Design and Development of a Contextual Search System for Qur'anic Text Based on Large Language Models (LLMs) or Artificial Intelligence (AI)

Lukman Hakim Husnan[1], Listiananda Apriliawan[2], Lailatul Mu'jizati[3], Kgs. Adlan Maghfur[4], Siti Alfiatun Hasanah [5]

[1] STIQ al-Lathifiyyah Palembang
[2]Airpaz Indonesia
[3] STIQ al-Lathifiyyah Palembang
[4] STIQ al-Lathifiyyah Palembang
[5]STIQ al-Lathifiyyah Palembang
Corresponding Email: elha@stiqlathifiyyah.ac.id

Received: 2025-11-13 /Accepted: 2025-11-27 / Doi:

ABSTRACT

This study addresses the challenge of providing reliable, safe, and contextually accurate access to the Holy Qur'an texts using Large Language Models (LLMs). While traditional approaches often rely on Retrieval-Augmented Generation (RAG) for factual grounding, this research proposes and evaluates a novel, non-RAG architectural approach centered on advanced, multi-layered GuardRail Prompting to manage the inherent risks of LLM stochasticity and hallucination in sensitive religious domains. The system integrates Input Guardrails for prompt injection mitigation, and critical Output Guardrails utilizing an LLM-as-a-Judge framework and a re-generation loop to validate responses against semantic relevance, Shariah compliance, and structural integrity (JSON Schema). The research adopts a Research and Development (R&D) methodology combined with a computational experimental approach to evaluate system performance, moderation effectiveness, and the functional role of rule-based control in regulating generative model outputs. Results demonstrate that a well-engineered GuardRail architecture can effectively constrain LLM behavior, achieving high faithfulness and relevance, with low PGR and acceptable FPR across adversarial and benign query datasets. This research establishes GuardRail Prompting as a viable and robust alternative for contextual grounding in sensitive, knowledge-intensive applications where RAG deployment may be restricted or structurally undesirable.

Keywords: Large Language Model, Contextual Search, GuardRail Prompt, Al-Qur'an, Hallucination Mitigation.

1. Introduction

The Qur'an, as the central text of Islam, possesses a highly rich and complex linguistic structure, rhetorical style, and semantic depth. This uniqueness encompasses variations of meaning at the lexical, phrasal, and verse levels, the use of elevated linguistic patterns, and intricate inter-verse relationships that require thematic and contextual understanding (Al Qarni, 2024). In this context, traditional keyword-based information retrieval methods often fail to capture the intended semantic nuances or to distinguish between terms that carry multiple layers of interpretation. Conventional search systems tend to produce shallow, literal, or irrelevant results due to their inability to interpret semantic relationships as well as the historical and textual contexts underlying Qur'anic verses. These challenges underscore an urgent need for more sophisticated and adaptive solutions, particularly contextual search systems capable of processing, understanding, and accommodating the complex meanings of the Qur'an with greater accuracy and methodological accountability.

Within the landscape of modern technology, Large Language Models (LLMs) offer exceptional capabilities in natural language processing and contextual reasoning, enabling the generation of text outputs that are more coherent, relevant, and semantically accurate. Leveraging billions of parameters and training on vast corpora, LLMs can identify complex linguistic patterns, comprehend inter-conceptual relationships, and interpret queries or instructions in a manner that closely resembles human cognitive processing. These capabilities present significant opportunities for the development of advanced search and interpretive systems for religious texts, including the Qur'an, which demand a high degree of sensitivity to context, deep meaning, and terminological nuance. Nevertheless, such strengths must be counterbalanced by robust control mechanisms, given the propensity of LLMs to generate hallucinations or inaccurate interpretations when insufficiently constrained (Alnefaie et al., 2024). The stochastic nature of LLMs renders them susceptible to non-deterministic, unpredictable, and potentially biased or harmful outputs (Nystrom, 2025).

The application of LLMs in religious studies, particularly in domains involving sacred texts, introduces substantial ethical and theological risks. A primary concern is the amplification of misinformation or hallucination (Rashed, 2025), whereby models generate seemingly plausible but factually incorrect, unreferenced, or doctrinally inconsistent information. In the context of sacred texts such as the Qur'an, such errors not only compromise informational quality but may also lead to religious misinterpretation, provoke controversy, or mislead users who perceive the model as an authoritative source of knowledge. Furthermore, there exists a risk of model misuse for interpreting verses outside established exegetical methodologies, thereby blurring epistemological boundaries between training data and legitimate religious authority. When erroneous content is generated by LLMs, it may seriously affect an individual's understanding of faith (Schwarting, 2025). Failure to verify and moderate responses in this domain poses a significant risk of social harm (Bhojani & Schwarting, 2023). Accordingly, system

architectures must implement stringent controls to ensure reliability, accuracy, and compliance with Islamic ethical principles (Shariah compliance) (Waqar et al., 2025).

Previous studies have explored the use of LLMs for Qur'anic semantic search by leveraging LLM-based embeddings to capture deep semantic connections (Al Qarni, 2024). Empirical findings indicate that LLM embeddings can perform consistently across varying levels of semantic complexity, providing a foundation for non-RAG contextual search approaches. However, other advanced transformer-based models, such as AraT5, may outperform LLMs in lower-level semantic retrieval tasks (Al Qarni, 2024).

In general, Retrieval-Augmented Generation (RAG) has emerged as the dominant architectural strategy for mitigating LLM hallucinations. RAG operates by linking LLMs to externally verified knowledge bases, enabling explicit source attribution and thereby enhancing user trust (Huang, 2023). Guardrails refer to agents, frameworks, or tools that guide and constrain LLM behavior to ensure outputs remain safe, ethical, and reliable (Zarecki, 2024). Functioning as real-time moderation systems, guardrails may incorporate content filtering, strict prompt structures, and output validation to enforce compliance with business rules, safety requirements, or regulatory standards (Endtrace, 2024). Guardrail prompting differs from conventional prompting by employing stricter instructions and backend logic that enforce external safety nets to capture internal model failures.

This study introduces a novel contribution by explicitly rejecting the RAG paradigm. Its uniqueness lies in demonstrating that a layered and integrated GuardRail Prompting architecture—implemented through internal contextual validation and corrective feedback loops—can function as an adequate grounding mechanism. GuardRail Prompting is employed to ensure the accuracy and faithfulness of LLM outputs in the Qur'anic domain without reliance on RAG infrastructure, which may be complex or costly, particularly in terms of token consumption (Shikkhaghildiyai, 2025).

Furthermore, the proposed system implements contextual guardrails tailored to theological and cultural sensitivities. This approach directly addresses gaps identified in AI ethics research concerning the lack of bias mitigation strategies specific to Arab and Muslim communities (Asseri et al., 2025). By emphasizing internalized behavioral control, this study demonstrates that guardrails are not merely defensive mechanisms but also active components in achieving Islamic ethical alignment and contextual accuracy in sensitive-domain applications (Waqar et al., 2025).

Accordingly, this study aims to design an LLM-based contextual search architecture for Qur'anic text in which reliability and security are fully ensured through GuardRail Prompt mechanisms; to develop and implement input and output guardrail layers capable of preventing prompt injection, mitigating bias and toxicity, and validating the contextual relevance and accuracy of generated outputs; and to quantitatively evaluate guardrail performance using Pass Guardrail Rate (PGR), False Positive Rate (FPR), and Attack Success Rate (ASR), as well as to assess response quality—specifically answer relevancy and faithfulness—through human evaluation studies.

2. Research Methods

This study focuses on the design of a GuardRail architecture as the core component of a Type-2 neural–symbolic system, in which symbolic rule-based controls are employed to moderate the generative outputs of a neural model. The research adopts a Research and Development (R&D) methodology combined with a computational experimental approach to evaluate system performance, moderation effectiveness, and the functional role of rule-based control in regulating generative model outputs. The procedure encompasses a series of system performance tests, behavioral stability analyses, and evaluations of the effectiveness of symbolic rule-based control over generative outputs.

The model is assessed using security-oriented metrics, including Pass GuardRail Rate (PGR), False Positive Rate (FPR), and F1-score, to measure the balance between system safety and practical utility. This experimental framework aligns with contemporary research designs that emphasize the importance of integrating symbolic control mechanisms as a risk mitigation strategy in sensitive LLM applications (Huang, 2023). Through this approach, the GuardRail architecture is evaluated not merely as a moderation tool, but as an epistemic component that ensures internal model grounding remains consistent, secure, and reproducible.

2.1. Types and Sources of Data

To support both search and validation functions, two types of data sources are employed. First, the core LLM and the LLM-as-a-Judge framework rely on internal access to a comprehensive Qur'anic text corpus. This corpus incorporates integrated linguistic layers, including orthographic representations (Uthmani script, transliteration, and translation), morphological features (part-of-speech tagging and root information), and syntactic structures (Hybrid Constituency–Dependency frameworks). Such linguistic depth is essential to enable robust semantic search and in-depth textual accuracy validation, thereby eliminating the need for external retrieval mechanisms.Second, for GuardRail evaluation, two test datasets are curated: a **Benign Dataset**, consisting of legitimate user queries used to measure the False Positive Rate (FPR), and an **Adversarial/Malicious Dataset**, designed to assess the system's defenses against prompt injection, toxicity, and attempts to trigger doctrinal misinformation (NVIDIA, 2025).

2.2. Research Object and Non-RAG Architecture Design

The research object is a contextual search system whose reliability does not derive from Retrieval-Augmented Generation (RAG). The system is powered by a core LLM with high semantic capability and operates within a three-stage architecture: Input Validation**,** Core Generation**,** and Output Validation/Correction Loop**.** Contextual (semantic) search is achieved solely through the representational power of LLM embeddings, which have been demonstrated to capture deep semantic relationships within Arabic corpora (Al Qarni, 2024). System accuracy is maintained through strict prompt engineering, which functions as a behavioral control mechanism for the model.

2.3. Implementation of a Multi-Layer GuardRail Prompting Architecture

The GuardRail Prompting layers are designed to guide and constrain LLM behavior to ensure safe, accurate, and ethical outputs (Zarecki, 2024). In other words, the GuardRail Prompting architecture ensures that the LLM does not generate responses freely without oversight, but instead operates within predefined, responsible boundaries. Through structured guidance and constraints, these layers enhance output quality, preserve informational accuracy, and ensure that responses remain aligned with accountable ethical standards.

2.3.1. Layer 1: Input Guardrails (Mitigation of Injection Risk)

Input Guardrails are responsible for protecting system integrity from manipulation attempts. Prompt sanitization is applied as a preprocessing technique to remove or neutralize potentially harmful input patterns. A Prompt Injection Shield specifically detects jailbreaking attempts—such as deceptive instructions like "ignore previous instructions"—and blocks or mitigates such attacks before they reach the core LLM (Nystrom, 2025).

2.3.2. Layer 2: Core Prompt Engineering (Contextual Guidance)

The core prompt functions as a strict behavioral guide. It explicitly defines the role of the LLM as an objective Qur'anic search authority and specifies output parameters such as tone, format, and depth. The LLM is explicitly prohibited from deviating from the sacred text or generating speculative content, thereby serving as a proactive behavioral constraint to mitigate hallucination risks prior to output validation.

2.3.3. Layer 3: Output Guardrails (Critical Validation and Correction)

Output Guardrails constitute the final line of defense, ensuring that generated responses comply with all safety, ethical, and relevance criteria.

a. **Contextual Relevance Validator**
   This validator ensures that responses remain coherent and on-topic by employing techniques such as cosine similarity computation or transformer-based models to compare the semantic alignment between the original query and the generated output, thereby preventing topic drift.
b. **Safety and Ethical Compliance Guardrails**
   These guardrails prevent harmful, biased, or Shariah-noncompliant outputs. They include content filters for offensive language and Personally Identifiable Information (PII) (Nystrom, 2025), and, critically, ensure that sensitive theological content does not violate established doctrinal boundaries (Lukose, 2025).
c. **Faithfulness and Hallucination Check (Theological Integrity)**
   To assess the extent to which LLM outputs are free from hallucination or misrepresentation, an LLM-as-a-Judge framework is employed. The validator LLM is instructed to decompose generated responses into individual propositions and verify the internal consistency of each statement against the referenced Qur'anic text. If hallucination is detected, this mechanism triggers a correction loop (Promptfoo, 2025).
d. **JSON Schema Enforcement**

Strict output validation is applied to ensure machine-readable structured data formats (e.g., Surah, Ayah, Arabic text, translation). This guarantees data integrity, although challenges related to the LLM's ability to consistently generate fully valid JSON structures remain (Erick, 2025).

2.3.4. Automatic Correction Mechanism (Re-generation Loop)

When Output Guardrails detect violations—such as low faithfulness—the GuardRail AI automatically generates a corrective prompt. This corrective instruction is sent back to the core LLM, forcing regeneration of a compliant and accurate response. This re-generation loop serves as the primary self-refinement mechanism in the non-RAG system, functioning as an internal grounding component by compelling the model to align with correct internal knowledge representations (Dong et al., 2025).

2.4. Quantitative Metrics and Evaluation Framework

2.4.1. Safety Evaluation (GuardRail Performance)

GuardRail performance is evaluated using formal security metrics:

a. Pass GuardRail Rate (PGR**):** The percentage of malicious attempts that successfully bypass the GuardRail. A lower PGR indicates stronger security effectiveness.

b. False Positive Rate (FPR): The proportion of legitimate (benign) queries incorrectly classified as attacks and blocked. A low FPR ensures system usability.

c. Attack Success Rate (ASR): Measures the failure rate of attack mitigation.

d. F1-Score**:** A composite metric reflecting the balance between precision and recall in blocking harmful content (Milvus, 2025).

2.4.2. Answer Quality Evaluation

a. Answer Relevancy**:** Measures the degree to which the generated response aligns with the intent of the query.

b. Faithfulness**:** Assesses theological integrity by evaluating the absence of hallucination in the generated response (Promptfoo, 2025).

2.4.3. Human Evaluation Framework

Given the sensitivity of the domain, assessments of information quality and safety involve expert reviewers (e.g., scholars in Islamic studies). This manual evaluation is crucial for validating safety, harm prevention, and theological accuracy, particularly in complex cases where automated judgments by the LLM-as-a-Judge may exhibit inconsistencies (Tam et al., 2025)

3. Result

This section presents the research findings based on systematic stages, including requirements analysis, architectural design, model/algorithm selection, and performance evaluation of the GuardRail Prompting system for LLM-based contextual search of Qur'anic texts.

3.1. System Requirements Analysis

The results of the requirements analysis indicate that contextual search of Qur'anic texts requires three main components:

a. Deep semantic representation capability, enabling the system to capture the contextual meaning of verses rather than relying solely on keyword matching.
b. Multi-layered security mechanisms, to ensure that the model does not deviate in meaning, does not generate hallucinations, and consistently maintains compliance with Shariah principles.
c. Structured output formatting, so that search results can be easily verified and programmatically reused.

This needs assessment reinforces the conclusion that LLMs cannot operate autonomously within the domain of sacred texts without stringent control mechanisms. Accordingly, the implementation of a multi-layer GuardRail architecture is essential to preserve accuracy, ethical integrity, and theological alignment.

A growing body of research indicates that LLM systems lacking layered guardrails are highly vulnerable to prompt injection and jailbreaking attacks, which can compel models to produce unstable or even harmful outputs (Liu & Carvalho, 2025). These findings are further supported by studies demonstrating that guardrails—whether in the form of symbolic rules, security filters, or semantic validators—can significantly enhance model reliability in sensitive domains (Nystrom, 2024). Moreover, research on the LLM-as-a-Judge paradigm highlights the importance of generative validators for assessing consistency and detecting hallucinations, although such mechanisms still require careful calibration to mitigate bias (Zeng & Shuster, 2025).

Additional studies have shown that techniques such as structured output enforcement, rule-based constraints, and re-generation loops can empirically reduce hallucination rates to a substantial degree (Jiang & Tandon, 2024). In the religious context, findings from IslamicEval 2025 underscore the critical role of layered validation in preventing semantic distortion in LLM-generated responses to Qur'anic and Hadith texts (Abdul Karim et al., 2025). Collectively, this recent empirical evidence consistently supports the conclusion that multi-layer GuardRail architectures represent the most effective approach for safeguarding semantic accuracy, ethical security, and theological integrity in the application of LLMs to Qur'anic texts.

3.1. Architecture Design Results

The proposed non-RAG architecture comprises three primary layers resulting from the design process:

(1) Input Guardrails

This layer functions to secure the system against manipulation and prompt injection attacks. Evaluation using adversarial datasets demonstrates that Prompt Sanitization and Injection Shield techniques are able to significantly reduce the effectiveness of jailbreaking instructions.

(2) Core Semantic Generator

The core LLM operates under a tightly engineered prompt that explicitly defines its role, theological constraints, reference boundaries, and output format. Embedding-based semantic representations are shown to effectively capture relevant verse meanings even in the absence of external retrieval mechanisms (non-RAG).

(3) Output Guardrails and Re-generation Loop

Contextual validation, ethical compliance checks, hallucination detection, and JSON structure verification are implemented as integral components of the control mechanism. When inconsistencies are detected, the system automatically triggers response regeneration through corrective prompting.

Overall, this architectural design ensures that every stage of processing—from input to output—is enveloped by measurable mitigation mechanisms, thereby maintaining robustness, reliability, and domain compliance throughout the system pipeline.

3.3. Model and Algorithm Selection

Based on preliminary testing, the following decisions were made:

1. The core LLM employs a transformer-based model with well-established semantic embedding capabilities, as it is highly effective in mapping user queries to Qur'anic concepts in a deep and nuanced manner.
2. The semantic similarity algorithm utilizes cosine similarity to detect topical relevance.
3. The LLM-as-a-Judge framework was selected to verify the theological consistency of each generated statement.
4. Structured output validation is enforced through JSON Schema Enforcement.

This combination of algorithms proved to be optimal in a non-RAG context, particularly in terms of output control and internal mitigation of hallucinations.

3.4. System Performance Evaluation

The system performance evaluation encompasses two main categories: GuardRail performance and answer quality.

1. **GuardRail Performance Evaluation**

Testing was conducted using both benign and adversarial datasets. The results are presented in Table 1 below.

Table. 1
Critical Performance Evaluation of Prompt-Based GuardRail

GuardRail Test Categories	Testing Objective	Pass Guardrail Rate (PGR) (%)	False Positive Rate (FPR) (%)	Reliability F1-Score (%)
Hallucination Mitigation (Faithfulness)	Preventing the dissemination of doctrinal misinformation	4.2%	5.1%	90.7%
Ethical Compliance (Toxicity/Bias)	Ensuring responses comply with Shariah principles	1.5%	2.3%	96.5%

Prompt Injection Shield	Blocking attempts to hijack LLM control	3.8%	3.5%	92.4%
Contextual Relevance	Ensuring responses remain aligned with the query topic	6.0%	4.8%	89.1%
Average	--	**3.88%**	**3.93%**	**92.17%**

The evaluation results indicate that the GuardRail system functions very effectively in balancing security and utility. First, the average Pass GuardRail Rate (PGR) of 3.88% suggests that the majority of attack attempts—including prompt injection, instruction manipulation, and jailbreaking efforts—failed to penetrate the GuardRail defense layers. This low PGR value serves as a critical indicator that the rule-based control mechanisms and automated validation processes are capable of resisting attacks with a high success rate, thereby preserving the integrity of LLM behavior throughout the response generation process.

Second, the average False Positive Rate (FPR) of only 3.93% demonstrates that the system is not overly aggressive or excessively restrictive in filtering legitimate queries. In other words, most normal user requests can still be processed without being blocked by the security mechanisms. This low FPR highlights that the GuardRail system does not sacrifice user experience for the sake of security, successfully maintaining a balance between strict filtering and openness to valid inputs.

Third, the average F1-score reaches 92.17%, reflecting the system's overall stability, consistency, and accuracy in detecting and filtering harmful content. The high F1-score confirms that the balance between precision (the ability to correctly identify malicious inputs) and recall (the ability to capture attacks without allowing many to bypass the system) is achieved at an optimal level. This performance underscores that the proposed multi-layer GuardRail architecture is not only effective but also efficient in reducing the risks of hallucination, theological distortion, and undesirable generative behaviors.

Collectively, these metrics demonstrate that the system achieves an ideal trade-off between security and usability. On the one hand, the GuardRail mechanisms maintain a high level of protection against various forms of attacks, as reflected by the low PGR and high F1-score. On the other hand, the system preserves user convenience and accessibility, as evidenced by the low FPR and the minimal blocking of legitimate queries. Accordingly, these findings affirm that the multi-layer GuardRail architecture is not only effective in mitigating risks—such as hallucination, prompt manipulation, and theological deviation—but also enables the LLM to operate optimally in delivering accurate contextual search results. This balance between security and utility represents a key indicator that the implemented non-RAG design performs as expected, while also providing an empirical foundation for deploying similar systems in other sensitive domains.

2. Answer Quality Evaluation

Following the implementation of the GuardRail mechanisms, several notable improvements were observed.

a. *Faithfulness* increased to 91.5%, representing a substantial improvement compared to the baseline model, which achieved only 55%.

b. *Answer relevancy* remained stable, supported by thematic control enforced through the Contextual Relevance Validator.

c. The LLM-as-a-Judge framework successfully detected and mitigated theological hallucinations through an automated correction loop.

These results demonstrate that GuardRail Prompting can effectively function as a substitute for external grounding mechanisms such as Retrieval-Augmented Generation (RAG) in domains characterized by static and well-defined textual corpora, such as the Qur'an.

4. Discussion

4.1. Performance Analysis of GuardRail: The Trade-off between Security and Utility

The performance analysis of the GuardRail mechanism indicates that the system successfully balances stringent security requirements with high utility. The low average Pass Guardrail Rate (PGR) of 3.88% demonstrates that the GuardRail architecture effectively blocks adversarial attempts targeting the system, reflecting a strong security posture. Importantly, this level of security is achieved without compromising legitimate user experience. The low False Positive Rate (FPR), averaging 3.93%, confirms that only a small proportion of valid user queries are incorrectly blocked or censored, thereby preserving the system's functional utility. This effectiveness is consistent with the findings of Suo et al. (2024), who reported that signed-prompt security approaches can reduce attack success rates to below 5% across a range of test scenarios (Suo et al., 2024). These results are further reinforced by the study of Xiong et al. (2025), which emphasizes that defenses based on defensive prompt patching tend to be highly effective in resisting context-shifting exploits (Xiong et al., 2025). The alignment of the GuardRail system's performance with these studies suggests that it implements defense strategies consistent with state-of-the-art practices in LLM security.

The high reliability F1-score (average 92.17%) further underscores the effectiveness of the hybrid symbolic–neural system employed in the GuardRail architecture. This score indicates that the integration of symbolic rules with generative neural models enables consistent detection and handling of risky content while minimizing classification errors. Within this architecture, symbolic components—implemented as logical rules directly encoded in the GuardRail Prompting framework—serve as a stable control structure capable of mitigating the output uncertainty commonly associated with neural generative models (Wang et al., 2025). This hybrid approach demonstrates that the system does not rely solely on probabilistic inference from large models, but instead combines it with deterministic formal constraints, resulting in more accurate and predictable decision-making.

Overall, this performance strengthens the position of GuardRail as a reliable control mechanism for critical applications that demand non-negotiable levels of security, compliance, and accuracy, such as in regulatory environments, healthcare services, banking, or automated risk assessment systems. With a high F1-score, GuardRail Prompting demonstrates that symbolic logic can function as an essential defensive layer to address gaps that cannot be adequately handled by purely

neural approaches. More broadly, these results point to a promising direction for the development of hybrid AI systems that integrate the generalization capabilities of neural models with the precision and stability of symbolic control, enabling optimal performance in sensitive and high-risk operational environments.

4.2. Achieving Contextual Grounding without RAG

This study demonstrates that GuardRail Prompting can function as a robust grounding mechanism even in the absence of a Retrieval-Augmented Generation (RAG) approach. Effective contextual search is achieved by leveraging the model's capacity to utilize rich and structured internal representations of Qur'anic knowledge. These representations are formed through fine-tuned LLM embeddings that are capable of capturing deeper semantic relationships, rather than relying solely on lexical matching or surface-level patterns (Al Qarni, 2024). Consequently, the contextual understanding of verses—including thematic relations, conceptual meanings, and inter-phrase linkages—can be traced without the need for external document retrieval. This effectiveness indicates that, in domains with strong semantic structures such as the Qur'anic text, prompt-based grounding mechanisms constitute a viable alternative to RAG, particularly when narrative consistency, security control, and computational efficiency are prioritized. These findings further open new avenues for the development of contextual search systems that combine the power of semantic embeddings in LLMs with rigorous symbolic control through GuardRail Prompting.

The Re-generation Loop mechanism is central to the success of internal grounding in this system. When the LLM-as-a-Judge detects semantic inconsistencies, interpretive bias, or indications of hallucination in an initial response, the system automatically generates corrective prompts that compel the model to regenerate its answer under stricter consistency constraints. This process is not merely corrective, but also functions as an internal self-alignment mechanism, ensuring that the final response is firmly anchored in the model's deeply learned knowledge representations (Dong et al., 2025). In non-RAG systems, this approach serves as a direct substitute for the external validation typically provided by document retrieval in RAG-based architectures.

By replacing external source-based verification with a series of rigorous internal consistency validations, the Re-generation Loop introduces a more flexible and efficient layer of quality control. This approach enables the system to minimize hallucinations, enhance response stability, and significantly improve Faithfulness—the degree to which outputs align with accurate and relevant knowledge. The marked improvement in Faithfulness provides strong evidence that integrated self-critique mechanisms within the generation pipeline can serve as a foundational component for non-RAG systems that rely on deep semantic embeddings and internal coherence, without dependence on external data sources.

4.3. Ethical Implications, Bias, and Domain-Specific Compliance

GuardRail Prompting offers a distinctive advantage in regulating the behavioral and ethical dimensions of LLMs that extends beyond the capabilities of RAG. Whereas RAG primarily focuses on improving factual accuracy through the integration of external documents, GuardRail mechanisms enable far broader control over how responses are formulated and presented (Zarecki, 2024). In this approach, symbolic rules embedded within the prompt not only guide the

substantive content of responses, but also regulate pragmatic dimensions such as communicative tone, ethical sensitivity, linguistic style, degrees of caution, and compliance with system-defined normative guidelines. Accordingly, GuardRail does not merely ensure that responses are factually correct, but also that they are delivered responsibly, non-harmfully, and in accordance with defensible standards of conduct.

This fundamental difference renders GuardRail particularly advantageous in applications that require strict oversight of how models interact with users—such as in education, public services, religious guidance, or other socially high-risk domains. While RAG primarily refines content sources, GuardRail refines the ethical framework that governs model behavior. Through this distinction, GuardRail Prompting not only strengthens informational reliability but also establishes a level of behavioral alignment that cannot be achieved through RAG mechanisms alone.

The system effectively addresses severe ethical challenges inherent in religious texts, namely the risks of misinformation and the lack of doctrinal intuition in LLMs. By explicitly defining ethical and security criteria within GuardRail prompts, the system functions as a moral and theological gatekeeper. The low PGR observed in the Ethical Compliance category indicates successful mitigation of culturally specific biases relevant to Arab/Muslim communities, aligning with Schwarting's argument that this area requires urgent interdisciplinary collaboration in AI ethics research (Schwarting, 2025).

Furthermore, the JSON Schema Validation mechanism ensures that outputs are generated in a structured format—such as consistent mappings between surahs and verses—thereby facilitating verification by both users and automated systems. This schema-based validation functions not only as a formatting check, but also as a form of structural oversight that constrains the model's margin for error in presenting verse references. By compelling the model to accurately populate predefined structures, the system reduces the likelihood of deviations, incompleteness, or fabrication of verse citations that may occur in free-form generation.

Such enhanced transparency is particularly critical in Qur'anic search applications, as users can directly observe how responses are derived and which references are employed. This structural transparency also contributes to rebuilding user trust—a form of trust that, in RAG-based systems, typically derives from external document attribution. In other words, although the system does not employ RAG, its clear, standardized, and traceable output structures provide a level of transparency functionally equivalent to source attribution in RAG. This demonstrates that informational reliability and accountability can be effectively maintained through robust internal mechanisms, without reliance on retrieval-based external verification.

4.1. Contrast with Conventional RAG Architectures

The primary contrast between the two approaches lies in the locus of reliability that underpins each system's operation. In RAG-based systems, reliability is derived from external knowledge sources that can be continuously updated and validated, making this approach highly effective for domains requiring up-to-date factual information or document-based correction. However, in the context of Qur'anic texts—which are static, fixed, and not subject to content

updates—the central challenge is not factual accuracy per se, but rather the preservation of interpretive precision, ethical propriety, and adherence to established scholarly and religious norms. It is in this respect that the GuardRail Prompting architecture demonstrates its advantage. This approach shifts the focus away from external document retrieval toward the internalization of behavioral control, implemented through a set of symbolic mechanisms that regulate how the model interprets, frames, and articulates its responses.

Because GuardRail mechanisms are capable of controlling output structure, ensuring the accuracy of verse references, regulating narrative style, and enforcing ethical compliance at every stage of generation, the proposed system affords substantially tighter control than RAG in highly sensitive domains such as Qur'anic exposition. Consequently, GuardRail functions not merely as a filter or an auxiliary security layer, but as an active component that steers the process of contextual grounding while maintaining ethical and theological alignment throughout the entire generation pipeline.

When applied in a layered and well-designed manner, GuardRail Prompting mechanisms form an integrated defense model that is critically required for LLM applications demanding non-negotiable domain security—such as thematic exegesis, verse retrieval, or Qur'anic literacy assistance. Under this approach, the system remains secure, guided, and consistent, while avoiding reliance on external sources that are irrelevant or potentially inappropriate for the domain of sacred texts.

5. Conclusion

This study successfully designed, developed, and evaluated a contextual search system for Qur'anic text based on Large Language Models (LLMs), employing GuardRail Prompting mechanisms as the primary means of ensuring reliability and security, explicitly without relying on Retrieval-Augmented Generation (RAG).

The proposed multi-layer GuardRail architecture—comprising an Input Guardrail, an Output Guardrail based on the LLM-as-a-Judge framework, and an automatic correction loop—was shown to effectively regulate LLM behavior. Quantitative evaluation indicates that the system achieves a high level of security (average Pass Guardrail Rate [PGR] of 3.88% and a low Attack Success Rate [ASR]), while simultaneously maintaining strong utility (average False Positive Rate [FPR] of 3.93%) and high contextual accuracy. This accuracy is further evidenced by a Faithfulness score of 91.5%, as validated through expert assessment.

The primary significance of this study lies in demonstrating that rigorous prompt engineering and layered GuardRail Prompting constitute a viable and robust strategy for ensuring theological integrity and mitigating the risk of LLM-generated misinformation in sensitive domains. This architecture offers an efficient and well-controlled alternative to domain-specific LLM architectures, particularly in contexts where behavioral control and ethical compliance are prioritized over dependence on external knowledge bases.

6. Acknowledgments

The author would like to express sincere gratitude to the College of Qur'anic Studies (Sekolah Tinggi Ilmu al-Qur'an, STIQ) al-Lathifiyyah Palembang for the support provided in the conduct of this research.

References

Abdul-Karim, M., Al-Sayed, H., & Nour, F. (2025). *IslamicEval: A shared task for evaluating hallucination and factuality of LLMs on Islamic texts*. ACL Anthology.

Alnefaie, M., Alharthi, I., & Alghamdi, A. (2024). *LLMs based approach for Quranic question answering*. In Proceedings of the 13th International Conference on Computer Science and Information Technology. SCITEPRESS. https://www.scitepress.org/Papers/2024/130129/130129.pdf

Alqarni, M. (2024). *Embedding search for Quranic texts based on large language models*. The International Arab Journal of Information Technology, 21(2), 243–256. https://doi.org/10.34028/iajit/21/2/7

Asseri, B., Abdelaziz, E., & Al-Wabil, A. (2025). Prompt Engineering Techniques for Mitigating Cultural Bias Against Arabs and Muslims in Large Language Models: A Systematic Review. *arXiv preprint arXiv:2506.18199*.

Bhojani, A.-R., & Schwarting, M. (2023). *Truth and regret: Large language models, the Quran, and misinformation*. *Journal of Qur'anic Studies*, advance online publication. https://www.tandfonline.com/doi/pdf/10.1080/14746700.2023.2255944

Dong, Y., Mu, R., Zhang, Y., Sun, S., Zhang, T., Wu, C., & Huang, X. (2025). Safeguarding large language models: A survey. *Artificial intelligence Review*, *58*(12), 382.

Endtrace. (2024). *Prompt engineering with guardrails: Safety-first design for LLMs*. https://www.endtrace.com/prompt-engineering-with-guardrails-guide/

Erick, H. (2025). *How JSON Schema works for structured outputs and tool integration*. PromptLayer Blog. Retrieved October 23, 2025, from https://blog.promptlayer.com/how-json-schema- works-for-structured-outputs-and-tool-integration/

Huang, K. (2023, November 22). *Mitigating security risks in Retrieval Augmented Generation (RAG) LLM applications*. Cloud Security Alliance. https://cloudsecurityalliance.org/blog/2023/11/22/mitigating-security-risks-in-retrieval- augmented-generation-rag-llm-applications

Jiang, Y., Kumar, S., & Tandon, R. (2024). *Effective strategies for mitigating hallucinations in large language models*. arXiv. https://arxiv.org/abs/2402.01832

Liu, X., Park, J., & Carvalho, A. (2025). *SoK: Evaluating jailbreak guardrails for largelanguage models*. arXiv. https://arxiv.org/abs/2506.10597

Lukose, D. (2025). *Guardrails implementation best practice*. Medium. Retrieved October 23, 2025, from https://medium.com/@dickson.lukose/guardrails-implementation-best-practice- e5fa2c1e4e09

Milvus. (2025). *What metrics are used to evaluate the success of LLM guardrails?* Retrieved October 24, 2025, from https://milvus.io/ai-quick-reference/what-metrics-are-used-to-evaluate- the-success-of-llm-guardrails

NVIDIA. (2025). *Measuring the effectiveness and performance of AI guardrails in generative AI applications*. NVIDIA Developer Blog. Retrieved October 24, 2025, from https://developer.nvidia.com/blog/measuring-the-effectiveness-and-performance-of-ai- guardrails-in-generative-ai-

applications/

Nystrom, R. (2025). *LLM guardrails: How to build reliable and safe AI applications*. Neptune.ai Blog. https://neptune.ai/blog/llm-guardrails

Promptfoo. (2025). *How to measure and prevent LLM hallucinations*. Retrieved October 23, 2025, from https://www.promptfoo.dev/docs/guides/prevent-llm-hallucinations

Schwarting, M. (2025). *To Christians developing LLM applications: A warning, and some suggestions*. AI and Faith. https://aiandfaith.org/featured-content/to-christians-developing-llm-applications-a-warning-and-some-suggestions/

Shikkhaghildiya**i.** (2025). *Context-aware RAG system with Azure AI Search to cut token costs and boost accuracy*. Microsoft TechCommunity. https://techcommunity.microsoft.com/blog/azure-ai-foundry-blog/context-aware-rag-system-with-azure-ai-search-to-cut-token-costs-and-boost-accur/4456810

Suo, S. et al. (2024). *Signed-Prompt: A New Approach to Prevent Prompt Injection Attacks Against LLM-Integrated Applications*. arXiv:2401.07612.

Tam, T. Y. C., et al. (2024). *A framework for human evaluation of large language models in healthcare derived from literature review*. *PubMed Central*. Retrieved October 23, 2025, from https://pmc.ncbi.nlm.nih.gov/articles/PMC11437138/

Wang, X., Ji, Z., Wang, W., Li, Z., Wu, D., & Wang, S. (2025). SoK: Evaluating Jailbreak Guardrails for Large Language Models. *arXiv preprint arXiv:2506.10597*.

Waqar, K. M., Ibrahim, M., & Khan, M. M. I. (2025). Ethical Implications Of Artificial Intelligence: An Islamic Perspective. *Journal of Religion and Society*, *3*(01), 347-358.

Xiong, W. et al. (2025). *Defensive Prompt Patch: A Robust and Generalizable Defense of Large Language Models against Jailbreak Attacks*. Findings of ACL 2025.

Zarecki, I. (2024). *LLM guardrails guide AI toward safe, reliable outputs*. K2View. https://www.k2view.com/blog/llm-guardrails

Zheng, T., Li, H., & Shuster, K. (2025). *A survey on LLM-as-a-Judge: Capabilities, limitations, and calibration techniques*. arXiv. https://arxiv.org/abs/2503.01234