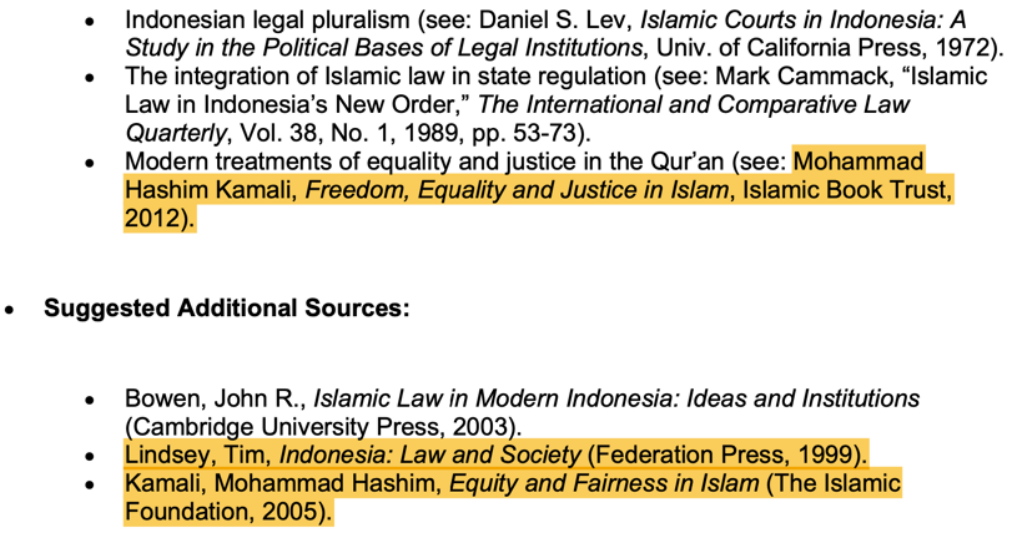
- Quranic citations: Qur'an 4:135 (chapter:verse)
- Hadith: Bukhari, *Sahih*, Kitab al-Iman, no. 37
- Classical texts: Author, *Title*, editor/translator, (Publisher, year), page number

Recommendation: Follow the *Journal of Islamic Studies* citation style or *Islamic Law and Society* guidelines for consistency in Islamic studies scholarship.

Figure 3. Example of errors found in Claude's feedback

The particular contextual error, situating the Hadith from Bukhari's Sahih (Kitab al-Iman, no. 37) which would be about fasting when the Hadith referred to was discussing faith and justice. Although this is a serious error that can misled students on what is Islam teaches, at least the lack of citation fabrication or representations evident from Islamic works suggests good prior knowledge of Islamic texts. Claude's performance is likely a product of Anthropic's constitutional AI training strategy, which emphasizes helpfulness and accuracy. In addition, to their apparent hesitation in making claims with Islamic content may explain why the number of errors is lower than ChatGPT, further research is needed to see if the reduction of hallucinations from reductions in their apparent risk or if it merely cost comprehensiveness or the level of detail in generating feedback.

For ChatGPT, all four identified hallucinations were in even, systematic citation errors, which were spread across the multiple Islamic scholarly sources. All four identified hallucinations were citations errors.



- Indonesian legal pluralism (see: Daniel S. Lev, *Islamic Courts in Indonesia: A Study in the Political Bases of Legal Institutions*, Univ. of California Press, 1972).
- The integration of Islamic law in state regulation (see: Mark Cammack, "Islamic Law in Indonesia's New Order," *The International and Comparative Law Quarterly*, Vol. 38, No. 1, 1989, pp. 53-73).
- Modern treatments of equality and justice in the Qur'an (see: Mohammad Hashim Kamali, *Freedom, Equality and Justice in Islam*, Islamic Book Trust, 2012).

- **Suggested Additional Sources:**

- Bowen, John R., *Islamic Law in Modern Indonesia: Ideas and Institutions* (Cambridge University Press, 2003).
- Lindsey, Tim, *Indonesia: Law and Society* (Federation Press, 1999).
- Kamali, Mohammad Hashim, *Equity and Fairness in Islam* (The Islamic Foundation, 2005).

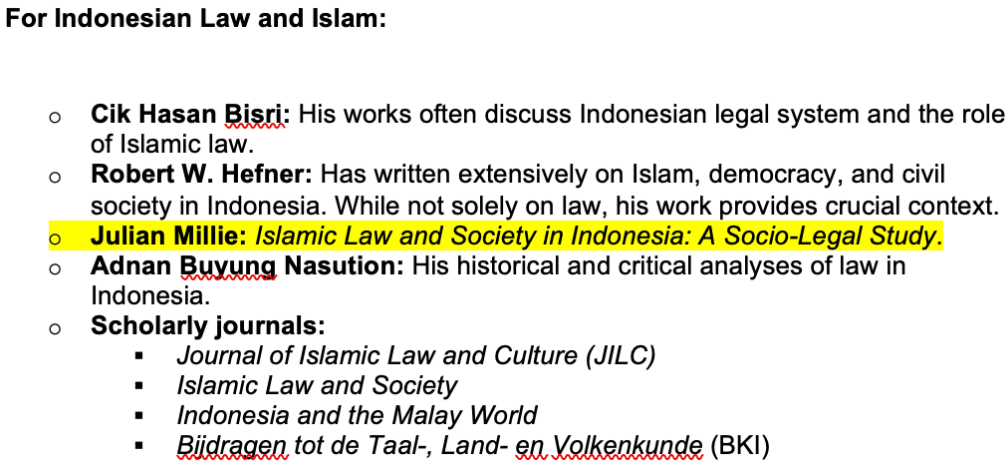
Figure 4. Example of errors found in ChatGPT's feedback

The citation errors had identified patterns: wrong publication dates (for Mohammad Hashim Kamali's work where ChatGPT identified the publication as 2012 instead of 2002), wrong publisher information (where Islamic Book Trust was identified by ChatGPT, when it should have been Islamic Text Society), and failures of

comprehensive bibliographic level accuracy, in some instances effecting multiple citation details at the same time. The error with Tim Lindsey's work (where ChatGPT provided 1999 instead of 2008 for the publication date), along with errors with the mis-identification of Simon Butt's work within their publication information from, Corruption and Law in Indonesia, 2011 instead of 2017 also knowns more systemic issues of maintaining accurate bibliographic repositories relating specifically to scholarship on Islamic and Indonesia scholarship.

Despite the citation inaccuracy challenges faced by ChatGPT, its use of sources was strong enough to indicate that issues with content fabrication and misrepresentation of the theological fact that, at the very least, its knowledge base contains accurate information about Islamic concepts and scholarly positions, even when the bibliographic references were incorrect.

Gemini's hallucination pattern was the most diverse of all reviewed systems as there were examples of both citation mistakes and contextual mistakes. The only citation error involved Julian Millie's invented journal article "Islamic Law and Society in Indonesia: A Socio Legal Study", which was incorrectly attributed to the journal "Islamic Law and Society" when there was no such publication by Millie.


For Indonesian Law and Islam:

- **Cik Hasan Bisri:** His works often discuss Indonesian legal system and the role of Islamic law.
- **Robert W. Hefner:** Has written extensively on Islam, democracy, and civil society in Indonesia. While not solely on law, his work provides crucial context.
- **Julian Millie:** *Islamic Law and Society in Indonesia: A Socio-Legal Study*.
- **Adnan Buyung Nasution:** His historical and critical analyses of law in Indonesia.
- **Scholarly journals:**
  - *Journal of Islamic Law and Culture (JILC)*
  - *Islamic Law and Society*
  - *Indonesia and the Malay World*
  - *Bijdragen tot de Taal-, Land- en Volkenkunde* (BKI)

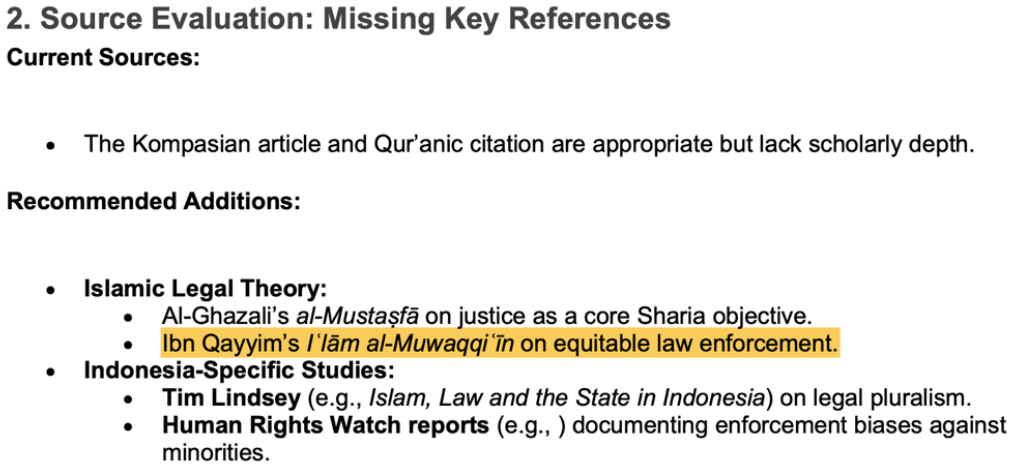
Figure 5. Example of errors found in Gemini's feedback

The two contextual mistakes represented Gemini's struggles with matching a source to the correct content area. In the first instance, the use of "Contributions to Linguistics, Land and Ethnology (BKI)", for a discussion about Indonesian law and Islam, is concerning, as the journal specifically states that it publishes research on linguistics, anthropology and Southeast Asian history. In the second instance, misusing "Indonesia and the Malay World" journal (the journal states it publishes articles on language, literature, art and culture and does not consider articles relating to legal studies), indicates serious and consistent challenges with source-context matching. This pattern of errors indicates that even though Gemini has some knowledge of relevant academic sources, it struggles to accurately understand the scope of those sources and their areas of focus, leading to inaccurate recommendations that could misinform students regarding their potential relevance and appropriateness.

The errors DeepSeek displayed consisted of a blend of content misrepresentation and context misapplication; one explanation would be that it struggled with both understanding the source content, and both content and context in using the source appropriately. In the same reference work 'Iʿlām al-Muwaqqiʿīn' by Ibn Qayyim, DeepSeek provided an extensive description of 'extensive commentary on the Qur'an

including interpretation, grammar, and linguistic nuances', when what was really presented had to do with Islamic jurisprudence and legal methodology.

The context misapplication also had to do with the reference work of Ibn Qayyim, as DeepSeek used information discussing adaptation of fatwa to different eras and contexts in order to support arguments made about equal legal in Indonesian context. This error not only demonstrates misunderstanding of the content of the source, but also wrong extrapolation on use of that information in the contemporary context about legal processes.



2. Source Evaluation: Missing Key References

Current Sources:

- The Kompasian article and Qur'anic citation are appropriate but lack scholarly depth.

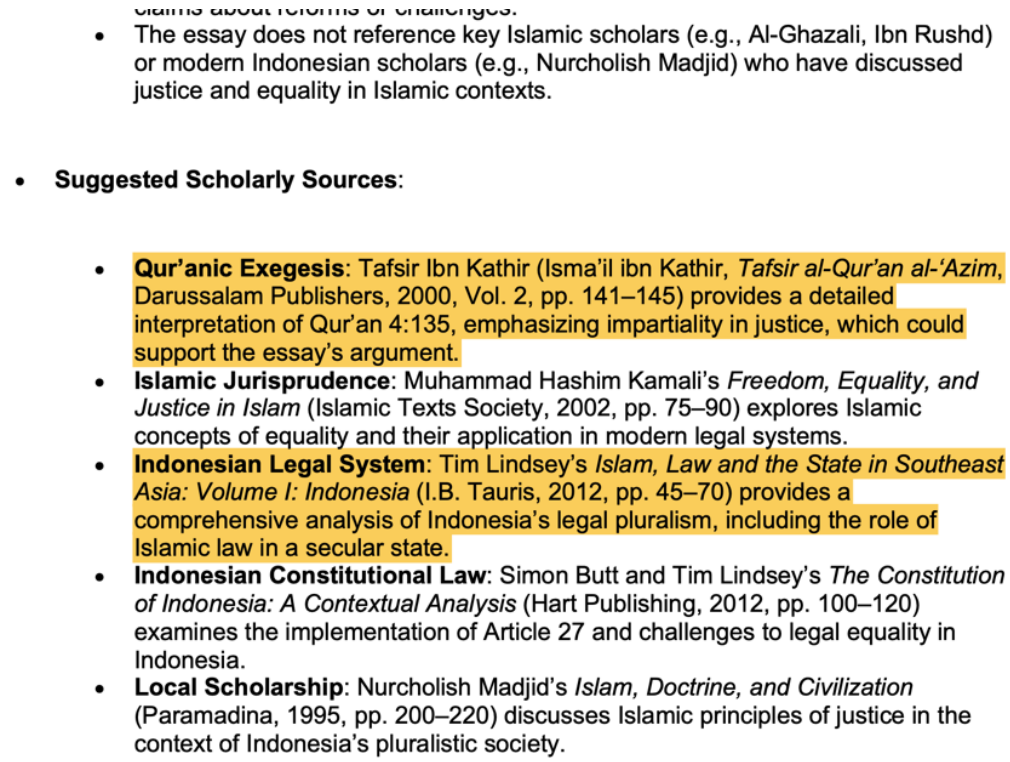
Recommended Additions:

- **Islamic Legal Theory:**
  - Al-Ghazali's *al-Mustaṣfā* on justice as a core Sharia objective.
  - Ibn Qayyim's *Iʿlām al-Muwaqqiʿīn* on equitable law enforcement.
- **Indonesia-Specific Studies:**
  - **Tim Lindsey** (e.g., *Islam, Law and the State in Indonesia*) on legal pluralism.
  - **Human Rights Watch reports** (e.g.,) documenting enforcement biases against minorities.

Figure 6. Example of errors found in DeepSeek's feedback

The pattern of errors displayed by DeepSeek would suggest an even greater lack of understanding of Islamic content than either ScholarAI or the other agencies, and this could simply be a reflection of the limitations of its training data from other sources and the limitations of its architecture to process specialized religious and legal content.

The hallucination patterns of Grok included some specific citation errors for classical Islamic texts and modern scholarship, or authors. In the error referring to Ibn Kathir's volume 'Tafsir al-Qur'an al-'Azim' volume 1, Grok incorrectly referenced specific pages 141-145, which he states were discussing verse 53 when in fact the pages referenced discussed Surah An-Nisa verses 23-29, and the accurate reference appeared on pages 166-173.



- The essay does not reference key Islamic scholars (e.g., Al-Ghazali, Ibn Rushd) or modern Indonesian scholars (e.g., Nurcholish Madjid) who have discussed justice and equality in Islamic contexts.

- **Suggested Scholarly Sources**:

  - **Qur'anic Exegesis**: Tafsir Ibn Kathir (Isma'il ibn Kathir, *Tafsir al-Qur'an al-'Azim*, Darussalam Publishers, 2000, Vol. 2, pp. 141–145) provides a detailed interpretation of Qur'an 4:135, emphasizing impartiality in justice, which could support the essay's argument.
  - **Islamic Jurisprudence**: Muhammad Hashim Kamali's *Freedom, Equality, and Justice in Islam* (Islamic Texts Society, 2002, pp. 75–90) explores Islamic concepts of equality and their application in modern legal systems.
  - **Indonesian Legal System**: Tim Lindsey's *Islam, Law and the State in Southeast Asia: Volume I: Indonesia* (I.B. Tauris, 2012, pp. 45–70) provides a comprehensive analysis of Indonesia's legal pluralism, including the role of Islamic law in a secular state.
  - **Indonesian Constitutional Law**: Simon Butt and Tim Lindsey's *The Constitution of Indonesia: A Contextual Analysis* (Hart Publishing, 2012, pp. 100–120) examines the implementation of Article 27 and challenges to legal equality in Indonesia.
  - **Local Scholarship**: Nurcholish Madjid's *Islam, Doctrine, and Civilization* (Paramadina, 1995, pp. 200–220) discusses Islamic principles of justice in the context of Indonesia's pluralistic society.

Figure 7. Errors found in Grok's feedback

The second error involved the author's manuscript Tim Lindsey and his book "Islam, Law and the State in Southeast Asia" as referencing a journal article when it was actually a book, and Grok fabricated a name for his journal, "Journal of Southeast Asian Studies, " but also fabrication of volume numbers, and page ranges. These instances suggest Grok also experienced difficulties distinguishing documents, specifically between book and article references and inaccurately representing bibliographical formats. It appears Grok's challenges in focused citation errors suggest Grok managed content and context well, and errors appeared related to accurate source representation and management for bibliographical details.

Source-Type Vulnerability Assessment

The assessment of vulnerability indicated identifiable patterns across varied kinds of Islamic sources, outlining vulnerabilities based on the types of Islamic scholarly works that AI systems encounter. Classical Islamic Texts were indicated to be very vulnerable to content misrepresentation and false citations. For instance, Iʿlām al-Muwaqqiʿīn by Ibn Qayyim, and, Tafsir al-Qurʾan al-ʿAzim by Ibn Kathir, both had misrepresentation in terms of inaccurate representation. This suggests that AI systems have difficulty grappling with classical types of scholarship that are complex and specialized types of inquiry. The nature of classical Islamic texts requires an advanced level understanding of Arabic terminology, historical contexts, and the methodology of the scholars, in this case, current AI systems lack an understanding of these crucial elements.

Contemporary Islamic Scholarship was generally vulnerable to false citations. For example, Abdul Rahman Azzam, Mohammad Hashim Kamali, Tim Lindsey and Simon Butt, all presented bibliographic errors in representation within their publications. This suggests that AI systems may contemplate the content and meaning of contemporary Islamic scholarship, however, they seem to struggle more with maintaining accurate bibliographies or publication databases for contemporary Islamic scholarship which is specialized.

Academic Journals of Islamic studies or related fields had a high level of susceptibilities to contextual misappropriation including the respective journals the AI systems did not correctly identify or understand the scope, intent, focus or subject areas of. This suggests that the AI systems struggle in understanding narrowly defined academic journals or literary forms, and are unable to apply scholarly works in the appropriate contexts.

Analysis

The study provides theoretical contributions to the understanding of AI hallucination in various ways and extends the previous features of hallucination in specific religious domains (Abbas et al., 2023). Overall, relative to the sources, when considering the degree of simulated, i.e., source fabrication and theological misstatements, our findings indicated that the three LLMs achieved sufficient training on their sizable Islamic contents so as not to exhibit the worst forms of hallucination. Additionally, it corroborates the work who outlined a degree of reliability of these frameworks wonder LLMs in specialized subjects, including religious domains (Tan, Cheng, & Ling, 2025). These findings indicate technology progress in the AI LLM for focused religious contents and supports findings in certain work which has implications of AI applications in culturally sensitive contexts (Kurata et al., 2025).

Nonetheless, the presence of citation errors and contextual misapplications showcases the continuing challenges of accurately representing finer details and the proper use of sources, exposing wider and denser issues related to AI training and knowledge representation, paralleling the evaluation put forth by Barnett et al (2024) of the issues associated with retrieval-augmented generation systems. The pervasive nature of citation errors emphasizes the limitations of current pathways for training AI in a way that properly prioritizes bibliographical accuracy, reinforcing the concerns put forward by Li et al (2024) that factual accuracy can also vary within domain-specific knowledge bases. In contrast to this, the examples of contextual misapplication illustrate the complexities and challenges of grappling with understanding the epistemological links between sources, and what constitutes proper "use" of the source.

Importantly, these findings represent the complex borderland between AI capability and specialized knowledge domains (Putri & Iswati, 2025), clearly demonstrating that general language ability does not equal accuracy in specialized contexts - a point of caution previously highlighted by Malik (2023) in the case of discussions of AI and Islamic thought; however, importantly, the impact of this nuance is not limited to Islamic knowledge, it extends to countless other specialized academic and professional contexts where very particular knowledges are unpacked with particular accuracy and cultural sensitivity, further supporting the growing body of literature questioning AI reliability by Abdalhussein (2024) and Abdelnour (2025).

Furthermore, in order to obtain more contextualization in our findings of AI hallucination in feedback related to Islamic matters, we have been keen to situate them within the existing literature related to AI using EFL contexts. The existing literature around AI in EFL contexts indicates that AI as a writing support tool is positively impacting EFL student performance. The report of Marzuki et al (2023) identified that teachers unanimously accepted that the use of AI tools helped improve students' writing quality, especially in the elements of content and organization. Similarly, Fitria (2023) presented evidence that EFL students with ChatGPT assistance experienced marked improvements in their writing skills overall.

Our study shows a clear void in the mostly favorable literature available on AI writing assistants by identifying the specific risks connected to hallucinations generated by culturally sensitive content. The preceding literature generally focused on the whole of writing related metrics showing improvement; however, our findings indicate that noted benefits from AI assistance in writing were potentially compromised when attempting state or convey specialized religious content where accuracy of the assistance has cultural and learning importance. This is a similar concern to certain study focusing investigation into the boundaries of the use of automated writing assistant tools in specialized areas (Mohamed et al, 2025).

The reference errors in our study are alarming, especially when already there is a clear emphasis of EFL education on proper attribution of sources and academic honesty. Previous researchers (Ghufron et al., 2024; Vieriu & Petrea, 2025) found accuracy issues with automated feedback systems as well, however, our findings seem to take into consideration the implications of the errors in a religiously sensitive context where bibliographic errors can result in significant cultural issues that are greater than academic performance indicators.

There is a significant impact on the influence of educators to help students develop EFL curriculum with Islamic content based on their research (Kurata et al., 2025) on how to teach religious studies with AI use. We think our findings conclude that the use of AI tools offers useful feedback to teachers on Islamic content, as it was written about by Achruh et al (2024) on AI in Islamic education, however, due to the problem of hallucinations it is important that teachers decide how to build implementation strategies.

The predominance of references errors in our own study affirms feelings in past research that AI-generated bibliographic references are problematic. In an Islamic studies context this is troubling, as it is important not only to properly refer from academic contexts to the Quran, Hadith collections, and classical texts, but also to be culturally considerate and respectful in ways Malik (2023) suggested, as reverence exists in the educational experience.

The errors in contextually applicable errors resonate with wider discussions of the AI systems inability to properly understand specialized areas. In their research on the use of LLMs with personalized learning Sharma et al (2025) indicated there were similar challenges in specialized areas. Overall, their findings indicated something much broader, that current training approaches in AI development may not be developing systems that utilize fine-tuned training for religious or culturally sensitive content.

Conclusion

The evidence has demonstrated that although all the LLM platforms exhibited some form of hallucination in the processing of Islamic content, the patterning of errors, frequency and severity of errors varied considerably across the systems. In the taxonomic analysis, it was found that citation errors were the most frequently observed form of hallucination, accounting for 58.3% of total identified errors, while contextual misapplications followed closely with a total identified percentage of 33.3%. This information reveals specific areas of vulnerability in how these AI systems are struggling: having accurate bibliographic information for Islamic scholarly sources and adequately matching sources to appropriate contextual applications.

On the other hand, the complete absence of source fabrication, attribution errors, and theological misstatements across all platforms implies that these models have encountered enough Islamic content, at the very least, to mitigate the more outrageous

hallucinations that could misinform students about Islamic beliefs, practices, or scholarly traditions. This study provides a platform for further research on the reliability of AI across multiple cultural and religious domains. Future research can explore longitudinal aspects of hallucination as AI models are updated and refined, to examine the diversity of types of Islamic content, and education settings, and build comparative studies with other types of religious and cultural content to understand commonalities across the domains and areas of specificity.

References

Abbas, A., Tirumala, K., Simig, D., Ganguli, S., & Morcos, A. S. (2023). SemDeDup: Data-efficient learning at web-scale through semantic deduplication. aRxIV: Journal of Computer Science. https://doi.org/10.48550/arXiv.2303.09540.

Abbas, N. F., Qasim, T. A., & Jasim, H. A. S. (2023). Request Constructions in Classical Arabic versus Modern Arabic: A Corpus-based Study. Journal of Ethnic and Cultural Studies, 10(5), 1–15. https://doi.org/10.29333/ejecs/1598

Abdalhussein, H. F. (2025). Comparative Analysis Of AI And Human Translations Of Qur'anic Legal Verses In Surah Al-Ma'idah. Ijaz Arabi Journal of Arabic Learning, 8(1). https://doi.org/10.18860/ijazarabi.v8i1.28804

Abdelnour, M. G. (2025). Artificial Intelligence and the Islamic Theology of Technology: From "Means" to "Meanings" and from "Minds" to "Hearts." Religions, 16(6), 796. https://doi.org/10.3390/rel16060796

Achruh, Rapi, M., Rusdi, M., & Idris, R. (2024a). Challenges and Opportunities of Artificial Intelligence Adoption in Islamic Education in Indonesian Higher Education Institutions. International Journal of Learning, Teaching and Educational Research, 23(11), 423–443. https://doi.org/10.26803/ijlter.23.11.22

Achruh, Rapi, M., Rusdi, M., & Idris, R. (2024b). Challenges and Opportunities of Artificial Intelligence Adoption in Islamic Education in Indonesian Higher Education Institutions. International Journal of Learning, Teaching and Educational Research, 23(11), 423–443. https://doi.org/10.26803/ijlter.23.11.22

Addzaky, K. U., Hasanah, U., Hasanah, I. F., & Asy'ari, M. (2025). Contextualizing Liberation Theology in Indonesian Religious Literature Through the Contributions of K.H. Abdurrahman Wahid. Jurnal Lektur Keagamaan, 23(1), 142–178. https://doi.org/10.31291/jlka.v23i1.1285

Ahmed Mohamed, F., Madihah Sheikh Abdul Aziz, & Murni Mahmud. (2023). Cues to Deception in Islamic Websites Text and Design. International Journal on Perceptive and Cognitive Computing, 9(1), 14–23. https://doi.org/10.31436/ijpcc.v9i1.299

Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., & Abdelrazek, M. (2024). Seven failure points when engineering a retrieval augmented generation system. Proceedings - 2024 IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI, CAIN 2024, 194–199. https://doi.org/10.1145/3644815.3644945

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., & Evans, O. (2024). the Reversal Curse: Llms Trained on "a Is B" Fail To Learn "B Is a." 12th International Conference on Learning Representations, ICLR 2024. https://doi.org/10.48550/arXiv.2309.12288.

Bobula, M. (2024). Generative artificial intelligence (AI) in higher education: a comprehensive review of challenges, opportunities, and implications. Journal of

Learning Development in Higher Education, (30), 30. https://doi.org/10.47408/jldhe.vi30.1137

Casal-Otero, L., Catala, A., Fernández-Morante, C., Taboada, M., Cebreiro, B., & Barro, S. (2023). AI literacy in K-12: a systematic literature review. International Journal of STEM Education, 10(1), 29. https://doi.org/10.1186/s40594-023-00418-7

Deiniatur, M., & Hasanah, U. (2024). Multicultural Education of EFL Classrooms in Indonesian Islamic Higher Education: Challenges and Strategies. Pedagogy : Journal of English Language Teaching, 12(2), 158–177. https://doi.org/10.32332/joelt.v12i2.9542

Fathia Baresh, E. (2024). Incorporating Cultural Background in Teaching EFL Writing in Libyan: A Critical Literature Review. Al-Lisan, Vol. 9, pp. 1–16. https://doi.org/10.30603/al.v9i1.3811

Fitria, T. N. (2023). Artificial intelligence (AI) technology in OpenAI ChatGPT application: A review of ChatGPT in writing English essay. ELT Forum: Journal of English Language Teaching, 12(1), 44–58. https://doi.org/10.15294/elt.v12i1.64069

Ghufron, M. A., Rohman, T., & Aditia, M. R. (2024). English Language Teaching Materials and Religious Moderation in Indonesian Primary Schools: A Narrative Case Study. Lingua Cultura, 18(1), 141–153. https://doi.org/10.21512/lc.v18i1.11982

Ilyas, M., & Afzal, N. (2021). Teaching literary texts through cultural model to efl students. Eurasian Journal of Applied Linguistics, Vol. 7, pp. 22–30. https://doi.org/10.32601/ejal.911502

Kimsesiz, F. (2024). Reflections from Learners of English as a Foreign Language in a Multicultural Non-native Context. Acuity: Journal of English Language Pedagogy, Literature, and Culture, Vol. 9, p. 2024. Retrieved from https://jurnal.unai.edu/index.php/acuity

Kurata, L., Ayanwale, M. A., Molefi, R. R., & Sanni, T. (2025). Teaching religious studies with artificial intelligence: A qualitative analysis of Lesotho secondary schools teachers' perceptions. International Journal of Educational Research Open, 8. https://doi.org/10.1016/j.ijedro.2024.100417

Lee, C., Kim, J., Lim, J. S., & Shin, D. (2025). Generative AI risks and resilience: How users adapt to hallucination and privacy challenges. Telematics and Informatics Reports, 19. https://doi.org/10.1016/j.teler.2025.100221

Li, J., Yuan, Y., & Zhang, Z. (2024). Enhancing LLM Factual Accuracy with RAG to Counter Hallucinations: A Case Study on Domain-Specific Queries in Private Knowledge-Bases. AirXv: Journal of Computer Science: Computation and Language. https://doi.org/10.48550/arXiv.2403.10446

Mahmud, M., & Abubakar, A. (2015). Investigating the act of deception in online-Islamic Content. Proceedings - 2014 3rd International Conference on User Science and Engineering: Experience. Engineer. Engage, i-USEr 2014, 242–247. Malaysia. https://doi.org/10.1109/IUSER.2014.7002710

Malik, S. A. (2023). Artificial Intelligence and Islamic Thought: Two Distinctive Challenges. Journal of Islamic and Muslim Studies, Vol. 8, pp. 108–115. https://doi.org/10.2979/jims.00020

Marzuki, Widiati, U., Rusdin, D., Darwin, & Indrawati, I. (2023). The impact of AI writing tools on the content and organization of students' writing: EFL teachers' perspective. Cogent Education, 10(2). https://doi.org/10.1080/2331186X.2023.2236469

Mohamed, A. M., Shaaban, T. S., Bakry, S. H., Guillén-Gámez, F. D., & Strzelecki, A. (2025). Empowering the Faculty of Education Students: Applying AI's Potential for

Motivating and Enhancing Learning. Innovative Higher Education, 50(2), 587–609. https://doi.org/10.1007/s10755-024-09747-z

Mozelius, P., Lindqvist, M. H., Cleveland-Innes, M., Jaldemark, J., & Sundgren, M. (2024). Educating the Educators on Generative Artificial Intelligence in Higher Education. Proceedings of the 4th International Conference on AI Research, ICAIR 2024, 4(1), 296–302. https://doi.org/10.34190/icair.4.1.3026

Putri, F. R., & Iswati, H. D. (2025). A qualitative exploration of students use of humata.ai to understand thesis literature: a cognitive load perspective. Wiralodra English Journal, 9(2), 166–179. https://doi.org/10.31943/wej.v9i2.471

Sharma, S., Mittal, P., Kumar, M., & Bhardwaj, V. (2025). The role of large language models in personalized learning: a systematic review of educational impact. Discover Sustainability, 6(1), 243. https://doi.org/10.1007/s43621-025-01094-z

Sun, Y., Sheng, D., Zhou, Z., & Wu, Y. (2024). AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. Humanities and Social Sciences Communications, 11(1), 1278. https://doi.org/10.1057/s41599-024-03811-x

Syuhda, N., Irwansyah, D., Adiwijaya, S. N., Harya, T. D., Deiniatur, M., & Septiyana, L. (2024). Contextualizing Islamic Traditions in English Language Teaching at Indonesian Islamic Higher Education. Forum for Linguistic Studies, 6(3), 260–273. https://doi.org/10.30564/fls.v6i3.6739

Tan, X., Cheng, G., & Ling, M. H. (2025). Artificial intelligence in teaching and teacher professional development: A systematic review. Computers and Education: Artificial Intelligence, 8. https://doi.org/10.1016/j.caeai.2024.100355

Vieriu, A. M., & Petrea, G. (2025). The Impact of Artificial Intelligence (AI) on Students' Academic Development. Education Sciences, 15(3), 343. https://doi.org/10.3390/educsci15030343

Zhang, K. (2025). Enhancing Critical Writing Through AI Feedback: A Randomized Control Study. Behavioral Sciences, Vol. 15. Basel, Switzerland. https://doi.org/10.3390/bs15050600