

Investigating Retrieval-Augmented Generation in Quranic Studies: A Study of 13 Open-Source Large Language Models

Zahra Khalila¹, Arbi Haza Nasution², Winda Monika³,
Aytug Onan⁴, Yohei Murakami⁵, Yasir Bin Ismail Radi⁶, Noor Mohammad Osmani⁷
Department of Informatics Engineering, Universitas Islam Riau, Pekanbaru 28284, Indonesia^{1,2}
Department of Library Information, Universitas Lancang Kuning, Riau 28266, Indonesia³
Department of Computer Engineering-College of Engineering and Architecture,
Izmir Katip Celebi University, Izmir, 35620 Turkey⁴
Faculty of Information Science and Engineering, Ritsumeikan University,
Kusatsu, Shiga 525-8577, Japan⁵
Faculty of AI-Quran & Sunnah, Universiti Islam Antarabangsa Tuanku Syed Sirajuddin (UniSIRAJ),
Kuala Perlis, Perlis 02000, Malaysia⁶
Department of Qur'an And Sunnah Studies-Ahas Kirkhs,
International Islamic University Malaysia, Malaysia⁷

Abstract—Accurate and contextually faithful responses are critical when applying large language models (LLMs) to sensitive and domain-specific tasks, such as answering queries related to quranic studies. General-purpose LLMs often struggle with hallucinations, where generated responses deviate from authoritative sources, raising concerns about their reliability in religious contexts. This challenge highlights the need for systems that can integrate domain-specific knowledge while maintaining response accuracy, relevance, and faithfulness. In this study, we investigate 13 open-source LLMs categorized into large (e.g., Llama3:70b, Gemma2:27b, QwQ:32b), medium (e.g., Gemma2:9b, Llama3:8b), and small (e.g., Llama3.2:3b, Phi3:3.8b). A Retrieval-Augmented Generation (RAG) is used to make up for the problems that come with using separate models. This research utilizes a descriptive dataset of Quranic surahs including the meanings, historical context, and qualities of the 114 surahs, allowing the model to gather relevant knowledge before responding. The models are evaluated using three key metrics set by human evaluators: context relevance, answer faithfulness, and answer relevance. The findings reveal that large models consistently outperform smaller models in capturing query semantics and producing accurate, contextually grounded responses. The Llama3.2:3b model, even though it is considered small, does very well on faithfulness (4.619) and relevance (4.857), showing the promise of smaller architectures that have been well optimized. This article examines the trade-offs between model size, computational efficiency, and response quality while using LLMs in domain-specific applications.

Keywords—Large-language-models; retrieval-augmented generation; question answering; Quranic studies; Islamic teachings

I. INTRODUCTION

Natural language processing (NLP) has been transformed as a result of the development of large language models (LLMs), which made it possible for these models to handle a wide range of activities. These include summarization and translation, as well as answering domain-specific questions [1]. Further, these models can even serve as a good anno-

tator for a number of NLP tasks [2]. Recent studies have explored the use of NLP in various domain-specific including Quranic studies, legal system, and medical field focusing on developing question-answering system. Alnefie et al. (2023) [3] has evaluated the effectiveness of GPT-4 in answering Quran-related questions and highlighting challenges in context understanding and answer accuracy. However, their work is limited by its reliance on a general-purpose LLM without domain-specific fine-tuning, which affects response precision for nuance religious queries. Retrieval-Augmented Generation (RAG) has been successfully applied in general knowledge such as medical domains to mitigate these issues [4], and legal domain such as research conducted by Pipitone and Alami (2024) [5] introduced LegalBench-RAG to evaluate retrieval accuracy in legal question-answering tasks. While significant progress has been made across these domains, challenges related to data quality, retrieval precision, domain-specific adaptation, and computational efficiency persist. This study builds upon these works by benchmarking open-source LLMs using the RAG framework to address the challenges in Quranic knowledge retrieval [6] while highlighting the balance between model size, performance, and efficiency. Using LLMs in religious or culturally sensitive environments has certain challenges [7], including ensuring the accuracy, contextual relevance, and authenticity of the generated responses. These issues are particularly significant when engaging with content derived from religious texts, as distortion or hallucination may result in misunderstandings and a loss of faith in AI systems [8].

This paper examines the role of LLMs in quranic studies [9]. We use a dataset from a book about the 114 surahs of the Qur'an [10], rather than the Qur'an text itself. This dataset provides thorough information about each surah, including meaning, provenance of revelation, and historical context. Descriptive insights are essential for offering relevant and accurate answers to questions regarding quranic studies.

Since Qur'an is the sacred revelation that remains intact in today's world without any human involvement, and the Qur'an is the only source to link human beings with their creator, an attempt to identify the reliable and trustworthy LLMs is really important. As they provide useful resources for the readers, accuracy and truthfulness must be given due importance while exploring the provided information from such sources.

In order to overcome the difficulties associated with hallucination and accuracy, we have implemented a framework known as Retrieval-Augmented Generation (RAG) [11], [12], [13]. This method integrates LLM semantics with a vector database to retrieve relevant data from descriptive datasets [14], [15]. The RAG method guarantees that responses derive from authoritative sources that are contextually appropriate, thereby reducing the probability of delivering content that is unsubstantiated or irrelevant. Citations are supplied in each response, which enables users to trace the information back to the descriptive dataset. This further enhances the level of trust and transparency [16].

The objectives of this research are threefold: (1) to compare 13 open-source LLMs in terms of their ability to respond accurately and faithfully to questions about quranic studies [17], (2) to assess the effectiveness of the RAG approach in reducing hallucination and ensuring response relevance [18], and (3) to provide insights into the use of descriptive datasets for religious education and AI-based knowledge systems [19]. The evaluation criteria consist of context relevance, response faithfulness, and answer relevance, and they are evaluated using human evaluation. [14].

This study provides a robust framework for integrating LLMs with descriptive datasets, thereby contributing to the expanding field of domain-specific AI applications [20]. It emphasizes the strengths and limitations of current LLMs in managing sensitive religious topics and establishes a foundation for future developments in AI-driven educational and informational tools.

The rest of this paper is organized as follows: Section I introduces the challenges of using LLMs for Quranic studies and outlines the research objectives. Section II reviews related work, discussing previous studies on LLM applications in religious text analysis and RAG-based retrieval systems. Section III provides a detailed description of the experimental setup, covering the dataset, NLP tasks and evaluation guidelines, dataset selection and curation, human evaluators, metrics for quality evaluation, large language models, and hardware and software configuration. Section IV presents the experimental results of various LLM models. Section V discusses key findings, including performance insights based on model size, the effectiveness of the RAG framework, the trade-off between computational resources and response quality, the surprising performance of Llama3.2:3b, and implications for domain-specific tasks. Section VI concludes the paper by summarizing the main contributions and providing suggestions for future research.

II. MATERIALS AND METHODS

This section outlines the methodical strategy employed in our research. We begin by analyzing the dataset, detailing its source, structure, and descriptive content. Subsequently,

the system's responsibilities are thoroughly delineated, encompassing the formulation of solutions to user concerns concerning Islamic doctrines [21]. In addition, we provide a summary of the rules that have been set for human evaluators who are responsible for evaluating the outputs of the system. To ensure that the evaluations are reliable and consistent, we have produced these guidelines.

A. Datasets

The dataset used in this research comes from a descriptive book providing a thorough study of the 114 surahs (Chapters) of the Qur'an. In this preliminary study, 20 surahs were selected and tested out of the total 114 surahs. The dataset includes numerous descriptive elements for each surah, such as but not limited to:

- Number of Verses: The total number of verses in each surah.
- Meaning of its Name: A description of the surah's title and its importance.
- Reason for its Name: An explanation for the designation of the surah, frequently linked to its subject matter or motifs.
- Names: Alternative titles or names linked to the surah, if relevant.
- General Objective: A brief explanation of the primary message or objective of the surah.
- Reason for its Revelation: The circumstances or context in which the surah was revealed, as available.
- Virtues: Key benefits or spiritual rewards associated with reciting or understanding the surah, often supported by hadith (sayings of the Prophet Muhammad, peace be upon him).
- Relationships: Insights into the connections between the beginning and end of the surah, or its relationship to preceding or succeeding chapters.

For instance, Surah Al-Hadid (Chapter 57) is described as follows:

- Number of Verses: 29.
- Meaning of its Name: "Al-Hadid" translates to "The Iron" in Arabic.
- Reason for its Name: It is the only chapter where the benefits of iron are mentioned, symbolizing strength and utility.
- General Objective: Encourages the virtue of spending in the cause of Allah as an appreciation of His favors.
- Virtues: Includes a hadith where the Prophet Muhammad (peace be upon him) recommends reciting this chapter as one of three glorifications of Allah.
- Relationships: Highlights thematic continuity between its verses and connections to preceding chapters, such as Surah Al-Waqi'ah.

The dataset was preprocessed and organized into structured fields to ensure efficient retrieval and usability in the study. Key steps included:

- **Field Segmentation:** Each descriptive element (e.g., "virtues," "relationships") was extracted and stored as a separate field for better semantic alignment with queries.
- **Vectorization:** Text data was transformed into high-dimensional vector embeddings through the use of cutting-edge NLP models, which facilitated the search for semantic similarity [22].
- **Storage in a Vector Database:** The organized and scalar data was kept in a scaled vector database so that it would be easy to find the right descriptions during query processing [23],[22].

This particular dataset provides lots of information that goes beyond the actual text of the Qur'an, including contextual and interpretative details. Based on this, it is a great instrument for evaluating the ability of LLMs to produce responses that are true, accurate, and relevant to the context in which they are being used. By emphasizing descriptive elements, the dataset ensures that responses align with established interpretations and scholarly perspectives.

B. NLP Tasks and Evaluation Guidelines

A Retrieval-Augmented Generation (RAG) architecture will be utilized in order to accomplish the objective of this study, which is to evaluate large language models (LLMs) in the context of answering questions related to quranic studies [24]. The RAG approach combines LLMs with semantic retrieval to provide contextually relevant and authoritative responses from a descriptive dataset. The primary tasks and evaluation guidelines used to assess the system's performance are outlined in full below [14].

1) *NLP Tasks:* The research's primary NLP task is to generate semantically pertinent and contextually accurate responses to inquiries regarding quranic studies. The system employs a Retrieval-Augmented Generation (RAG) architecture, combining retrieval-based and generative methodologies, to ensure that responses are both dataset-based and linguistically coherent. The system executes the following tasks:

- **Semantic Search and Retrieval:** Upon a user's query submission, the system does a semantic similarity search over the vectorized dataset obtained from Qur'anic surah descriptions [19]. This procedure determines the most contextually pertinent entries from the dataset to respond to the query.
- **Response Generation:** The retrieved descriptions are submitted to the LLMs, which produce a comprehensive response [14]. This response integrates the retrieved information and provides explanatory content to address the query.
- **Citations and Contextualization:** Each generated response includes references to the original dataset entries (e.g., surah descriptions or specific virtues), allowing users to trace the information back to its source [25].

2) *Evaluation guidelines:* To assess the quality of the responses generated by the system, human evaluators followed a structured set of evaluation guidelines. These guidelines provided a consistent framework for scoring responses across three key dimensions: Context Relevance, Answer Faithfulness, and Answer Relevance [14]. Each dimension is explained below, along with its calculation method and examples.

- **Context Relevance** evaluates how precisely the retrieved and generated responses align with the user query while avoiding irrelevant or extraneous information. The relevance score is calculated using the precision@k metric, where k represents the number of top retrieved results considered as shown in Equation 1.

$$\text{Precision@k} = \frac{\text{No. of relevant results in the top-k responses}}{k} \quad (1)$$

Example:

- **Query:** "What is the reason for Surah Al-Fatihah being named Umm Al-Kitab?"
- **Retrieved Information:**
 - 1) Surah Al-Fatihah is named Umm Al-Kitab because it summarizes the essence of the Qur'an (relevant).
 - 2) It is recited in every unit of prayer (relevant).
 - 3) Surah Al-Baqarah discusses laws and stories (irrelevant).
 - 4) Surah Al-Fatihah has seven verses (relevant).
 - 5) Surah An-Nas is the last chapter of the Qur'an (irrelevant).

If $k = 5$, then 3 out of the 5 retrieved results are relevant:

$$\text{Precision@5} = \frac{3}{5} = 0.6$$

The context relevance score for this response is therefore 0.6.

- **Answer Faithfulness** ensures that the generated responses accurately represent the retrieved information without introducing unsupported content or hallucinations. Evaluators compare the generated response with the dataset to verify factual consistency.

Example:

- **Query:** "What does Surah Al-Fatihah emphasize?"
- **Retrieved Information:** Surah Al-Fatihah emphasizes monotheism, gratitude, and seeking guidance from Allah.
- **Faithful Response:** Surah Al-Fatihah highlights the themes of monotheism, gratitude, and the importance of seeking Allah's guidance.
- **Non-Faithful Response:** Surah Al-Fatihah emphasizes the stories of past prophets.

The faithful response adheres strictly to the retrieved information, while the non-faithful response introduces unsupported content.

- **Answer Relevance** measures whether the response directly addresses the query while maintaining se-

semantic and theological appropriateness. It assesses completeness, clarity, and alignment with the question. Example:

- Query: “Why is Surah Al-Fatihah called Umm Al-Kitab?”
- Relevant Response: Surah Al-Fatihah is called Umm Al-Kitab because it summarizes the central teachings of the Qur’an and is recited in every unit of prayer.
- Irrelevant Response: Surah Al-Fatihah has seven verses and is the first chapter of the Qur’an.

The relevant response directly answers the query, providing reasoning, while the irrelevant response, though factually correct, fails to address the specific question.

3) *Evaluation process:* Human evaluators assessed the system-generated responses through a web-based platform, where they were presented with prompts, the corresponding responses, and an interface for scoring. The evaluation process included the following steps:

- **Reviewing Responses:** The process of reviewing responses was a critical step in evaluating the quality of the outputs generated by the large language models (LLMs). Human evaluators carried out this task through a web-based platform specifically designed to facilitate structured and unbiased assessments.
- **Scoring System:** For each response, evaluators assigned scores on a Likert scale (1 to 5) for the three evaluation criteria: Context Relevance, Answer Faithfulness, and Answer Relevance. In addition to numerical scores, evaluators could provide written feedback to justify their evaluations [16]. This qualitative feedback emphasized specific faults or applauded features of the response, providing more insight into the system’s performance.
- **Reevaluation and Calibration:** Since numerous replies were provided for each query, evaluators could compare the quality of outputs from various LLMs. This comparative approach was instrumental in identifying the models’ relative strengths and limitations, thereby enabling a more thorough evaluation [26]. To verify the dependability of their judgments, evaluators returned to a subset of previously evaluated responses on a regular basis and reassessed them. This consistency check allowed evaluators to reflect on their scoring processes and ensure they were in line with the rating criteria.

The structured evaluation guidelines guaranteed that the assessment process was meticulous, consistent, and transparent. The guidelines established a comprehensive framework for assessing the system’s performance by emphasizing context relevance, answer faithfulness, and answer relevance [27]. This method enable a thorough comparison of several LLMs and provided vital insights into their performance in answering Islamic queries with contextual precision and faithfulness [15], [27]. The evaluations were submitted through the platform after all responses to a specific query were evaluated, scored,

and commented on. In order to provide a comprehensive dataset for the purpose of investigating the LLMs, the platform logged and stored the data for research [28].

C. Dataset Selection and Curation

The dataset used in this study was carefully selected and organized to guarantee it aligns with the goals of assessing large language models (LLMs) within the framework of quranic studies. The process of selection and curation included identifying a reliable source, organizing the data, and confirming its alignment with the research goals [15], [28].

1) *Selection criteria:* The dataset was chosen according to these specific criteria:

- **Authenticity:** The source underwent a thorough review to confirm its compliance with recognized Islamic scholarship and the absence of speculative interpretations [26], [27].
- **Descriptive Richness:** The dataset must deliver comprehensive, contextually rich descriptions that can be effectively employed for semantic search and response generation [27], [29].
- **Clarity and Accessibility:** The content needed to be created in a structured and clear manner, facilitating both manual review and computational processing [15], [28].
- **Relevance:** The dataset was meticulously curated to facilitate the process of addressing inquiries related to quranic studies, emphasizing themes that are frequently observed in these discussions [26].

2) *Curation process:* A comprehensive curation procedure was carried out on the dataset in order to get it ready for integration with the retrieval-augmented generation (RAG) system and LLMs:

- **Data Digitization:** The text from the source book was digitized to generate a dataset that can be accessed by machines. Optical Character Recognition (OCR) tools were employed where necessary to convert printed material into digital text [15], [28].
- **Data Structuring:** The content was segmented into surah name, number of verses, reason for the name, general objective, virtues, and relationships. Each field was carefully labeled to facilitate precise retrieval [27].
- **Content Validation:** The digitized and structured dataset was reviewed by experts in Islamic studies to verify its accuracy and alignment with the original source [26].
- **Preprocessing:** Unnecessary or redundant information was removed, and inconsistencies were corrected. Tokenization was performed to split the text into smaller, manageable units for processing by the semantic search system.
- **Vectorization:** The structured data was transformed into high-dimensional vector embeddings using pre-trained language models [30]. This step allowed for

efficient and accurate semantic similarity searches within the dataset.

- **Storage in a Vector Database:** The vectorized dataset was stored in a scalable and efficient vector database, enabling quick retrieval of relevant entries based on user queries [22].

3) *Dataset integrity:* To ensure the integrity and reliability of the dataset, multiple layers of validation were employed, including manual review and automated consistency checks, regular audits of the data were conducted to identify and rectify any errors or discrepancies, and a backup of the raw and processed datasets was maintained for reproducibility and future reference.

4) *Strengths of the dataset:*

- **Richness in Context:** The dataset goes beyond literal translations, providing thematic, historical, and theological insights.
- **High Relevance:** The information directly supports answering user queries about quranic studies.
- **Scalability:** The vectorized format enables integration with modern NLP systems and future upgrades.

This curated dataset ensures that the responses generated by the system are accurate, faithful, and contextually relevant, thereby serving as the foundation for the investigating and evaluation of the LLMs

D. Human Evaluators

In this research, human evaluators were instrumental in evaluating the responses produced by the large language models (LLMs) [31]. The evaluations were carried out using a specially designed website that focused on optimizing the evaluation process and maintaining consistency. The website offered evaluators with questions and responses from several LLMs, allowing for a systematic assessment using preset criteria which are context relevance, answer faithfulness, and answer relevance.

1) *Evaluator selection:* The evaluators were selected with careful consideration to guarantee that they had the proper knowledge and comprehension of quranic studies, given that the study centers on inquiries pertaining to Islamic content. Criteria for selection included:

- **Knowledge of quranic studies:** Evaluators who had either formal education or significant experience in Islamic studies were prioritized.
- **Analytical Skills:** In order to evaluate the quality of responses across multiple dimensions, evaluators were required to possess strong analytical skills.
- **Familiarity with Evaluation Tasks:** It was considered beneficial to have prior experience analyzing textual data or utilizing NLP technologies.

Evaluators from a variety of backgrounds were included to make sure the replies were evaluated fairly and without bias.

2) *Evaluation platform:* The evaluation process was conducted through a dedicated website designed to facilitate efficient and user-friendly assessments. The platform included the following features:

- **Query-Response Display:**
 - Each evaluation session displayed a prompt (query) along with responses generated by different LLMs.
 - Responses were anonymized to prevent bias, ensuring that evaluators were not influenced by the identity of the LLM responsible for generating a response.
- **Scoring Interface:** Evaluators rated each response based on the three evaluation criteria:
 - **Context Relevance:** Precision and alignment of the response with the query.
 - **Answer Faithfulness:** Accuracy of the response in relation to the retrieved dataset content.
 - **Answer Relevance:** Appropriateness and direct pertinence of the response to the query.

A Likert scale (1–5) was used for scoring, where 1 indicated poor performance and 5 indicated excellent performance.

- **Feedback Mechanism:** Evaluators could provide written comments to justify their scores or highlight specific issues in the responses. This feature allowed the identification of nuanced errors that might not be captured by numerical scores alone.

3) *Significance of human evaluations:* The use of human evaluators provided an essential layer of validation for the study, ensuring that the generated responses were assessed not only for technical accuracy but also for their theological and contextual integrity [31]. By leveraging a well-structured platform and robust evaluation criteria, the study ensured that the investigation of LLMs was both rigorous and comprehensive, offering valuable insights into their performance in responding to Islamic queries [32].

E. Metric for Quality Evaluation

This study employs Inter-Evaluator Agreement (IEA) as the primary metric to ensure the quality, reliability, and consistency of human evaluations [33]. Since human evaluators, rather than annotators, were tasked with assessing the generated responses, IEA provides a robust measure of agreement across evaluators, validating the credibility of the evaluation process and the results.

IEA measures the level of consistency between evaluators when scoring responses based on predefined criteria: context relevance, answer faithfulness, and answer relevance. An elevated IEA score signifies uniform application of evaluation criteria by assessors, whereas a diminished score reveals inconsistencies that may necessitate recalibration.

Fleiss' Kappa, as presented in Eq. (2), was employed to calculate IEA due to its appropriateness for evaluating agreement among multiple evaluators concurrently [33]. The consistent monitoring of IEA scores allowed the research team to detect discrepancies promptly and facilitate recalibration

sessions as needed, ensuring evaluators' comprehension and interpretation of the scoring guidelines were aligned. This methodical approach guaranteed that the evaluation process was dependable, replicable, and aligned with the study's goals.

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \quad (2)$$

where:

$$\bar{P} = \frac{1}{N} \sum_{i=1}^N P_i \quad \text{and} \quad \bar{P}_e = \sum_{k=1}^K p_k^2 \quad (3)$$

with:

$$P_i = \frac{1}{n(n-1)} \sum_{k=1}^K n_{ik}(n_{ik}-1), \quad p_k = \frac{\sum_{i=1}^N n_{ik}}{Nn} \quad (4)$$

where:

- κ : Fleiss' Kappa value.
- \bar{P} : Average observed agreement across all items.
- \bar{P}_e : Expected agreement based on chance.
- P_i : Proportion of agreement for item i .
- p_k : Proportion of ratings in category k .
- n : Total number of ratings per item.
- n_{ik} : Number of raters who assigned category k to item i .
- N : Total number of items.
- K : Total number of categories.

F. Large Language Models

The efficacy of numerous large language models (LLMs) in responding to inquiries regarding quranic studies is assessed in this study using a Retrieval-Augmented Generation (RAG) framework [34]. A comprehensive comparison of the capabilities of the LLMs selected for investigation is possible due to the fact that they represent a variety of architectures and parameter scales. More information about the models, how they are put together, and how they relate to this study is given below.

1) *Llama*: Meta AI has developed the Llama (Large Language Model Meta AI) family of models, which are state-of-the-art transformer-based architectures that are optimized for natural language understanding and generation [35]. Llama models are trained on vast, diversified corpora to do various NLP tasks such as contextual reasoning, question answering, and text summarization. They are available in a variety of parameter values, which provides a degree of flexibility in terms of computational requirements and performance. The Llama models were incorporated in this investigation due to their adaptability and superior performance across various parameter sizes. A thorough investigation of how model size affects the capacity to produce faithful, accurate, and contextually relevant

replies was made possible by the range of configurations. The comparison of Llama generations (e.g., Llama3 with Llama3.1) yielded insights on the impact of incremental architectural enhancements on performance.

2) *Gemma*: Google's DeepMind developed the Gemma family of large language models, which are a set of transformer-based designs that are best for understanding and creating natural language. The Gemma models are engineered to provide superior performance while ensuring efficiency, rendering them adaptable for various jobs. These models have undergone pre-training on varied and comprehensive datasets, enabling them to generalize effectively across multiple domains [36]. Gemma models are offered in many parameter scales, including 27b, 9b, and 2b, providing flexibility to optimize performance and computational demands. Their versatility renders them appropriate for applications from resource-intensive jobs to real-time implementations in limited surroundings.

3) *QwQ*: The QwQ model developed by Alibaba Cloud, which has 32 billion parameters, is a large-scale transformer-based language model designed to handle complex natural language processing tasks [37]. While specific information about the QwQ model's architecture or pre-training details is limited in comparison to more established models like Llama and Gemma, its parameter scale positions it as a powerful model capable of capturing complex relationships in textual data. Its large size enables it to perform effectively on tasks requiring nuanced comprehension, contextual reasoning, and content generation across a wide range of subjects.

4) *Phi*: Microsoft created the Phi family of language models, which are a group of lightweight transformer-based models that work best for jobs that involve processing natural language. Even though the Phi models have lower parameter sizes compared to other large-scale models such as Llama and Gemma, they are engineered to exceed expectations, providing robust performance while ensuring computational efficiency [38]. They are pretrained on rigorously selected datasets that prioritize high-quality information, enabling effective generalization across tasks despite a reduced number of parameters. This method guarantees that Phi models provide an exceptional equilibrium between performance and resource demands, rendering them especially appropriate for resource-limited settings and real-time applications.

5) *Key Features of the models*: From smaller-scale models (e.g., 1 billion parameters) to large-scale ones (e.g., 70 billion parameters), the chosen models encompass a wide spectrum of parameter sizes. This variation enables a study of the trade-offs between response quality and computing efficiency. Smaller models might lose accuracy and contextual depth, yet producing faster responses with reduced processing expenses. On the other hand, it is anticipated that larger models will produce more nuanced and high-fidelity outputs, although at the expense of increased computational demands.

The models use transformer architectures, which are efficient in comprehending and producing natural language content. Their architecture allows for capturing the intricate patterns, correlations, and contextual subtleties within the dataset. This work emphasizes the models' capacity to adjust to specific domains, such as quranic studies, despite being

trained on varied datasets. The assessment analyzes the efficacy of these models when enhanced with domain-specific data through the RAG methodology. The study evaluates the models in a zero-shot context, without any fine-tuning on the unique dataset. This method emphasizes the models' intrinsic capacity to generalize and appropriately respond by utilizing external information obtained from the descriptive dataset.

6) *Integration with the RAG framework:* The selected LLMs were integrated with the RAG framework, allowing them to generate contextually relevant responses based on the dataset. The RAG framework enhances the performance of the models by providing contextual input, reducing hallucination, and facilitating citations.

G. Hardware and Software Configuration

The experiments were conducted using a high-performance computing system equipped with an Intel(R) Xeon(R) Gold 5318Y CPU operating at 2.10 GHz with 24 cores. The system featured four NVIDIA RTX A6000 GPUs, each providing 48 GB of VRAM, enabling efficient handling of computationally intensive tasks, particularly those involving deep learning models. Additionally, the system was supported by 128 GB of RAM, ensuring smooth execution of memory-intensive operations and facilitating large-scale data processing. This configuration provided the computational resources necessary to run and evaluate the models effectively.

III. RESULTS

The results of this study are based on the implementation of a RAG architecture, designed to evaluate the performance of 13 LLMs in answering questions related to Quranic studies. By leveraging a descriptive dataset of Quranic surahs, the RAG system facilitates the integration of external knowledge to address the limitations of standalone models. The comparative analysis focuses on assessing the relevance, faithfulness, and contextual accuracy of the responses generated by the LLMs within this framework.

A. Experimental Results

The experimental evaluation assessed the capacity of a variety of large language models (LLMs) to respond to queries related to quranic studies. The models were assessed based on three critical metrics: context relevance, answer faithfulness, and answer relevance. The models were categorized into three categories based on their parameter sizes: large models (marked in red), medium models (marked in yellow), and small models (marked in green). Therefore, the results were analyzed. The following is a comprehensive analysis of the performance of each category.

1) *Context relevance:* Context relevance as shown in Fig. 1 evaluates how well the generated responses align with the query's context.

- **Large Models (Red):** The large models outperformed other categories, with Llama3.3:70b achieving the highest score of 0.583, followed by Llama3.2:3b (0.508), despite being categorized as a small model. Both Llama3.1:70b and Gemma2:27b achieved competitive scores of 0.492 and 0.429, respectively, while

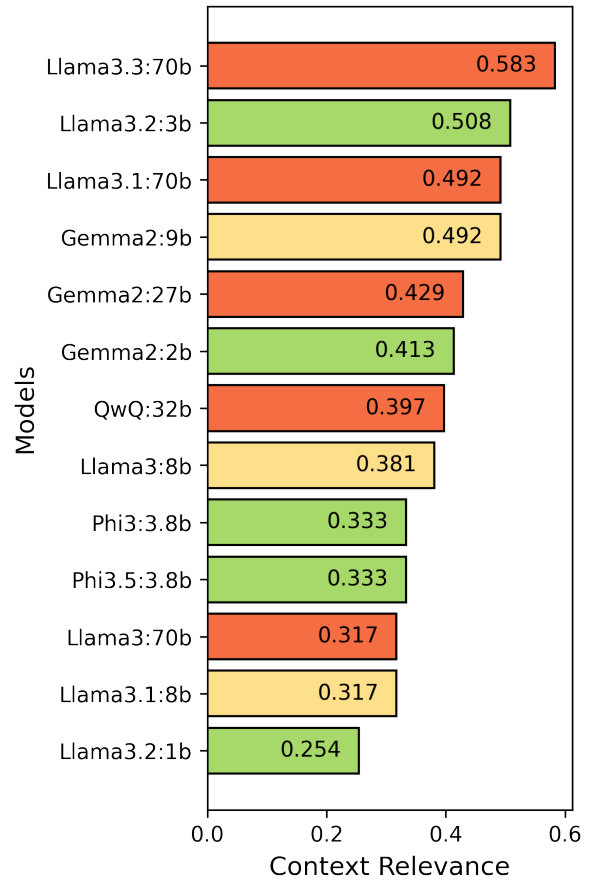


Fig. 1. Context relevance by the 13 LLMs.

QwQ:32b recorded 0.397. These models excel at retrieving relevant information and aligning responses with the query intent due to their larger parameter size.

- **Medium Models (Yellow):** Among the medium models, Gemma2:9b performed best with a score of 0.492, comparable to some large models. Llama3:8b and Llama3.1:8b followed with scores of 0.381 and 0.317, respectively. These models demonstrated decent performance but lagged behind the large models in handling complex or nuanced queries.
- **Small Models (Green):** The small models struggled overall, with Llama3.2:1b achieving the lowest score of 0.254. Phi3.5:3.8b and Phi3:3.8b performed moderately with scores of 0.333, while Gemma2:2b achieved 0.413. Notably, Llama3.2:3b outperformed all expectations with a score of 0.508, surpassing even some medium and large models.

2) *Answer faithfulness:* Answer faithfulness as shown in Fig. 2 measures whether the responses remain consistent with the retrieved content, avoiding inaccuracies or hallucinations.

- **Large Models (Red):** The large models dominated this metric, with Llama3.2:3b (exceptionally performing despite its small category) achieving the top score of 4.619. Llama3:70b and Llama3.1:70b scored

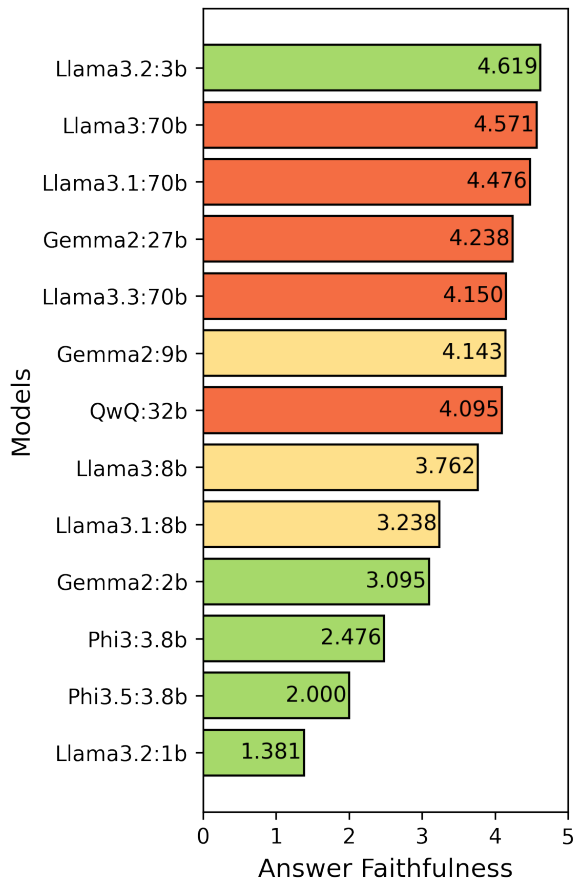


Fig. 2. Answer faithfulness by the 13 LLMs.

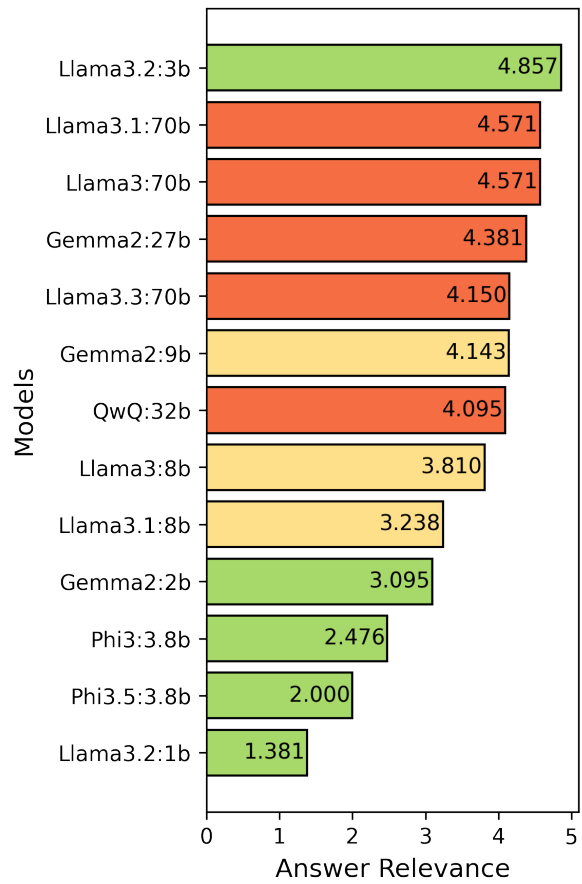


Fig. 3. Answer relevance by the 13 LLMs.

4.571 and 4.476, respectively, while Gemma2:27b and QwQ:32b followed with scores of 4.238 and 4.095. Their ability to maintain faithfulness highlights the advantages of larger parameter sizes.

- **Medium Models (Yellow):** The medium models showed reliable performance, with Gemma2:9b scoring 4.143 and Llama3:8b achieving 3.762. Llama3.1:8b, while consistent, lagged slightly behind with 3.238. These models balanced faithfulness and efficiency but struggled with queries requiring deep contextual reasoning.
- **Small Models (Green):** The small models faced significant challenges. Llama3.2:1b recorded the lowest faithfulness score of 1.381, and Phi3.5:3.8b achieved 2.000, indicating frequent inconsistencies. While Phi3:3.8b scored slightly higher at 2.476, Llama3.2:3b stood out with a score of 4.619, showcasing exceptional faithfulness that rivaled larger models.

3) *Answer relevance:* Answer relevance as shown in Fig. 3 assesses whether the generated responses address the query's intent effectively.

- **Large Models (Red):** The Llama3.2:3b, while classified as small, achieved the highest relevance score of 4.857, followed closely by Llama3:70b and Llama3.1:70b, both scoring 4.571. Gemma2:27b and

QwQ:32b continued to perform well, with scores of 4.381 and 4.095, respectively. These models consistently delivered relevant responses aligned with the intent behind complex queries.

- **Medium Models (Yellow):** The medium models provided strong performance, particularly Gemma2:9b, which achieved 4.143. Llama3:8b followed with 3.810, and Llama3.1:8b recorded a slightly lower score of 3.238. These models addressed moderately complex queries effectively but occasionally lacked depth.
- **Small Models (Green):** The small models exhibited varied performance, with Llama3.2:1b scoring the lowest at 1.381. Phi3.5:3.8b and Phi3:3.8b recorded scores of 2.000 and 2.476, respectively, indicating challenges in providing fully relevant responses. However, Llama3.2:3b once again stood out, achieving the highest score of 4.857, performing on par with the best large models.

4) *Intercoder agreement:* We assessed intercoder agreement, which measures the extent to which evaluators within the same group report the same evaluation for a given instance. To compute this metric, we examined the percentage of instances where both evaluators assigned identical evaluation scores independently. This measure enabled us to evaluate the degree

of concordance between evaluators and assess their consistency in evaluation.

TABLE I. THE KAPPA VALUES FOR THE EVALUATION

Creator	Model	Evaluator
Meta	Llama 3.3 70B	0.80
Meta	Llama 3.2 3B	0.90
Meta	Llama 3.2 1B	0.82
Meta	Llama 3.1 70B	0.82
Meta	Llama 3.1 8B	0.85
Meta	Llama 3 70B	0.92
Meta	Llama 3 8B	0.80
Google	Gemma 2 27B	0.90
Google	Gemma 2 9B	0.92
Google	Gemma 2 2B	0.83
Microsoft	Phi 3.5 3.8B	0.83
Microsoft	Phi 3.3 3.8B	0.93
Alibaba	QwQ 32B	0.89

In Table I, we delve into the inter-annotator agreement analysis, which measures the level of agreement between human evaluators across different models using Fleiss' Kappa. This indicates a similar level of agreement between the evaluators.

IV. DISCUSSION

This section discusses the experimental findings, analyzing the performance of various LLMs categorized into large, medium, and small models across the three evaluation metrics: context relevance, answer faithfulness, and answer relevance [14]. This study examines the relationship among model size, response quality, and computational trade-offs, as well as the behavior of models within the Retrieval-Augmented Generation (RAG) framework.

A. Performance Insights Based on Model Size

The experimental results show that the quality of the responses across all three evaluation criteria is significantly affected by the model size.

- **Large Models (Red):** Large models, including Llama3:70b, Llama3.1:70b, Llama3.3:70b, Gemma2:27b, and QwQ:32b, consistently demonstrate superior performance compared to medium and small models. The models demonstrated superior context relevance due to their larger parameter sizes [39], which facilitate a deeper semantic understanding and enhance their ability to retrieve and integrate pertinent information. Furthermore, their enhanced performance in answer faithfulness and relevance indicates that large models are more adept at minimizing hallucinations [40] and producing responses that accurately address user queries. However, their computational demands remain a major trade-off, requiring substantial memory and processing power. This limits their accessibility in resource-constrained environments, making them more suitable for high-performance systems.
- **Medium Models (Yellow):** Medium-sized models like Gemma2:9b, Llama3:8b, and Llama3.1:8b demonstrated strong performance relative to their parameter sizes, particularly in answer faithfulness and

relevance. These models provide a balance between computational efficiency and response quality, making them ideal for systems where resources are limited but accuracy cannot be compromised. However, their performance in context relevance was slightly lower than that of large models, indicating limitations in capturing deeper relationships within the data. Medium models present a viable option for applications requiring moderate precision while maintaining resource efficiency.

- **Small Models (Green):** The small models, including Llama3.2:3b, Llama3.2:1b, Gemma2:2b, Phi3:3.8b, and Phi3.5:3.8b, faced significant challenges in delivering high-quality responses. Models like Llama3.2:1b and Phi3.5:3.8b scored the lowest across all metrics, reflecting their inability to process complex queries effectively due to their smaller parameter size. Interestingly, Llama3.2:3b emerged as an outlier, achieving performance levels comparable to the large models, particularly in answer faithfulness (4.619) and answer relevance (4.857). This unexpected performance demonstrates the efficacy of architectural improvements and pre-training techniques, even on smaller models. Although small models are computationally efficient and well-suited for lightweight tasks [41], their overall limitations render them less suitable for complex queries that necessitate a deep comprehension of semantics.

B. Effectiveness of the RAG Framework

The implementation of the Retrieval-Augmented Generation (RAG) framework significantly enhanced the response quality of the evaluated models [42]. By incorporating external knowledge from the descriptive Qur'anic dataset, the models can retrieve pertinent information prior to formulating responses. This method alleviated the prevalent issue of hallucination, where language models produce factually inaccurate or irrelevant responses.

The results demonstrate that the larger models derived the greatest advantage from the RAG framework, as their higher parameter sizes facilitated more effective integration of the retrieved context, resulting in enhanced performance across all metrics. Medium and small models showed enhancements, nevertheless, their capacity limitation affected the integration of retrieved knowledge, leading to lower context alignment and fewer accurate responses.

C. Trade-Off Between Computational Resources and Response Quality

The results of this research highlight the trade-off between response quality and computational efficiency throughout several model sizes.

- **Large Models** deliver outstanding performance but they are less practical for real-time or cost-sensitive installations since they need large computational resources.
- **Medium Models** fit for uses with intermediate hardware availability since they offer a useful compromise between dependability of performance and low resource usage.

- Small Models are quick and light-weight, yet they usually underperform on jobs needing sophisticated thinking. Nonetheless, the unexpected findings in Llama3.2:3b suggest that smaller models can still show good performance.

This trade-off emphasizes the need of choosing the suitable model size depending on particular application criteria including accuracy, speed, and resource availability.

D. Surprising Performance of Llama3.2:3b

This study's outstanding discovery is the exceptional performance of Llama3.2:3b, despite its classification as a small model. It outperformed a number of medium and even big models, achieving the highest results in answer faithfulness and answer relevance. According to this finding, factors including pre-training quality, data efficiency, and architectural upgrades can have an impact on model performance, in addition to parameter size. The potential for smaller models to produce high-quality outputs in resource-efficient environments is underscored by the robust performance of Llama3.2:3b when used in conjunction with effective frameworks such as RAG.

E. Implications for Domain-Specific Tasks

The research emphasizes both the difficulties and potential benefits of utilizing general-purpose LLMs for specialized tasks, including responding to inquiries about quranic studies. Although large models excelled in aligning responses with the given dataset, their effectiveness is significantly dependent on the quality and organization of the retrieved content. This research demonstrates how important it is to add domain-specific knowledge [4], like curated descriptive datasets, to improve the models' abilities and lower the risk of hallucinations. The RAG framework proved to be an effective technique for ensuring that responses were contextually correct and based on credible sources.

The experimental results offer important insights into the performance of LLMs of different sizes. Large models provide exceptional accuracy and relevance, though they require significant resources, whereas medium models offer a practical compromise between performance and efficiency. Smaller models, while typically less powerful, demonstrated surprising potential, especially Llama3.2:3b, which excelled across various metrics. The use of the RAG framework enhanced the models' performance by reducing hallucinations and grounding responses in reliable data. These findings highlight the importance of model selection, optimization strategies, and retrieval mechanisms when applying LLMs to domain-specific tasks. Future work will include conducting an ablation study to compare the performance of models with and without the RAG framework, evaluating additional large language models (LLMs), and leveraging LLMs for automatic evaluation of answer faithfulness and answer relevance metrics. Additional fine-tuning strategies and assessments in other specialized domains can also be explored.

V. CONCLUSIONS

This study evaluated multiple large language models (LLMs) of different sizes in responding to Quranic studies-related queries using a Retrieval-Augmented Generation

(RAG) framework. The findings indicate that large models, such as Llama3:70b, Llama3.1:70b, and Gemma2:27b, consistently delivered superior performance in context relevance, answer faithfulness, and answer relevance. However, their computational demands pose challenges for practical deployment. Medium-sized models, including Gemma2:9b and Llama3:8b, demonstrated a balance between efficiency and performance, making them suitable for moderately complex tasks. Interestingly, Llama3.2:3b, a small model, performed comparably to larger models in certain aspects, particularly in answer faithfulness and relevance, suggesting that architectural optimizations can enhance the capabilities of smaller models.

The study also highlights the importance of the RAG framework in improving response quality by grounding answers in external domain-specific knowledge, reducing hallucinations, and ensuring more reliable outputs. These findings emphasize the trade-offs between model size, performance, and computational efficiency, indicating that while large models are ideal for high-accuracy tasks, smaller models, when optimized, can serve as viable alternatives. Future research can focus on further optimizations, fine-tuning strategies, and expanding dataset diversity to enhance model performance across different applications.

ACKNOWLEDGMENT

This research is supported by Universitas Islam Riau.

REFERENCES

- [1] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam, "A review on large language models: Architectures, applications, taxonomies, open issues and challenges," *IEEE Access*, 2024.
- [2] A. H. Nasution and A. Onan, "Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language nlp tasks," *IEEE Access*, 2024.
- [3] S. Alnefaie, E. Atwell, and M. A. Alsalka, "Is gpt-4 a good islamic expert for answering quran questions?" in *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, 2023, pp. 124–133.
- [4] Q. Zhou, C. Liu, Y. Duan, K. Sun, Y. Li, H. Kan, Z. Gu, J. Shu, and J. Hu, "Gastrobot: a chinese gastrointestinal disease chatbot based on the retrieval-augmented generation," *Frontiers in Medicine*, vol. 11, p. 1392555, 2024.
- [5] N. Pipitone and G. H. Alami, "Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain," *arXiv preprint arXiv:2408.10343*, 2024.
- [6] A. Alrayzah, F. Alsolami, and M. Saleh, "Challenges and opportunities for arabic question-answering systems: current techniques and future directions," *PeerJ Computer Science*, vol. 9, p. e1633, 2023.
- [7] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, and Q. Li, "A survey on rag meeting llms: Towards retrieval-augmented large language models," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 6491–6501.
- [8] Z. Sun, X. Zang, K. Zheng, Y. Song, J. Xu, X. Zhang, W. Yu, and H. Li, "Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability," *arXiv preprint arXiv:2410.11414*, 2024.
- [9] S. Patel, H. Kane, and R. Patel, "Building domain-specific llms faithful to the islamic worldview: Mirage or technical possibility?" *arXiv preprint arXiv:2312.06652*, 2023.
- [10] Y. B. I. Radi, *Al-Bitaqat: Chapters of the Noble Quran Explored in 114 Cards*. Dakwah Corner Bookstore (M) Sdn. Bhd, 2023.

- [11] C. Njeh, H. Nakouri, and F. Jaafar, "Enhancing rag-retrieval to improve llms robustness and resilience to hallucinations," in *International Conference on Hybrid Artificial Intelligence Systems*. Springer, 2024, pp. 201–213.
- [12] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.
- [13] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [14] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.
- [15] E. Kamaloo, N. Dziri, C. L. Clarke, and D. Rafiei, "Evaluating open-domain question answering in the era of large language models," *arXiv preprint arXiv:2305.06984*, 2023.
- [16] Y. Zhou, Y. Liu, X. Li, J. Jin, H. Qian, Z. Liu, C. Li, Z. Dou, T.-Y. Ho, and P. S. Yu, "Trustworthiness in retrieval-augmented generation systems: A survey," *arXiv preprint arXiv:2409.10102*, 2024.
- [17] H. K. Shabaz Patel and R. Patel, "Building domain-specific llms faithful to the islamic worldview: Mirage or technical possibility?" *arXiv preprint arXiv:2312.06652*, 2023.
- [18] S. Gupta, R. Ranjan, and S. N. Singh, "A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions," *arXiv preprint arXiv:2410.12837*, 2024.
- [19] Y. L. *et al.*, "Datasets for large language models: A comprehensive survey," *arXiv preprint arXiv:2402.18041*, 2024.
- [20] M. R. Rizqullah, A. Purwarianti, and A. F. Aji, "Qasina: Religious domain question answering using sirah nabawiyah," in *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*. IEEE, 2023, pp. 1–6.
- [21] M. S. Abubakari, W. Shafik, and A. F. Hidayatullah, "Evaluating the potential of artificial intelligence in islamic religious education: A swot analysis overview," in *AI-Enhanced Teaching Methods*. IGI Global, 2024, pp. 216–239.
- [22] Y. Han, C. Liu, and P. Wang, "A comprehensive survey on vector database: Storage and retrieval technique, challenge," *arXiv preprint arXiv:2310.11703*, 2023.
- [23] Z. Jing, Y. Su, Y. Han, B. Yuan, H. Xu, C. Liu, K. Chen, and M. Zhang, "When large language models meet vector databases: A survey," *arXiv preprint arXiv:2402.01763*, 2024.
- [24] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1–17, 2023.
- [25] A. Y. Alan, E. Karaarslan, and Ö. Aydin, "A rag-based question answering system proposal for understanding islam: Mufassirqas llm," *arXiv preprint arXiv:2401.15378*, 2024.
- [26] D. Tam, A. Mascarenhas, S. Zhang, S. Kwan, M. Bansal, and C. Raffel, "Evaluating the factual consistency of large language models through summarization," *arXiv preprint arXiv:2211.08412*, 2022.
- [27] W. Zhou, S. Zhang, H. Poon, and M. Chen, "Context-faithful prompting for large language models," *arXiv preprint arXiv:2303.11315*, 2023.
- [28] C. Wang, S. Cheng, Q. Guo, Y. Yue, B. Ding, Z. Xu, Y. Wang, X. Hu, Z. Zhang, and Y. Zhang, "Evaluating open-qa evaluation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [29] Y. Huang, S. Chen, H. Cai, and B. Dhingra, "Enhancing large language models' situated faithfulness to external contexts," *arXiv e-prints*, pp. arXiv–2410, 2024.
- [30] S. S. Monir, I. Lau, S. Yang, and D. Zhao, "Vectorsearch: Enhancing document retrieval with semantic embeddings and optimized search," *arXiv preprint arXiv:2409.17383*, 2024.
- [31] A. Elangovan, L. Liu, L. Xu, S. Bodapati, and D. Roth, "Considers-the-human evaluation framework: Rethinking human evaluation for generative large language models," *arXiv preprint arXiv:2405.18638*, 2024.
- [32] K. Feng, K. Ding, K. Ma, Z. Wang, Q. Zhang, and H. Chen, "Sample-efficient human evaluation of large language models via maximum discrepancy competition," *arXiv preprint arXiv:2404.08008*, 2024.
- [33] F. Moons and E. Vandervieren, "Measuring agreement among several raters classifying subjects into one-or-more (hierarchical) nominal categories. a generalisation of fleiss' kappa," *arXiv preprint arXiv:2303.12502*, 2023.
- [34] S. B. Islam, M. A. Rahman, K. Hossain, E. Hoque, S. Joty, and M. R. Parvez, "Open-rag: Enhanced retrieval-augmented reasoning with open-source large language models," *arXiv preprint arXiv:2410.01782*, 2024.
- [35] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.
- [36] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé *et al.*, "Gemma 2: Improving open language models at a practical size," *arXiv preprint arXiv:2408.00118*, 2024.
- [37] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.
- [38] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv preprint arXiv:2404.14219*, 2024.
- [39] S. Badshah and H. Sajjad, "Quantifying the capabilities of llms across scale and precision," *arXiv preprint arXiv:2405.03146*, 2024.
- [40] S. Tonmoy, S. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, and A. Das, "A comprehensive survey of hallucination mitigation techniques in large language models," *arXiv preprint arXiv:2401.01313*, 2024.
- [41] L. Chen and G. Varoquaux, "What is the role of small models in the llm era: A survey," *arXiv preprint arXiv:2409.06857*, 2024.
- [42] X. Su and Y. Gu, "Implementing retrieval-augmented generation (rag) for large language models to build confidence in traditional chinese medicine," 2024.