Journal of Integrated Sciences
Volume 6, Issue 2, March 2026
ISSN: 2806-4801

EVALUATING AN AI MODEL'S QUR'ANIC RECALL AND CITATION IN ARABIC AND SAHIH INTERNATIONAL ENGLISH: A REAL-DATA PILOT ETHICAL AND ANALYTICAL REVIEW

Alexandra V. Maragha, Ph.D. Islamic Sciences

Assistant Professor, Department of Islamic Studies,
International Islamic University of Minnesota (IIUMN)

ABSTRACT

Large Language Models (LLMs) in Artificial Intelligence (AI) are commonly used to query religious texts, including the Holy Qur'an, raising questions about accuracy, reliability, and contextual fidelity, as minor deviations in Quranic recall can carry significant theological, ethical, and educational implications. This quantitative study presents a reproducible bilingual, pilot grounded evaluation of ChatGPT-5's Quranic recall, citation, and thematic search abilities based on Arabic and Sahih International English translation operating three distinct task categories: (a) Verse Completion, (b) Citation-from-Text, (c) Thematic Retrieval in themes such as Patience, Charity, Fasting, and Mercy. The results displayed near-perfect accuracy across both languages for verse completion, with a reading of 1.00 in Arabic and English. The results of citation-from-text recall achieved 0.97 in Arabic and 1.00 in English. Regarding thematic retrieval, evaluated by Precision@10, the themes of Patience, Charity, and Mercy were identified, achieving 1.00 in Arabic and 0.90 for Fasting, while maintaining 1.00 in English. The implications of this study demonstrate that AI chatbot LLMs, such as ChatGPT-5, can provide accurate Quranic recall when grounded; however, Arabic syntax and morphology remain key challenges for AI systems, while ethical and bias considerations leave Quranic thematic analysis (tafsīr) to be avoided.

Journal of Integrated Sciences
Volume 6, Issue 2, March 2026
ISSN: 2806-4801

Keywords: *Artificial Intelligence (AI), Qur'an, Large Language Models (LLMs), Arabic, English.*

Corresponding author: Dr. Alexandra V. Maragha can be contacted at avmaragha@gmail.com

Acknowledgement: This study acknowledges the Tanzil Project for providing verified Quranic texts and the broader Quranic NLP community for resources and methods that informed this work.

1. INTRODUCTION

The Holy Qur'an is the primary source, holding supreme authority in Islamic epistemology, legal knowledge, and Divine knowledge and guidance. The language and precision of the wording, lexical, grammar, and miraculous structure are revered and studied by classical and contemporary scholars within Islamic and Secular contexts. Allah The Almighty has revealed in the Qur'an, heeding the challenge of Quranic replication, stating,

"وَإِن كُنتُمْ فِي رَيْبٍ مِّمَّا نَزَّلْنَا عَلَىٰ عَبْدِنَا فَأْتُوا بِسُورَةٍ مِّن مِّثْلِهِۦ وَٱدْعُوا۟ شُهَدَآءَكُم مِّن دُونِ ٱللَّهِ إِن كُنتُمْ صَٰدِقِينَ"

"And if you are in doubt about what We have sent down [i.e., the Qur'ān] upon Our Servant [i.e., Prophet Muḥammad (ﷺ)], then produce a *sūrah* the like thereof and call upon your witnesses [i.e., supporters] other than Allāh, if you should be truthful" (Quran 2:23, Sahih International Translation).

Muslims of all ages strive to recite and memorize the Qur'an as an everyday act of worship and learning, from which the oral

tradition of Quranic memorization (*ḥifẓ*) and verification (*isnād*) of the chain of teaching sources, tracing back to the Prophet Mohammad (ﷺ), is unique to Islam and the Islamic tradition. Within this context, artificial intelligence (AI) systems, specifically Large Language Models (LLMs), are used to recall or quote Quranic verses due to their benefits of rapid information retrieval, widespread user access, and natural-language querying. At the same time, the risks of using AI, such as misquotations, inaccurate *sūrah*: *āyah* attributions, and conflation of translations with paraphrasing presented as verbatim scripture, exist. These risks pose major implications of a larger problem of misguidance, as users who do not have strong Quranic knowledge could inevitably propagate misinformation within public discourse, causing ethical misrepresentation of the Divine word of Allah The Almighty. Likewise, AI systems lack the ability to retrieve authoritative sources unless grounded sources are provided to ensure credibility. Moreover, Arabic language recall should be cautioned within AI structures relating to translation, linguistic, and computational biases that occur in non-grounded uses of AI LLMs (Habash, 2010).

Therefore, any evaluation of Quranic recall by AI must differentiate between two methods: (1) closed book, based on internal parameters; and (2) grounded, based on a provided corpus during recall analysis. Closed-book provides insights into a model's parametric memory, while grounded models offer a responsible demonstration of ethical assurance for authoritative texts such as the Holy Qur'an.

This study provides a bilingual (Arabic and Sahih International English) pilot evaluation of Quranic recall in a grounded context. A verse-aligned Qur'an corpus was used to ensure synchronization of Arabic and English translation *sūrah* and *āyah* indices and to measure the performance of three categories: verse completion, citation-from-text, and thematic retrieval. Additionally, the challenge of Arabic computational orthographic variability, where the same lexical entry can differ across forms such as diacritics, *alif/hamza*, *tāʾ marbūṭah*, and yāʾ, among others, was addressed by normalizing for fair evaluation and retrieval outcomes in keyword methods and exact-string matching. Likewise, it provides a reproducible language-specific evaluation of grounded Quranic recall with high-precision results, achieving near-perfect accuracy in structured tasks, meeting the reliability needs of Islamic education and scholarship.

At the same time, it highlights the necessity of grounded approaches, especially when Arabic language is tested for thematic retrieval, showing the lack of Arabic language depth and understanding without internal computational bias among AI LLMs within grounded and closed book or common AI LLM usage with further education among general and scholarly users to be aware of the limitations and lack of accuracy and reliability of Quranic recall of AI LLMs to which insightful development in adding depth to Arabic language recall and understanding is needed to remove ethical bias and basic recall abilities while quantifying confidence and risk.

2. REVIEW OF LITERATURE

2.1 AI Reliability, Hallucination, and Digital Quranic Studies

The reliability of LLMs remains a major concern in the Natural Language Processing (NLP) community. Studies report that even the most advanced models can produce fluent but inaccurate or correct statements, called hallucinations, even when prompted or when pressured to answer despite uncertainty (Ji et al., 2023). Bender & Koller (2020) emphasize that form does not guarantee meaning fidelity, while a model may generate convincing text while deviating from the source. Likewise, Alqahtani & Atwell (2011) state, "several deficiencies exist with the Quranic verses (*Āyāt*) retrieved for a query using the existing keyword search techniques" (p.1), relating to problems in irrelevant verses being retrieved with relevant verses being unordered or not retrieved, relating to misunderstandings in input meaning and processing retrieval. Likewise, Atwell et al. (2011) emphasize, "artificial Intelligence research on language and text is generally based on a Corpus, a machine-readable dataset of the text being researched, enriched with metadata and tags or annotations showing morphological analyses, Part-of-Speech tags, etc." (p.1), emphasizing a grounded approach to AI study and use.

With the recent digitization of classical religious texts and the Qur'an corpus, multiple projects have produced high-quality digital resources. The Tanzil project curates verified Arabic text and translations of the Qur'an, including Sahih International English. At the same time, the variety offers version control and

verse-level alignment for use by researchers, software developers, and digital contexts.

Studies and projects such as The Quranic Arabic Corpus, completed by Dukes & Habash (2010), introduced morphological annotation that supports parsing and linguistic analysis. QurSim, produced by Sharaf & Atwell (2012), is a dataset for measuring semantic similarity in short texts derived from the Qur'an. These sources and projects have established best practices in digital Quranic scholarship and use, ensuring digital accuracy through canonical editions, data organized by verse, and content published to support reproducibility while maintaining Islamic ethical guidelines and considerations.

2.2 Arabic Natural Language Processing (NLP) Challenges and Normalization

The Arabic language presents distinctive computational challenges, including considerations of diglossia, templatic morphology, and orthographic variation. Scholars such as Habash (2010) detail how diacritics (*tashkīl*), *alif/hamza* forms (ٱ/أ/إ/آ), *tāʾ marbūṭah* (ة), *yāʾ* vs. *alif maqṣūrah* (ى/ي), and tatwīl (ـ) complicate exact matching. As such, these elements of the Arabic language have been removed in normalization for digital preprocessing within retrieval and classification tasks. This same framework has been applied to this study to ensure keyword-based methods behave consistently across varying elements within the digital sphere, such as editions and fonts. Without normalization, Arabic thematic retrieval may appear weaker than English within AI Quranic recall settings.

2.2.1 Arabic Orthography, Diacritics, and Computational Ambiguity

The removal of diacritics (*tashkīl*) in Arabic natural language processing (NLP) is routine, resulting in orthographic ambiguity. In contrast to languages with explicit vowel representation, such as English, Arabic usually preserves only consonantal forms, leaving short vowels and other grammatical markers. This produces words that could correspond to multiple lexical, grammatical, or semantic interpretations, depending on context (Habash, 2010). Most Quranic corpora use simplified scripts without diacritics for search and computational processing (Zerrouki & Balla, 2017). Likewise, AI models infer grammatical rules and meanings using probabilistic algorithms rather than learned determination, ultimately degrading AI performance on tasks that require fine linguistic judgement, including part-of-speech tagging and semantic classification (Farghaly & Shaalan, 2009). Understandings such as normalization, as demonstrated in the present study, can improve consistency and retrial performance but, at the same time, remove syntactic and semantic disambiguation.

2.2.2 Morphological Depth and Root-Pattern Structure

The Arabic language is characterized by a non-concatenative Morphological system where lexical meaning reflects the strong interaction between consonantal roots and vocalic patterns. Numerous verbal forms, nouns, participles, and verbal nouns (*maṣdar*) can derive from a single root, which can increase in complexity in computational modeling (Habash, 2010). Previous studies, such as that of Farghaly & Shaalan (2009), recognize the limitations of keyword-based retrieval systems used in thematic

analysis, resulting in under-recall, but these limitations can be mitigated through morphology-conscious indexing, which creates a dependency on high-quality analyzers. The depth of Arabic root forms is simplified to a single lexical item, creating asymmetries in Arabic and English recall performance, and explaining why thematic retrieval can achieve perfect precision in English but not in Arabic for Quranic recall.

2.2.3 Arabic Syntax, Word Order, Parsing Limitations, and Rhetoric

Arabic syntax is complex, with a flexible word order that allows verb-subject-object (VSO) constructions and subject-verb-object (SVO) and other variations for rhetorical effect. Agreement patterns differ depending on word order and the role of syntax that may not be overly marked (Maamouri et al., 2004). Such properties can complicate syntax, especially in vocalized text.

For AI LLMs performing thematic retrieval without explicit Arabic lexical markers, this can result in the nonrecognition of syntactic structures such as conditional clauses, exemptions, or narrative structures, even when processing Modern Standard Arabic (Diab, 2009). Projects such as the Penn Arabic Treebank (Maamouri et al., 2004) have advanced parsing research; however, challenges remain in advanced systems. Additionally, AraBERT, developed by Antoun et al. (2020), has improved Arabic-language understanding in computational contexts but has reduced performance on syntax-heavy and out-of-domain tasks (Antoun et al., 2020).

The rhetorical structures in Quranic Arabic are displayed through implicit and legal conditions, sequencing, and thematic cohesion throughout the Qur'an, in contrast to the use of repetitive lexical markers. AI is unable to recognize these elements unless they are modelled, leaving only the human tradition of interpreting meaning (*tafsīr*) to fully consider them (Dukes & Habash, 2010; Sharaf & Atwell, 2012). For example, keyword-based methods are known to under-represent theoretically relevant verses, when meaning is across syntax and context.

2.3 Translation of a Computational Mediation Layer

Translation is an essential element in AI performance. As the Holy Qur'an is originally in Arabic, English translations, such as Sahih International, when provided in a grounded context, offer higher retrieval precision in computational tasks. This is rooted in reduced morphological variation, fixed work orders, and explicit grammatical indications in English. Scholars of translation studies caution that computational convenience should not be mistaken for linguistic superiority or theological primacy (Kidwai, 2011). In contrast, it highlights the need to treat Arabic as the authoritative source while recognizing that this requires greater computational demands. As such, challenges such as asymmetry can arise when comparing Arabic and English recall performance, and calls for AI systems to be designed with ethical imperatives that account for linguistic complexity, such as in Arabic, rather than overlook it.

2.4 AI Biases in Arabic Language Processing

While AI systems and LLMs are rooted in programmed knowledge that has evolved over time, they also reflect intentional or ideological biases in their structural and computational contexts. A broader view of Natural Language Processing (NLP) suggests that bias develops through disparities in training data, model architecture, tokenization strategies, and evaluation benchmarks, rather than through explicit linguistic or cultural prejudice (Bender & Friedman, 2018). This understanding is critical for assessing AI recall performance in Arabic, particularly Quranic Arabic.

English and other Indo-European languages are the predominant language structures of web-scale corpora from which LLMs are trained and are thus overrepresented. At the same time, developers are not concerned with Islamic ethical principles and may even intend to produce untruthful results through unmitigated responses that summarize or reference unauthoritative corpora or sources, rather than a structured authority universally. Likewise, the political intentions of such technology skew results and usage to align with a creator's worldview rather than actual truth, leading LLMs and AI tools such as ChatGPT-5 to hold internal biases against Islam and Islamic content by default (Hemmatian, 2023). Likewise, Arabic content is often unevenly distributed across dialects, registers, and genres, and is completely scarce in high-quality annotated datasets (Habash, 2010; Farghaly & Shaalan, 2009). These contextual limitations and references create an imbalance that leads to statistical exposure bias, leaving languages rich in morphology, such as Arabic, underprocessed. This lack of

Arabic data volume is one cause of infrastructural constraints on data availability, which in turn leads to LLM biases.

At the same time, model architecture itself reveals structural bias. Arabic uses non-concatenative root-pattern morphology, extensive cliticization, and flexible word order, bringing multiple linguistic layers of meaning (Habash, 2010). This contrasts with English-language characteristics, such as concatenative structure, mostly fixed word order, limited inflection, and explicit grammatical markers. When seeking distinction within transformer-based models that avoid tokenization and morphological breakdown of meaning structures and inferred meanings, a challenge of representation arises, particularly when NLP benchmarks prioritize surface lexical matching and keyword-based relevance metrics, which favor lower morphological variability and more explicit semantic encoding (Bender & Koller, 2020). At the same time, tokenization, such as subword tokenizers and Byte Pair Encoding (BPE), often fragments Arabic during translation or Arabic recall within LLMs, producing semantically opaque tokens (Antoun et al., 2020).

In Quranic Arabic, such challenges are heightened by the high rhetorical density and elliptical expression. Prior studies have noted that Quranic NLP and existing computational models have difficulty recalling meaning without integrating tafsīr-based representations (Dukes & Habash, 2010; Sharaf & Atwell, 2012). Likewise, in a similar study scholars such as Bhojani & Schwarting (2023) analyzed ChatGPT-3 and other AI LLMs in their ability to accurately recall the verse in the Qur'an, within an open-book context, without a grounded source, noting the

operational challenges of AI in that “nearly all deployed LLMs operate stochastically; that is, repeated identical queries will yield nonidentical responses sampled from an underlying probability distribution” (p.558) when running prompt tasks. This adds to the unreliability of ungrounded AI use, especially when used for Quranic recall, where overall trends are employed for reliability assessment.

As such, the limitations presented in bilingual AI recall provide a backdrop for the current performance differences and motivate the research gap in this study, thereby supporting the analytic framework to be adopted.

2.5 Research Gap and Conclusion

The literature exemplifies the challenges that AI-based text retrieval systems face due to the orthographic ambiguity, morphological depth, cliticization, and syntactic flexibility that Arabic possesses. Studies in NLP research have indicated such challenges however, the present study builds.

3. RESEARCH METHODOLOGY

3.1 Data

To establish grounding, verse-aligned texts of the Arabic Qur’an (simple script) and the Sahih International English translation, each with 6,236 ayahs from Tanzil, were used. Both files include explicit *sūrah* and *āyah* identifiers, enabling a clean merge into a bilingual CSV with the following fields: {Surah, Ayah, Arabic, English_Sahih}. This alignment guarantees one-to-one correspondence across languages for scoring and analysis.

3.2 Preprocessing

Arabic normalization was applied to reduce orthographic variation. Diacritics were removed and mapped to forms of canonical equivalents: : ي→ئ ,و→ؤ ,ى→ي ,ا→أ/إ/آ/ٱ, and ه→ة, and removed tatwīl (ـ). Table 1 indicates normalization. English text was lowercased for keyword matching only; exact-quote tasks used original case-preserving strings.

Table 1. Arabic text normalization rules applied in preprocessing

Original form	Normalized form
أ ,إ ,آ ,ٱ	ا
ى	ي
ؤ	و
ئ	ي
ة	ه
Remove	Diacritics, tatwīl ـ

Source: The Author's own work.

3.3 Tasks

Three families of tasks were defined as follows:

a. Verse Compilation: Given the first four words (Arabic) or five words (English), the system must return the exact verse text to establish recall boundaries evaluated by exact string match.

b. Citation-from-text: Given a verse (Arabic or English), the system must return (*Sūrah*, *Āyah*) to test reference linking and indexing fidelity.

c. Thematic Retrieval: Given a theme label (e.g., Patience/*Ṣabr*; Charity/*Infāq*; Fasting/*Ṣiyām*; Mercy/*Raḥmah*), the system retrieves top-10

relevant verses by keyword scoring. Transparent keyword lists were used per language, with Arabic matching performed on normalized text.

3.4 Metrics

The areas of verse completion and citation-from-text were scored with accuracy within the proportion of exact matches. The third area of thematic retrieval was scored using Precision@10 (P@10), which indicates the fraction of the top-10 results that contain one or more theme keywords. P@10 with keyword ranking is a baseline in which future work should consider human judgment to measure semantic relevance more deeply.

3.5 Procedure

Random sampling was used to select 120 verses (n=120) for the verse completion and citations tasks, ensuring a broad representation within the Qur'an corpus. Thematic retrieval used the full corpus and computed P@10 per theme and language. Results were collected and reported separately for Arabic and English, and addressed the relevance of normalization on Arabic retrieval.

4. RESULTS

The results of this study identify necessary findings and frameworks for AI LMMs to be further evaluated for use in sacred scripture, such as the Holy Qur'an and the Quranic Arabic language, with emphasis on the differences between grounded and open-book execution contexts within AI models.

4.1 Structured Tasks

In the grounded context, verse completion achieved 1.00 accuracy in both Arabic and English, indicating perfect accuracy in quotation recall when grounded with the verified corpus. Citation-from-text reached 1.00 in English and 0.97 in Arabic, accounting for technical loading errors during a reload step rather than true look-up errors. Enforcing a strict output schema in future studies may eliminate such issues.

Table 2. Accuracy results for structured tasks (completion and citation)

Task	Language	Accuracy
Verse Completion	Arabic	1.00
Verse Completion	English (Sahih)	1.00
Citation	Arabic	0.97
Citation	English (Sahih)	1.00

Source: The Author's own work.

4.2 Thematic retrieval

With Arabic normalization, P@10 reached 1.00 for Patience, Charity/Spending, and Mercy in both Arabic and English. For the fasting theme, thematic recall was 0.90 in Arabic and 1.00 in English.

Bilingual evaluation likewise sampled the following verses and translations used in thematic analysis:

Patience/*Ṣabr*: (Qur'an 2:153)

Arabic: "يَا أَيُّهَا الَّذِينَ آمَنُوا اسْتَعِينُوا بِالصَّبْرِ وَالصَّلَاةِ ۚ إِنَّ اللَّهَ مَعَ الصَّابِرِينَ

English (Sahih): O you who have believed, seek help through patience and prayer. Indeed, Allah is with the patient.

Fasting/*Ṣiyām*: (Qur'an 2:183)

Arabic: "يَا أَيُّهَا الَّذِينَ آمَنُوا كُتِبَ عَلَيْكُمُ الصِّيَامُ كَمَا كُتِبَ عَلَى الَّذِينَ مِن قَبْلِكُمْ لَعَلَّكُمْ تَتَّقُونَ

English (Sahih): O you who have believed, decreed upon you is fasting as it was decreed upon those before you that you may become righteous.

Charity/*Infāq*: (Qur'an 2:261)

Arabic: مَّثَلُ الَّذِينَ يُنفِقُونَ أَمْوَالَهُمْ فِي سَبِيلِ اللَّهِ كَمَثَلِ حَبَّةٍ أَنبَتَتْ سَبْعَ سَنَابِلَ فِي كُلِّ سُنبُلَةٍ مِّائَةُ حَبَّةٍ ۗ وَاللَّهُ يُضَاعِفُ لِمَن يَشَاءُ ۗ وَاللَّهُ وَاسِعٌ عَلِيمٌ

English (Sahih): The example of those who spend their wealth in the way of Allah is like a seed [of grain] which grows seven spikes; in each spike is a hundred grains. And Allah multiplies [His reward] for whom He wills. And Allah is all-Encompassing and Knowing.

Mercy/*Raḥmah*: (Qur'an 39:53)

Arabic: "قُلْ يَا عِبَادِيَ الَّذِينَ أَسْرَفُوا عَلَىٰ أَنفُسِهِمْ لَا تَقْنَطُوا مِن رَّحْمَةِ اللَّهِ ۚ إِنَّ اللَّهَ يَغْفِرُ الذُّنُوبَ جَمِيعًا ۚ إِنَّهُ هُوَ الْغَفُورُ الرَّحِيمُ

English (Sahih): Say, "O My servants who have transgressed against themselves [by sinning], do not despair of the mercy of Allah. Indeed, Allah forgives all sins. Indeed, it is He who is the Forgiving, the Merciful."

Table 3. Precision@10 results for thematic retrieval

Theme	Language	Precision@10	Results
Patience	English (Sahih)	1.00	10
Patience	Arabic (normalized)	1.00	10
Charity/Spending	English (Sahih)	1.00	10
Charity/Spending	Arabic (normalized)	1.00	10
Fasting	English (Sahih)	1.00	10
Fasting	Arabic (normalized)	1.00	9
Mercy	English (Sahih)	1.00	10
Mercy	Arabic (normalized)	1.00	10

Source: The Author's own work.

These examples demonstrate how grounded bilingual alignment supports both exact quotation tasks and thematic analysis and recall across both Arabic and English.

5. DISCUSSION

5.1 Arabic Morphological Richness and Computational Constraints

This study demonstrates the contrasts that arise when incorporating linguistic recall in rich morphological languages, such as Arabic, in which meaning is produced through root connections and patterns, compared to English. For example, the root S-W-M (ص - و - م) produces both *ṣiyām* (صيام) and *ṣawm* (صوم), which hold the same meaning related to fasting, but appear in different grammatical constructions.

Thematic retrieval reflects this discrepancy in morphological variation, with fasting attaining Precision@10 = 0.90 in Arabic, compared to 1.00 in English. Within the grounded framework,

fasting is emphasized through *ṣiyām*, with at least one relevant verse displaying an alternative morphological understanding or conveying it contextually. In contrast, English reduces the discrepancies in accurate recall by AI LLMs because the single lexical item “fasting” is defined, allowing grounded keyword retrieval to achieve higher accuracy.

This illustrates a broader need for morphological analyzers to be further integrated into LLM-based AI models, which are computationally expensive but may resolve orthographic variation when diacritics are absent (Farghaly & Shaalan, 2009).

In this study, the AI system did not perform full syntactic parsing for thematic retrieval. As a result, it could not reliably infer conceptual relevance when a theme was expressed indirectly through syntactic structure rather than through explicit lexical items. This limitation is well documented in Arabic NLP literature and persists even in transformer-based models such as AraBERT (Antoun et al., 2020).

The normalization process accounted for the procedures that were grounded for this study. As such, the removal of clitics, the flexibility of word order, and syntactic ambiguity result in the understanding of subjects as objects, or in identifying the scope of negation and conditional clauses as probabilistic rather than deterministic (Maamouri et al., 2004).

5.2 Quranic Recall Complexity

Quranic Arabic presents additional challenges of computational recall due to its rhetorical density and semantic compression. Concepts, such as fasting, are expressed or referenced through discussions of exemptions, compensatory acts, or temporal

markers such as Ramadan, without restating the exact term *ṣiyām*. In contrast to existing LLMs, human readership and tradition allow these patterns and implications to be recognized without exact work-match correlations, demonstrating their lack of depth unless explicitly grounded in a reference to an encoding modeling structure that allows for semantics and syntax contexts.

5.3 English Translation Sahih International as a Computational Advantage

The results of this study show that English translations often outperform Arabic translations in AI retrieval tasks, due to their computational simplicity. Sahih International's translation style also contributes to consistency by using standardized terminology across verses. When placed within a computational encoding context, such as AI retrieval, it is optimal for such tasks. At the same time, this lack of computational encoding is needed to achieve the same understanding and recall results in Quranic Arabic computational tasks, treating the Arabic Qur'an corpus as an authoritative source.

5.4 Bias and Ethical Approaches to AI Arabic and Quranic Analysis (*Tafsīr*)

This study also seeks to understand the bias within AI LLMs, such as ChatGPT-5, in how they recall information. Due to the grounded nature of the study, the quantitative results on pure corpus recall indicate minimal computational bias related to intentional or ideological elements but highlight critical structural and linguistic challenges in Arabic NLP. As such, structured tasks resulted in perfection, while thematic retrieval

displayed a Precision@10 of 0.90 in Arabic, compared to 1.00 in English, aligning with the literature indicating that morphological realizations, such as *ṣiyām and ṣawm*, are unrecognizable in the context of explicit repetition. As such, these results indicate AI models lack morphology and syntax-aware mechanisms to generalize across implicit and variant expressions. This brings further discussion into the true abilities of computerized LLMs, such as AI, into the ability to fully understand thematic retrieval, even in a grounded setting, as the Quran was revealed in Arabic, and the rules of Arabic play a key role in understanding the meaning and content of the Quran well (Selim, 2018). The discrepancy is that AI infrastructures are not sufficiently equipped to model Arabic's linguistic complexity.

Ethically and pedagogically, this distinction matters, as recognizing limitations becomes crucial to awareness and methodology in Quranic recall and understanding. Scholars emphasize that understanding Arabic helps in appreciating the nuances, metaphors, and deep meanings of the Quran and allows for the more accurate use of *tafsīr* (interpretation) (Osman & Hassan, 2022), to the extent that, without such abilities, one cannot accurately and fully recognize the meaning and perform accurate Quranic analysis (*tafsīr*). Scholars such as Bhojani & Schwarting (2023) ultimately argue, "that LLM's are not trustworthy and thus should not be employed without a critical disposition and due attention to truth" (p.557). Islamic ethical considerations likewise test AI chatbots such as ChatGPT-5 to ultimately examine the source of the AI worldview portrayed in their training, and the examination of the ethics and intentions of the creators of ultimately secular

tools with no Islamic oversight will leave room for false interpretation and reproduction (Hemmatian, 2023).

As such, Arabic becomes an essential foundation for maintaining the integrity and proper understanding of the Quran, ensuring that the message and teachings of Islam contained therein are interpreted correctly (Sya'bani & Has, 2023). Likewise, thematic understandings are not accessible within current AI systems.

6. CONCLUSION

The Arabic language is not easily understood by AI due to its meaning across multiple interdependent linguistic layers, including orthography, morphology, syntax, and rhetoric, many of which cannot be recognized within a digital text framework. This study was conducted within a grounded theory framework using the Quranic corpus in Arabic and English (Sahih International), in which Arabic's linguistic properties contribute to near-perfect performance on structured tasks and limitations in Arabic thematic recall compared to English. These limitations are essential for understanding and improving AI performance while maintaining ethical and scholarly integrity in Quranic applications. Implications for educational use ultimately infer that only grounded modes in which uploaded Quranic corpus are used will maintain reliability in structured tasks; however, they should not be relied upon for thematic *tafsīr* analysis in Arabic, as the limitations of computational understanding are not structured to support the humanistic practice and tradition of Quranic recall and analysis. As such, this study proves to remind within Islamic ethics from the Qur'an in *sūrah Al-Ḥujurāt*, "O you who believe! If a *fāsiq* (sinful person) comes to you with

news, verify it, lest you harm people in ignorance and then become regretful for what you have done" (Qur'an, 49:6), of the misinformation that is relevant even among technological tools, such as AI, in which ethical considerations and education must be made.

REFERENCES

Alqahtani, M., Atwell, E. (2016). Arabic Quranic Search Tool Based on Ontology. In: Métais, E., Meziane, F., Saraee, M., Sugumaran, V., Vadera, S. (eds) Natural Language Processing and Information Systems. NLDB 2016. Lecture Notes in Computer Science(), vol 9612. Springer, Cham. https://doi.org/10.1007/978-3-319-41754-7_52

Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. https://doi.org/10.48550/arXiv.2003.00104

Atwell, E., Brierly, C., Dukes, K., (ﷺ)alha, M., Sharaf, A. (2011). An artificial intelligence approach to Arabic and Islamic content on the Internet. In *Proceedings of NITS 3rd National Information Technology Symposium*.https://doi.org/10.13140/2.1.2425.9528

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*, pp. 587–604. https://doi.org/10.1162/tacl_a_00041

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5185–5198, Online. Association for Computational Linguistics. 10.18653/v1/2020.acl-main.463

Bhojani, A. & Schwarting, M. (2023). Truth and Regret: Large Language Models, the Quran, and Misinformation.

Theology and Science, *21*(4), 557-563, https://doi.org/10.1080/14746700.2023.2255944

Diab, M. T. (2009). Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chinking. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*. http://www.elda.org/medar-conference/pdf/56.pdf

Dukes, K., & Habash, N. (2010). Morphological annotation of the Qur'an. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pp.2530-2536. http://www.lrec-conf.org/proceedings/lrec2010/pdf/276_Paper.pdf

Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing*, *8*(4), 1–22. https://doi.org/10.1145/1644879.1644881

Habash, N. (2010). Introduction to Arabic Natural Language Processing. Morgan & Claypool.

Hemmatian, B. (2023). Muslim-Violence Bias Persists in Debiased GPT Models. ArXiv (Cornell University). https://doi.org/10.48550/ARXIV.2310.18368

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), Article No. 248. https://doi.org/10.1145/3571730

Kidwai, A. R. (2011). Translating the Untranslatable: A Critical Guide to 60 English Translations of the Quran. Sarup Book Publishers Pvt. Ltd.

Maamouri, M., Bies, A., & Buckwalter, T. (2004). The Penn Arabic Treebank: Building a large-scale annotated Arabic corpus. *Proceedings of NEMLAR*.

Osman, R. A. H., & Hassan, M. I. A. (2022). Keistimewaan bahasa Arab sebagai bahasa al-Qurandan kepentingan menguasainya bagi para mufassirīn: The privilege of Arabic as the language of the Qur'an and the importance of mastering it for the mufassirīn. *Al-Hikmah: International Journal of Islamic Studies and Human Sciences*, *5*(2), 325–342. https://doi.org/10.46722/hikmah.v5i2.260

Selim, N. (2018). Arabic, Grammar and Teaching: An Islamic Historical Perspective. *International Journal of Islamic Thought*, *13*(1), 80–89. https://doi.org/10.24035/ijit.13.2018.008

Sharaf, A., & Atwell, E. (2012). QurSim: A corpus for evaluation of similarity in short texts. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 2295-2302. http://www.lrec-conf.org/proceedings/lrec2012/pdf/190_Paper.pdf

Sya'bani, M. Z., & Has, Q. A. B. (2023). Relevansi Bahasa Arab dalam Dakwah: Refleksi atasKedudukan Bahasa Arab sebagai Bahasa Al-Quran (Tinjauan Literatur). *Ath-Thariq: JurnalDakwah Dan Komunikasi*, *7*(1), 97–111. https://doi.org/10.32332/ath-thariq.v7i1.6532

The Qur'an (M.A.S. Abdel Haleen, Trans.). (2008). Oxford University Press.

Tanzil Project. (n.d.). Qur'an Texts & Translations (Arabic; Sahih International English). Retrieved from tanzil.net.

Zerrouki T., & Balla, A. (2017). Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems.*Data in brief*,*11*, 147–151. https://doi.org/10.1016/j.dib.2017.01.011.