



An-Najah National University
Faculty of Graduate Studies

**NEBRAS: A RAG-BASED QUESTION
ANSWERING SYSTEM FOR ISLAMIC AND
LEGAL GUIDANCE**

By
Samer Nitham Al-Huwari

Supervisor
Dr. Hamed Abdelhaq

**This Thesis is Submitted in Partial Fulfillment of the Requirements for the Degree
of Master of Artificial Intelligence, Faculty of Graduate Studies, An-Najah
National University, Nablus - Palestine.**

2025

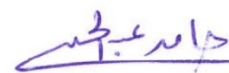
NEBRAS: A RAG-BASED QUESTION ANSWERING SYSTEM FOR ISLAMIC AND LEGAL GUIDANCE

By

Samer Nitham Al-Huwari

This Thesis was Defended Successfully on 27/02/2025 and approved by

Dr. Hamed Abdelhaq
Supervisor



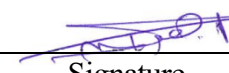
Signature

Prof. Mohammed Awad
External Examiner



Signature

Dr. Ajmad Hawwash
Internal Examiner



Signature

Dedication

To my incredible family—my parents and siblings—whose unwavering love, patience, and encouragement have been a constant source of strength throughout this journey. Your faith in me has been my greatest motivation, and your words have guided me forward.

To Dr. Hamed, my mentor and guide, who not only supervised this research but also opened doors to opportunities I never imagined. Your belief in my potential has shaped my academic journey profoundly.

Lastly, to myself, for demonstrating unwavering determination and confronting this journey with courage. This achievement stands as a testament to resilience, perseverance, and an enduring belief in the promise of brighter days ahead.

Acknowledgements

I would like to express my heartfelt gratitude to all those who contributed to the success of this research. My sincere thanks go to the Palestinian Dar Al-Ifta'a for their invaluable assistance in evaluating the fatwas generated by the system, providing critical insights that greatly enriched this work. I am also deeply thankful to An-Najah National University for providing a sample of Islamic fatwa related to the local Palestinian community, which served as a resource for implementing the Islamic Fatwa answer generation. Without the support and collaboration of these esteemed institutions, this research would not have been possible.

Declaration

I, the undersigned, declare that I submitted the thesis entitled:

NEBRAS: A RAG-BASED QUESTION ANSWERING SYSTEM FOR ISLAMIC AND LEGAL GUIDANCE

I declare that the work provided in this thesis, unless otherwise referenced, is the researcher's own work, and has not been submitted elsewhere for any other degree or qualification.

Student's Name: Samer Nitham Al-Huwari

Signature: 

Date: 27/02/2025

List of Contents

Dedication.....	III
Acknowledgements.....	IV
Declaration.....	V
List of Contents.....	VI
List of Tables.....	VIII
List of Figures.....	IX
List of Appendices	X
Abstract.....	XI
Chapter One: Introduction and Theoretical Background.....	1
1.1 Theoretical Background.....	5
1.1.1 Question Answering.....	5
1.1.3 Prompt Engineering	7
1.1.4 Retrieval Augmented Generation.....	8
1.1.5 Agents in Large Language Models	14
1.2 Literature Review	14
1.3 Problem Statement.....	18
1.4 Aims of Study	20
1.5 Hypotheses of Study	21
1.5.1 Accuracy Hypothesis	21
1.5.2 Adaptability and Scalability Hypothesis.....	21
1.5.3 Language-Specific Performance Hypothesis.....	22
Chapter Two: Methods.....	23
2.1 Data Collection	23
2.1.1 Islamic Fatwa Dataset Collection	24
2.1.2 University Help-desk Dataset Collection	24
2.2 Data Pre-processing and Structuring	26

2.2.1 Islamic Fatwa Dataset Pre-processing and Structuring	27
2.2.2 NNU Dataset Pre-processing and Structuring	31
2.3 Implementation	32
2.3.1 Vector Database	32
2.3.2 Indexing and Chunking.....	34
2.3.3 QA Pipeline.....	35
Chapter Three: Experimentation and Results	42
3.1 Indexing and chunking.....	43
3.2 Implementation	47
3.3 Experiment with Islamic Fatwa Dataset	47
3.3.1 Baseline Evaluation: Responses Without Retrieval.....	47
3.3.2 Context Effectiveness on Fatwa Answer Generation.....	49
3.3.3 Employing RAG	52
3.4 Experiment with An-najah National University Dataset	60
3.4.1 Baseline Evaluation: Responses Without Retrieval.....	61
3.4.2 Hybrid Retrieval Pipeline Evaluation with NNU Dataset	62
Chapter Four: Discussions and Conclusion	64
List of Abbreviations	66
References.....	67
Appendices.....	77
المخلص	ب

List of Tables

Table 1: Islamweb Fatwa Fields Mapping.....	24
Table 2: NNU Academic Majors Fields Mapping	25
Table 3: NNU Academic Courses Fields Mapping.....	26
Table 4: Token Counts from the Scraped Islamweb Dataset	27
Table 5: Frequency of Start and Trailing Words in Introductory Sentences in Fatwa	29
Table 6: Frequency of First Words After Removing Introductory Sentences in Fatwa ..	29
Table 7: Token Counts from the Scraped Islamweb Dataset After Cleaning.....	30
Table 8: Embedding Models Evaluation.....	46
Table 9: Summary of Implementation Technologies	47
Table 10: Palestinian Dar Al Ifta'a Evaluation of Generated Fatwas.....	50

List of Figures

Figure 1: Traditional RAG Pipeline.....	9
Figure 2: Two-Level Hierarchical Structure Indexing for One Document.....	10
Figure 3: Query Expansion Using HyDE	13
Figure 4: Dataset Field Mapper	34
Figure 5: QA Agentic RAG Pipeline	36
Figure 6: Hybrid Retrieval Pipeline.....	41
Figure 7: Ground Truth Evaluation With and Without Context.....	51
Figure 8: Comparing GT Automatic Metrics & Human Evaluation.....	58
Figure 9: Ground Truth Comparison Between Nebras and Baseline Models.....	60
Figure 10: NNU Dataset Ground Truth Scores Comparison	63

List of Appendices

Appendix A: Google Gemini Response to Accounting Major Required GPA	77
Appendix B: GPT-4o Response to Accounting Major Required GPA	80
Appendix C: LLM and Mufti Responses to Fatwa on Promoting Products.....	81
Appendix D: Dataset Scraped Fields Mapping.....	86
Appendix E: Fatwa Dataset Category Distribution	88
Appendix F: Field Augmentation for Structuring NNU Dataset	89
Appendix G: Query Decomposition Agent's Prompt.....	90
Appendix H: Query Classification Agent's Prompt	92
Appendix I: Candidate Answer Agent Prompt	95
Appendix J: Context Relevance Prompt.....	96
Appendix K: Answer Generation Agent Prompt	99
Appendix L: Evidence Extraction Agent Prompt	101
Appendix M: Irrelevant Query Response Agent Prompt.....	104
Appendix N: Islamic Fatwa with No Context (Baseline Evaluation).....	106
Appendix O: Prompt for Generating Context-based Answers.....	107
Appendix P: PDR Evaluation	108
Appendix Q: Retrieval with Ranking Model.....	109
Appendix R: HyDE.....	110
Appendix S: Query to Question.....	111
Appendix T: Query to Topic	112
Appendix U: Hybrid Retriever	113
Appendix V: Ground Truth Evaluation for 100 Fatwas.....	114
Appendix W: NNU Baseline.....	115
Appendix X: NNU Hybrid Retrieval	116
Appendix Y: Experimentation Download Link	117

NEBRAS: A RAG-BASED QUESTION ANSWERING SYSTEM FOR ISLAMIC AND LEGAL GUIDANCE

By

Samer Nitham Al-Huwari

Supervisor

Dr. Hamed Abdelhaq

Abstract

Question answering (QA) systems are essential tools in natural language processing (NLP), designed to interpret user queries and generate relevant answers. These systems have evolved over time from rule-based models to advanced machine-learning-based approaches. The emergence of the transformers architecture and Large Language Models (LLMs) have set the stage for modern QA systems.

LLMs have transformed QA by leveraging vast datasets to generate human-like responses across various domains and their ability to understand complex linguistic patterns. However, LLMs often generate plausible but incorrect answers particularly in specialized domains like law and religion where accuracy is critical. This phenomenon is known as “hallucination”. The risk of “hallucination” is increased when dealing with a complex language like Arabic. Arabic language, with its rich morphology, diverse dialects, and its dependency on diacritics, present significant challenges for LLMs primarily trained on Western languages.

Fine-tuning LLMs for domain-specific tasks is time-intensive, and computationally-expensive, given their massive parameters size, demanding innovative approaches to mitigate the LLMs hallucination issue without extensive re-training.

This thesis introduces Nebras, a generic multi-domain QA system leveraging a Retrieval-Augmented Generation (RAG) framework, LLM agents, and a hybrid retrieval approach. Nebras’s knowledge base can be dynamically extended by following simple guidelines and using its built-in mapping component, enabling it to adapt to any textual dataset. By employing an Agentic RAG pipeline, Nebras optimizes each processing stage using specialized agents. Furthermore, it utilizes pre-trained LLMs without fine-tuning, enhancing scalability and reducing computational costs.

Experimentation results demonstrated Nebras's performance in Arabic domain-specific QA. In the Islamic fatwa domain, it achieved a BERTScore-F1 of 70.94%, a METEOR of 13.49%, with 9 accepted fatwas compared to only 7 accepted from GPT-4o. In the university help-desk domain, Nebras achieved a BERTScore-F1 of 75.80%, METEOR of 40.20%, and BLEU of 9%, significantly outperforming the BLEU score of 2.3% from GPT-4o's. These results highlight Nebras's ability to enhance factual accuracy, confirming its potential as a scalable Arabic QA solution.

Keywords: NEBRAS, Multi-domain Arabic question-answering, Agentic Retrieval-Augmented Generation, Large Language Models

Chapter One

Introduction and Theoretical Background

Research in language modeling began in the early 1950s with Shannon's work exploring the predictive and compressive capabilities of simple n-gram models on natural language [1]. Over time, statistical language modeling became a foundational technique in various natural language understanding and generation tasks, such as speech recognition, information retrieval, and question answering [2].

The introduction of neural network-based approaches represented a paradigm shift in language modeling, moving beyond traditional statistical methods [3]. Early deep learning approaches laid the groundwork for future progress by enabling the capture of more complex textual patterns and contextual dependencies [4]. This advancement led to the development of transformer architectures which revolutionized the field by enabling the effective handling of long-range dependencies and parallel processing [5]. These advancements marked a new era of NLP and paved the way for the emergence of LLMs[6].

LLMs have significantly advanced NLP by demonstrating their ability to understand and generate human-like text [7]. LLMs scale, architecture, and their training on extensive and diverse datasets, enabled them to capture hidden and complex patterns in language [8]. These capabilities allowed the models to achieve state-of-the-art performance across a variety of NLP tasks, including but not limited to text summarization, sentiment analysis, machine translation, and question-answering [9] .

LLM capabilities are often tested by the complexities of other languages like Arabic [10,11]. Unlike most Western languages, Arabic is characterized by its rich and complex morphology, its dependence on diacritics, and its multiple verb variations and noun forms [12]. This complexity makes accurate language processing a challenge. Diacritical marks (harakat) are often omitted in most standard texts in written Arabic, such as newspapers, books, and other general writing. These marks indicate short vowels and certain grammatical features, and their absence can introduce ambiguity in word meaning and pronunciation [13]. This ambiguity arises because many Arabic words are written as skeletons without diacritics, making the intended word dependent on context. For example, the root "كتب" (k-t-b) can mean "he wrote" (kataba), or "books" (kutub),

depending on the diacritics and context. This ambiguity becomes even more noticeable when paired with the wide variety of Arabic dialects, which often differ significantly from Modern Standard Arabic in vocabulary, syntax, and pronunciation [11]. Combining these factors with the lack of high-quality Arabic datasets, lead to weaker performance and frequent misunderstandings in Arabic-language tasks compared to languages with richer datasets [14].

In the field of QA, LLMs typically rely on the knowledge acquired during training and stored within their parameterized memory to generate responses [15]. This reliance makes them exposed to knowledge gaps in specialized domains, and difficulties in processing diverse languages, increasing the risk of generating irrelevant or factually incorrect answers —a phenomenon commonly referred to as hallucination [16,17]. Moreover, the inherent complexity of natural language often results in models misinterpreting questions, leading to responses that are either irrelevant or incorrect [18]. For instance, when posed with the Arabic question:

ما هو معدل القبول لتخصص المحاسبة في جامعة النجاح الوطنية؟

(English translation: What is the required GPA for the Accounting major at An-Najah National University?)

Google's Gemini 1.5 *Pro* didn't provide direct answer about the minimum required GPA, instead it explained why academic majors do not have a fixed minimum required GPA at An-najah national University, resulting in an irrelevant response. The generated response is detailed in Appendix A.

When OpenAI's GPT-4o was presented with the same question, it demonstrated a clear understanding of the query but incorrectly identified the minimum required GPA for admission as being between 80–85% while in 2024, the actual minimum GPA for admission to the Accounting major at An-Najah National University (NNU) is 65%. The complete response from GPT-4o can be found in Appendix B.

These challenges are further increased when dealing with specialized domains in general, particularly in sensitive areas like legal counseling and religion, which often require deep knowledge understanding and reasoning, as well as access to information that may not have been available to the LLM during its creation [19].

In Islamic religious context for example, Fatwas serve as a crucial guide for Muslims on legal, ethical, and religious matters. Fatwas are formal statements issued by qualified Islamic scholars (Muftis) after analyzing Quranic verses, Hadith narrations (teachings attributed to the Prophet Muhammad), and the established reasoning of past Islamic scholars and Imams to establish an answer for a specific case [20]. The significance of Fatwas lies in their ability to translate broad Islamic principles into practical guidance for individuals facing unique situations. They offer clarification on permitted actions, ethical dilemmas, and religious obligations in a constantly evolving world. Therefore, inaccurate or misleading Fatwas can have serious religious and legal repercussions for the user. Moreover, Islamic Fatwas can differ among scholars due to variations in jurisprudential schools of thought (madhahib), resulting in a diversity of perspectives on the same issue[21,22].

These factors highlight the importance of accurate and reliable resources when seeking guidance through Islamic Fatwas. Accurately understanding the context of Qur'anic and Hadith verses, and reasoning within Islamic text is crucial for navigating the complexities of Islamic law [11]. Current LLMs might struggle with this depth of domain-specific knowledge. For example, Google's Gemini 1.5 Pro and OpenAI's GPT-4o were asked the following Arabic fatwa question:

حكم الترويج لمنتجات لشركة مقابل عمولة مع شرط مسبق على المروج بدفع مبلغ مالي

(translated to English): What is the Islamic Fatwa ruling on promoting a company's products in exchange for a commission, with a prior condition requiring the promoter to pay a sum of money?

Both the Gemini and GPT-4o responses presented multiple rulings for various scenarios, deeming it جائز (permissible) in some cases, whereas the Mufti's response explicitly declared it حرام (forbidden). The detailed responses from both Google's Gemini 1.5 *Pro* and OpenAI's GPT-4o, along with the Mufti's ground truth answer, are provided in Appendices C - C.1, C.2, and C.3.

In contrast, academic domains often involve factual queries requiring precision, clarity, and contextual understanding, such as university policies, course requirements, or research guidelines. These inquiries are typically based on carefully crafted institutional policies and thoroughly reviewed documentation [23,24]. Specialized academic domains

pose unique challenges for LLMs due to the diversity, complexity, and constant evolution of educational systems and their policies. Research highlights that LLMs struggle with reasoning tasks requiring specialized knowledge which necessitate efforts in domain-specific data acquisition and fine-tuning [25]. The dynamic nature of academic content and institutional guidelines further complicates matters, as static training data risks outdated or biased responses in such domains [26,27].

Fine-tuning is one approach to mitigate LLM hallucinations, either by expanding the model's knowledge or enhancing its linguistic capabilities. However, updating a model's parameterized memory through fine-tuning is challenging, as there's no established method for directly overwriting existing knowledge, leading to potential knowledge conflicts—an ongoing research area [28,29]. Expanding a model's language capabilities within a specific domain typically involves fine-tuning on data relevant to that domain. If the model lacks initial support for the target language, pre-training on a large corpus in that language is usually necessary before domain-specific fine-tuning [30]. Furthermore, given the large number of parameters in LLMs, fine-tuning demands significant computational resources, making it a time-consuming and expensive process. However, advancements in parameter-efficient fine-tuning (PEFT) methods, such as Low-Rank Adaptation (LoRA) or adapter modules, can enable targeted updates without requiring full re-tuning of the model, mitigating some of these challenges but requiring careful tuning of their hyperparameters, which can impact performance if not optimized [35].

An alternative approach to mitigating hallucination in LLMs is RAG. RAG integrates a language model with an external retriever that fetches relevant documents from a knowledge base, enabling the model to generate accurate responses by referencing up-to-date information [19]. This method can be more efficient than fine-tuning because it avoids altering the model's internal parameters, reducing computational costs, and enabling dynamic updates to the knowledge base [32]. However, RAG has its challenges, including ensuring retrieval accuracy, managing outdated or conflicting information, and maintaining retrieval latency at scale [33]. Additionally, applying RAG to languages like Arabic requires effective retrieval mechanisms for morphologically rich languages [34]. Recent advancements in RAG, such as the incorporation of intelligent agents, have improved retrieval and synthesis processes, addressing limitations like retrieval accuracy and latency while expanding the system's capabilities [35–37].

Agents in LLMs leverage their extensive reasoning capabilities to autonomously perform tasks by interpreting complex instructions and execute multi-step processes, acting as intermediaries between users and computational resources [38,39]. These agents demonstrate versatility across diverse domains such as social and natural sciences and engineering, but challenges like maintaining consistent performance, accurate contextual understanding, and seamless tool integration persist [38,40].

Traditional RAG systems lack decision-making and validation mechanisms. In contrast, Agentic RAG integrates intelligent agents to dynamically select, process, and validate relevant information, enhancing response accuracy, contextual relevance, and robustness compared to traditional RAG systems [15,35,36]. In addition, employing agents into the RAG approach, implementing Agentic RAG, gives the ability to handle complex, multi-step reasoning tasks more effectively and deliver contextually appropriate responses [37,41].

This research introduces Nebras, a novel multi-domain Arabic question-answering system designed for adaptability across various domains. Unlike traditional models that require domain-specific fine-tuning, Nebras processes textual datasets provided by users to expand its domain coverage. This capability allows the system to incorporate new datasets provided by administrators without the need for costly and time-consuming re-training processes. Nebras leverages a hybrid retrieval approach that combines multiple techniques to ensure relevant information retrieval from large-scale corpora. It employs Agentic RAG pipeline which consist of multiple agents, each performing a specific task in the system's proposed pipeline. These agents work in coordination with pre-trained LLMs, enabling Nebras to deliver accurate, factually grounded, and context-aware responses. This design ensures that Nebras remains flexible, scalable, and capable of addressing diverse Arabic language applications across various sectors.

1.1 Theoretical Background

This section presents an overview of the theoretical concepts in the field of question-answering systems and recent advancements in the field.

1.1.1 Question Answering

QA is a core task within the field of NLP, which deals with enabling machines to understand and process human language. QA systems are designed to answer questions

posed by users in natural language aiming to provide relevant, accurate answers by analyzing the input query and generating appropriate responses [42,43].

Based on their reliance on different data structures, QA systems can be categorized into text-based and knowledge-based [44]. Knowledge-based QA utilizes structured knowledge bases (KBs) that store information as triples in the format (subject, predicate, object) [45]. These systems answer questions by querying single or multi-relation facts in the KB, with single-relation questions relying on one fact and multi-relation questions requiring reasoning over multiple connected facts [46,47].

In contrast, Text-based QA retrieves answers from unstructured text, such as documents or articles, by identifying and extracting the most relevant passages that matches the query [45]. These systems typically follow a three-steps pipeline [43]:

1. *question processing*: involves query formulation and answer type detection using classifiers.
2. *document and passage retrieval*: which employs information retrieval (IR) models to extract relevant text segments.
3. *answer extraction*: where the system measures the similarity between the query and candidate answers to determine the most appropriate response.

Advanced deep neural models are often employed to enhance text-based QA by accurately matching questions with potential answers [44,45].

1.1.2 Large Language Models

LLMs are computational systems designed to understand and generate human language by leveraging statistical methods to predict word sequences or generate responses based on input [48]. These models achieve remarkable performance in tasks such as text generation, translation, and summarization (to name a few) due to their large-scale training on extensive datasets and their implementation of the Transformer architecture [49]. Central to this architecture is the self-attention mechanism, which enables efficient parallel processing and assigns varying importance to input tokens. This allows the model to capture long-range dependencies effectively [5]. Transformers have powered state-of-the-art language models like Google's BERT [6], Facebook's RoBERTa [50], and OpenAI's GPT-3 and GPT-4 [8,51].

The massive size of LLMs, billions to trillions of parameters (hence the name "Large"), is a critical factor in their performance. This scale enables LLMs to learn complex language patterns and develop sophisticated linguistic abilities [49]. Moreover, LLMs have demonstrated the emergence of novel capabilities, such as in-context learning [52]. This ability allows the models to adapt to specific tasks and generate contextually relevant responses, making them well-suited for a wide range of applications, including dialogue systems [48], step-by-step reasoning [53], even processing none-textual input like images and audio (known as Multi-modal Large Language Models) such as Google's PaLM [54].

As knowledgeable and plausible-sounding they are, LLM-generated responses can be nonsensical or factually incorrect, therefore cannot always be trusted. This phenomenon is common in LLMs and known as "hallucination" [55,56].

LLMs typically rely on the knowledge they learned from the training process and stored in their parameterized memory [15]. Hallucinations often arise from models attempting to fill gaps in knowledge by generating responses based on probabilistic associations rather than verifiable knowledge [24]. The hallucination problem becomes particularly concerning in specialized domains, such as legal, or medical fields, where incorrect information can have serious consequences and can mislead users, undermining trust in LLM-generated responses. Even with their enhanced capabilities in understanding, reasoning, and generation, Multi-modal LLMs are not immune to generate hallucinated content that may appear plausible [57].

One established approach for interacting with LLMs is prompt engineering. This technique involves crafting specific textual prompts to guide the LLM's response generation process [62]. Users can steer the LLM's output towards desired outcomes and tasks by carefully designing these prompts.

1.1.3 Prompt Engineering

Prompt engineering enables LLMs to perform a wide array of tasks without requiring retraining nor fine-tuning. Practitioners can guide LLMs toward generating contextually relevant and accurate outputs by designing the input prompts thoughtfully, leading to enhancing the LLMs performance [63].

Prompt engineering involves crafting specific instructions or queries (prompts) that encourage the model to generate responses aligned with the user's goal. Recent studies have expanded the landscape of prompt engineering by exploring methods such as zero-shot and few-shot prompting [60,61]. Zero-shot prompting allows LLMs to perform new tasks without any task-specific training by relying entirely on the model's pre-trained knowledge base [61]. This technique is widely used for large models like GPT-3 where a well-structured prompt can enable the model to perform tasks it has never encountered before [51]. On the other hand, few-shot prompting has shown improvements in handling more complex tasks by providing the model with a few example inputs and outputs, even with minimal additional input data [62].

A notable advancement in prompt engineering is Chain-of-Thought (CoT) prompting which was introduced by Wei et al. in 2022 [53]. CoT prompts guide the model through logical steps, enhancing its ability to process and produce logical, reasoned outputs, which makes it particularly useful for complex reasoning tasks such as mathematical problem solving and commonsense reasoning. Further improvements to this approach, such as Auto-CoT, automate the generation of reasoning chains, thereby enhancing robustness and reducing human effort in creating example-based prompts [63].

The introduction of role-prompting has also improved the specificity of model outputs by assigning a "role" to the model in the prompt, such as "acting as an expert" or "a friendly assistant". This helps guide the model towards more contextually appropriate and accurate responses in various domains [64].

Despite these advancements, challenges persist in optimizing prompts for more complex tasks due to the influence of multiple factors, including task complexity, model biases, and token limitations [65,66].

1.1.4 Retrieval Augmented Generation

RAG is a method that addresses factual hallucination and limitations in domain-specific knowledge for LLMs by incorporating external knowledge through information retrieval [32]. As illustrated in Figure 1, a traditional RAG-based QA system consists of three steps[23]:

1. *Indexing*: storing vector representations (embeddings) of text chunks. These embeddings allow efficient retrieval of relevant text based on similarity and often

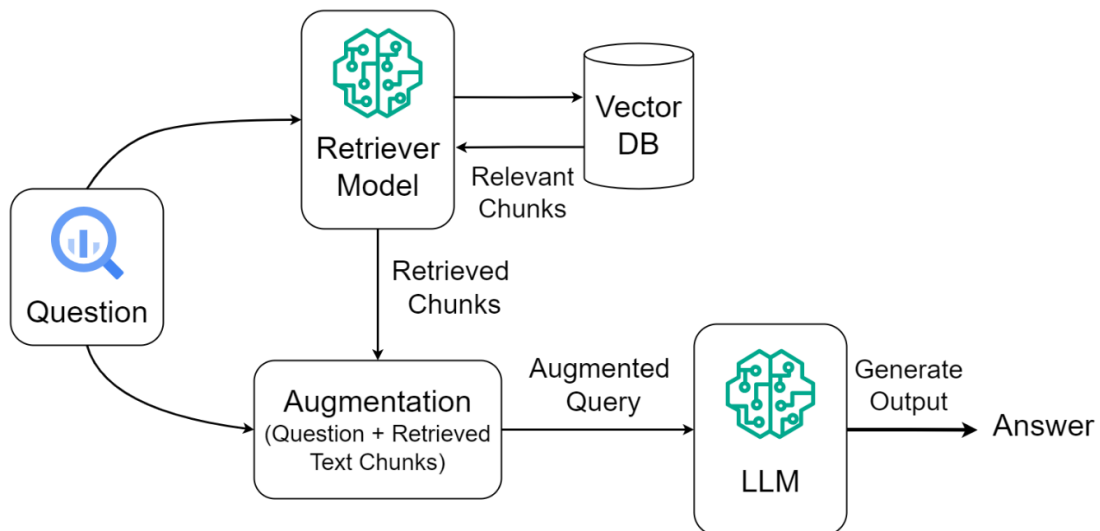
stored in a vector database. Vector databases are specialized data management systems designed for storing, indexing, and querying high-dimensional vectors. They support similarity search by using nearest-neighbor search algorithms, enabling efficient retrieval of semantically related data [67]. Unlike traditional relational databases, which organize data in tables with rows and columns and rely on structured query languages (SQL), vector databases handle unstructured data by representing it as numerical vectors. While regular databases perform exact-match searches, vector databases perform approximate nearest-neighbor searches.

2. *Retrieval*: identifies the most relevant text chunks for a given question. It uses the same embedding model as the vector storage to find similar text based on the question embedding.
3. *Generation*: generating answers from the retrieved text segments using a language model.

The traditional RAG (also known as Naive RAG) adapts the Retrieve-Read method [68] which takes the user's query, matches it against indexed documents, then retrieve the most relevant k documents.

Figure 1

Traditional RAG Pipeline



The Naive RAG approach faces significant drawbacks across its retrieval, generation, and augmentation stages. During retrieval, it struggles with selecting relevant or well-aligned content and may fail to retrieve essential information, leading to misaligned (inaccurate)

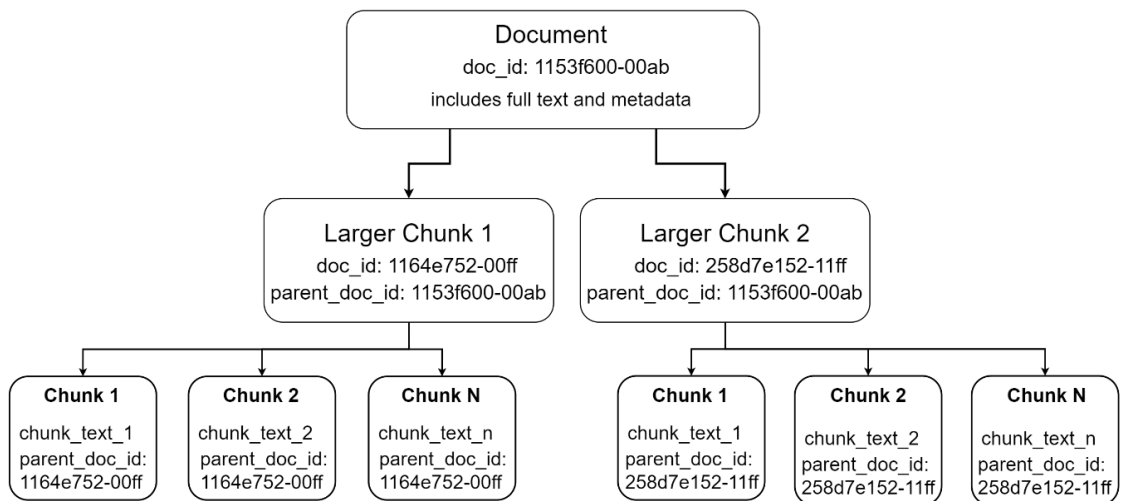
responses [76]. In the generation phase, the model often produces hallucinated or unsupported content, which compromises response quality and reliability. Additionally, issues of irrelevance in generated outputs can further reduce effectiveness [19,69]. The augmentation process presents challenges in addressing coverage errors and ensuring retrieved content aligns effectively with task-specific requirements [71]. Moreover, generation models can produce repetitive outputs or fail to integrate retrieved content meaningfully, resulting in responses that reflect surface-level repetition rather than deep contextual understanding [19]. To address these drawbacks, several modifications have been introduced:

Index Structuring

A structural index enhances retrieval in RAG systems by organizing documents hierarchically, creating multilevel parent-child relationships among document chunks instead of indexing text chunks independently without any relationships or links between them. This method is called Hierarchical Structure Indexing [19] where each child document stores its parent document id to retrieve the entire parent document instead of the text chunk, reducing issues associated with redundant or contextually-disjointed chunks. Figure 2 illustrates a two-level Hierarchical Indexing for a single document.

Figure 2

Two-Level Hierarchical Structure Indexing for One Document



Another indexing structure method is Knowledge Graph Structure Indexing, where document chunks are represented as nodes, and the relation between chunks are represented as edges. Adding a knowledge graph (KG) index further strengthens this

structure by linking concepts and entities within the documents. This approach not only minimizes errors in retrieval but also translates the process into steps the LLM can interpret, leading to more accurate and contextually relevant responses. Methods like Knowledge Graph Prompting (KGP) [72] use KGs to represent document sections as nodes (such as pages or tables) and their connections as edges. This representation allows capturing semantic relationships and enabling coherent knowledge retrieval and reasoning across multiple documents [19].

Chunking Optimization

Chunking in RAG is essential for efficient and accurate query answering. It balances providing enough relevant context with minimizing irrelevant data, thus improving retrieval quality and computational performance [73]. Fixed chunk sizes are often used in RAG pipelines and can sometimes lead to insufficient or excessive information within each chunk. Techniques such as recursive chunking and sliding windows address this issue by segmenting content based on natural language structures, like punctuation and sentence boundaries, or by overlapping chunks to preserve coherence [74]. For documents with clear structures, like financial reports, more advanced chunking methods like element-based chunking provides a tailored approach by using structural cues like titles and tables to create chunks, leading to more accurate retrieval [74]. While semantic and agentic chunking strategies provide improved contextual alignment, their increased complexity and computational requirements highlight the trade-off between retrieval accuracy and processing efficiency [73].

Metadata Attachment

Incorporating metadata into document chunks can contribute to enhancing retrieval performance in RAG systems, particularly in multi-document contexts [75]. Attaching metadata such as page numbers, document titles, authorship information, timestamps, and other relevant identifiers allows precise filtering and prioritize most recent information, thereby improving retrieval relevance and minimizing the potential for confusion between similar chunks originating from distinct documents [76]. Furthermore, artificially constructed metadata bridges the semantic gap between user queries and document content, like paragraph summaries or hypothetical questions generated by LLMs, resulting into more accurate responses [75,76]. Incorporating metadata annotations also add contextual layers to each chunk therefore improving the RAG system's capacity to retrieve and present coherent information from diverse sources therefore [19].

Re-ranking

Re-ranking models are essential in RAG systems, as they refine document retrieval by applying a secondary prioritization based on relevance (secondary to the similarity score) [19]. By reorganizing retrieved chunks, re-ranking ensures that the most relevant content appears at the top, optimizing the document pool that is provided to the language model [19,76]. This prioritization process can be rule-based, relying on metrics such as relevance, diversity, and mean reciprocal rank (MRR), or it can employ model-based methods leveraging advanced natural language processing models [19,76].

Re-ranking addresses the limitations of initial retrieval methods, which often prioritize similarity (e.g., through cosine similarity scores) without fully assessing relevance [19,76]. Advanced re-ranking models, like cross-encoders, are particularly effective in accurately scoring chunk relevance for a given query, often outperforming simpler bi-encoder models in this domain [19]. However, these approaches are computationally intensive, especially those that use pre-trained language models (PLMs) [19,77].

Generative LLMs, such as GPT-3, can further enhance re-ranking by generating synthetic queries for domain-specific training, enhancing the accuracy of the relevance-based ordering without requiring vast amounts of new labeled data [77]. Despite their performance, these re-ranking models are resource-intensive, emphasizing a trade-off between retrieval precision and computational cost [19].

Context Compression

Context compression is a technique in RAG systems to optimize performance and reduce inference costs. RAG systems retrieve relevant documents from an external datastore to augment a language model's response, but incorporating full documents as context can quickly lead to excessive token usage, exceeding the model's context length limits and increasing processing time [78]. Instead of simply concatenating numerous documents, context compression selectively simplifies information to minimize noise and highlight essential data, allowing the language model to focus on the most relevant content [78].

Several strategies for context compression have proven effective. One method employs small language models (SLMs) to filter out less important tokens in order to create a compressed prompt. Although the compressed result might seem disjointed to humans, it remains interpretable by LLMs and achieves compression without requiring further LLM

training. Other methods train information extractors to identify relevant content within large documents [19].

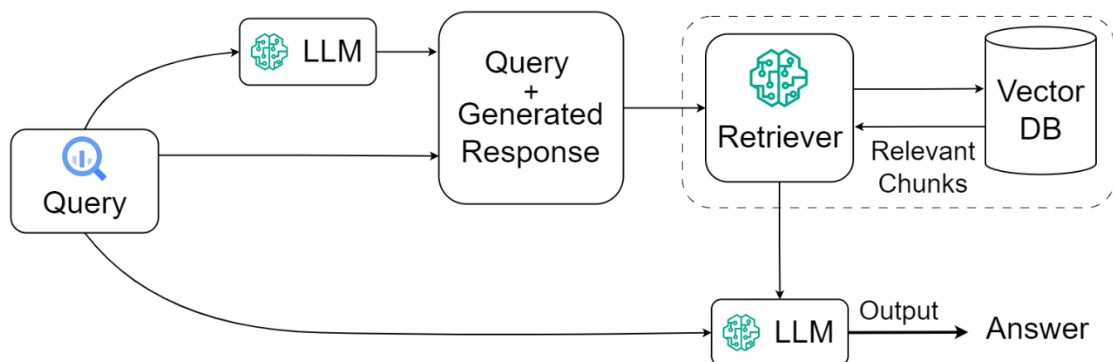
Combining document reduction with context compression further improves model accuracy [19]. The "Filter-Re-ranker" paradigm, for example, uses SLMs as filters and LLMs as re-rankers to prioritize relevant content. LLMs can also evaluate and critique retrieved content before generating a response, discarding irrelevant information and focusing on key details. These compression techniques are crucial for balancing relevance, token limits, and processing costs in RAG systems while maintaining language integrity [19].

Query Expansion and Re-writing

In RAG systems, query expansion enhances retrieval accuracy by adding relevant context to the input query [19]. Hypothetical Document Embeddings (HyDE) achieves this by using LLMs to generate hypothetical contexts, which are then embedded with the original query to improve retrieval precision (Figure 3). This method is particularly beneficial when limited labeled data or explicit knowledge is available, enabling RAG systems to create richer embeddings by incorporating potential relevant details [79].

Figure 3

Query Expansion Using HyDE



However, query expansion using LLMs, including methods like HyDE, can sometimes introduce inaccuracies or hallucinations—where hypothetical content diverges from factual information—especially if the LLM lacks knowledge of the query topic. To address this, multi-query approaches expand the initial query into several targeted queries, while sub-query methods break down complex questions into simpler prompts [19].

1.1.5 Agents in Large Language Models

LLM agents are designed to autonomously perform tasks by leveraging the models' reasoning abilities [38]. These agents can interpret complex instructions and execute multi-step processes, acting as intermediaries between users and computational resources [39]. Agents demonstrated their adaptability and potential for transformative impact in various fields, including social sciences, natural sciences, and engineering [40]. However, challenges remain in maintaining consistent performance, ensuring accurate contextual understanding, and achieving seamless integration with external tools [38]. Ongoing research focuses on improving the agents' autonomy, reliability, and human-like interaction [80].

Agentic RAG leverages LLM agents to enhance the retrieval of information and overcome traditional RAG systems' limitations [15] by incorporating intelligent agents that dynamically select and process relevant information, thereby improving response accuracy and contextual relevance [35]. This approach enables the system to handle complex, multi-step reasoning tasks more effectively than naive RAG, which lacks such dynamic decision-making capabilities [37]. Furthermore, by leveraging agents with access to various tools, Agentic RAG can route queries to specialized knowledge sources, leading to more accurate and contextually appropriate responses [36]. In contrast, naive RAG systems directly use retrieved information without additional processing, which may result in coherence issues in generated responses [15].

1.2 Literature Review

QA systems research began in the early 1960s, marking a significant area of study within natural language processing [81]. Early efforts often adapted rule-based methodologies, exemplified by the system proposed by Riloff and Thelen in 2000 [82]. However, their work [82] highlighted several inherent limitations of this approach, including the resource-intensive nature of rule creation, sensitivity to variations in wording and sentence structure, and limited ability to draw inferences. Further challenges arise from coreference resolution, contextual interpretation, and ambiguity management. Scalability is also a constraint, and errors from earlier processing stages can cascade, negatively impacting overall system performance. The manual crafting required by rule-based systems complicates their maintenance and updates, and their inability to generalize can lead to performance decline with larger datasets [83].

In the pre-transformer era, machine learning and NLP techniques were combined in various models to address the challenges of rule-based systems. Poon et al. [84] advanced machine reading by combining probabilistic reasoning with NLP to infer meaning from text, aiming to improve structured knowledge extraction from unstructured data using statistical methods. Lende and Raghuwanshi [85] similarly used NLP techniques like part-of-speech tagging, named entity recognition, and syntactic pattern matching to build question-answering systems for educational texts, focusing on interpreting and retrieving relevant information.

IR techniques were employed for extracting relevant text segments from extensive document collections in addition to machine learning-driven methods in order to enhance the overall performance. Dwivedi and Singh [86] conducted a comprehensive review of question-answering systems showing how IR systems employed different techniques such as document indexing, keyword matching, and ranking algorithms to locate relevant information. The field was further advanced by the development of probabilistic IR models and the introduction of statistical methods for relevance estimation [87]. Early efforts, such as vector space models [88] and latent semantic analysis [89], laid the groundwork for more sophisticated retrieval systems. Despite their contributions, these early systems were often constrained by their reliance on complex NLP pipelines that required accurate linguistic annotations and handcrafted rules. This dependency not only increased the computational overhead but also limited their scalability and adaptability across diverse domains and languages, as emphasized by Lende and Raghuwanshi [85].

Since the introduction of the Transformer architecture [5] and the introduction of BERT [6], a few systems were introduced as extractive question answering systems [90–93]. The methodologies [90–93] involved fine-tuning BERT or its variants (like RoBERTa) on domain-specific datasets, often with enhancements like hybrid architectures, semantic layers, or de-biasing strategies to address specific challenges in extractive QA. Most focus on leveraging transformer-based embeddings for efficient answer extraction. A common set of challenges emerges across research on extractive question answering using BERT and its variants. Many models exhibit domain and dataset dependence, performing well only on datasets similar to those they are fine-tuned on, which limits their generalization ability to diverse or unseen contexts [91,94]. Bias issues, such as position bias, cause models to overly rely on the location of answers within passages rather than their content [92]. Additionally, advanced architectures and mechanisms for improving performance

often come with increased computational overhead, making them less practical for real-world applications [90]. Language limitations are significant, particularly in low-resource languages, where models struggle due to insufficient pre-training data and linguistic complexity [91]. Finally, models often face challenges with contextual understanding, such as handling complex sentence structures or reasoning across multiple sentences, reducing their effectiveness in answering complex queries [90,93].

A methodology proposed by Alkhurayyif and Sait [95] for Arabic question answering involves data preprocessing, named entity relationship identification using the Multinomial Naïve Bayes algorithm and Named Entity Recognition (NER), and response retrieval leveraging ELMo embeddings and Quaternion Long Short-Term Memory (QLSTM) networks. However, the system faces several challenges, including sensitivity to real-world variability, difficulty handling complex Arabic sentence structures, limited support for non-textual queries, and inefficiencies in training time. The complex morphology of Arabic, particularly verb-noun patterns, hinders contextual understanding, and scalability remains a concern due to its design being tailored for Arabic-specific tasks.

Building on the foundational advancements of BERT, LLMs emerged as a transformative innovation in NLP, characterized by their unprecedented scale and capability [96]. These models significantly enhanced their ability to understand and generate contextually relevant text, achieving remarkable performance across a wide array of NLP tasks [97]. The adoption of transformer-based architectures, as outlined by Gillioz et al. [98], laid the foundation for these advancements, enabling LLMs to handle tasks such as machine translation, text summarization, and sentiment analysis with unprecedented accuracy and scalability. The vast pre-training of LLMs on diverse datasets across multiple domains has provided the models with extensive general knowledge, making them highly effective for open-domain applications. Kamaloo et al. [99] demonstrated LLMs' ability to provide contextually rich and accurate answers in open-domain question answering, leveraging their training on large datasets. However, despite their capabilities, LLMs are prone to hallucinations, particularly in specialized and closed-domain scenarios [100].

While LLMs exhibit impressive capabilities, they are prone to hallucinations, particularly in specialized and closed-domain contexts. Fine-tuning with domain-specific data has been shown to enhance performance by improving knowledge representation and accuracy within those domains. For instance, Guo and Hua [101] employed continuous

training and instruction fine-tuning to adapt Meta's Llama 2 base models to the Chinese medical domain, achieving performance comparable to OpenAI's GPT-3.5-turbo in medical question answering. Similarly, Singhal et al. [102] proposed fine-tuning LLMs with specialized medical datasets to improve their ability to address expert-level medical inquiries accurately. Huang et al. [103] further demonstrated this principle by introducing "Lawyer LLaMA", a legal domain LLM that incorporates domain knowledge through continuous training and acquires professional skills via supervised fine-tuning, effectively adapting the model to legal-specific queries. These fine-tuning approaches enhance domain accuracy, challenges remain, notably in supporting multilingual queries. Many LLMs are initially trained on primarily English datasets, which may limit their ability to handle languages with less training data [104]. The research by Xu et al. [105] discusses significant limitations in multilingual LLMs, such as language imbalance and multilingual alignment, which can decline the performance in low-resource languages.

Additionally, LLMs require frequent updates to their knowledge base to keep up with rapid advancements and evolving information in specialized domains. Since the knowledge in an LLM is "learned" within its parameters, updating the model's knowledge typically involves computationally expensive retraining or fine-tuning [72] and LLMs may struggle to integrate new factual knowledge effectively, instead reinforcing their pre-existing knowledge, which can lead to increased hallucinations [106]. Chang et al. [107] highlight that adding multilingual data can improve low-resource language modeling performance, but as dataset sizes increase, adding more data may begin to decrease the performance due to limited model capacity, known as the "curse of multilinguality" [107,108]. This limitation poses significant challenge in making LLMs fully reliable and adaptable across diverse and dynamic specialized domains.

RAG has emerged as a promising alternative to traditional fine-tuning approaches for addressing the limitations of parameter-based language models [15]. While fine-tuning involves adapting a model by updating its parameters on specific datasets, it is often constrained by the static nature of learned knowledge and the risk of forgetting. In contrast, RAG dynamically integrates external knowledge sources, allowing models to access and leverage up-to-date information beyond their initial training data [15,19]. Several question-answering systems have been proposed using RAG, demonstrating its adaptability. For example, the English question-answering system DPR-RAG [109] integrates dense passage retrieval to enhance the accuracy of generated responses,

offering significant improvements over traditional approaches. However, its reliance on dense embeddings can result in poor generalization to domains or languages with limited training data, and irrelevant passages retrieved during the process can degrade output quality.

A research focusing on Islamic question answering is MufassirQAS [110], which employs RAG to enhance Arabic question answering, particularly within Islamic studies domain by integrating a vector database of Turkish-translated Islamic texts and prompt engineering to semantically search for relevant information and provided it to the LLM. Nonetheless, the authors didn't show any evaluation criteria, metrics, or scores in their research to reflect the system's performance, although they mentioned that the system's effectiveness is limited when faced with larger contexts. Integrating and summarizing a large number of context chunks can disrupt coherence and the flow of information, resulting in fragmented responses. This challenge suggests a need for improvements in connecting and synthesizing retrieved data to produce cohesive outputs.

Existing research in QA systems demonstrates significant advancements but also reveals notable limitations. The MufassirQAS LLM [110] struggles with larger contexts, leading to fragmented responses and limited applicability to broader questions. Similarly, the Arabic QA system leveraging deep learning techniques [95] shows strong performance but suffers from scalability issues, inefficiencies in handling complex Arabic morphology, and a dependency on extensive training and fine-tuning processes. Lastly, there is a noticeable lack of generic QA systems that can seamlessly adapt to new knowledge bases without requiring significant reconfiguration or retraining, highlighting the limited flexibility and scalability of current approaches.

1.3 Problem Statement

LLMs like OpenAI's ChatGPT and Meta's Llama are often pre-trained on vast datasets of internet text enabling the models to learn world information, grammatical phrasing, vocabulary, and lingual context [111]. While these LLMs impress with their natural language abilities [9], they encounter distinct challenges regarding query understanding or generating response when applied in some specialized domains. The challenges are caused by: the broad dataset they were trained on, conflicting facts in the dataset, or outdated information, among other factors [24]. These challenges are even increased when the LLM is dealing with complex languages like Arabic [112].

The phenomenon of generating nonsensical, inaccurate, or factually-incorrect responses is known as "hallucination" [24]. In the field of question-answering, some sensitive domains like legal counseling or Islamic fatwa, such inaccuracies are not tolerated. In Islamic fatwa for instance, where interpretations of Qur'an and Hadith verses are crucial, even minor errors are unacceptable. Additionally, user queries may involve culturally-specific and geographically-dependent matters. For instance, one may ask about the permissibility of using a specific local bank or investing in a regional company. Such inquiries require a deep understanding of local culture, including events and institutions. While current LLMs trained on localized data show some cultural awareness, they often lack in-depth comprehension, hindering accuracy in specialized topics [113]. Moreover, LLMs like ChatGPT and Gemini often use content filtering to avoid sensitive, ethical, or potentially harmful topics [114]. While intended for responsible use, these filters can limit the scope and effectiveness of LLMs for specialized applications like Fatwa inquiries, where culturally specific responses are crucial.

Fine-tuning LLMs for specialized domains is the traditional solution. However, the enormous parameter counts of these models requires significant computational resources for fine-tuning. Robust LLM that can understand the text consists of billions of parameters, for example OpenAI's GPT-3 models have over 175 billion parameters, Google's PaLM has over 540 billion parameters [7], and Meta's Llama 3.1 family has 405 billion parameters in some models [115]. Training the GPT-3 davinci model took over 3 years and it cost about 4 million USD to train its 175B parameters. OpenAI's GPT-4 on the other hand, has trillions of parameters, took about 3 years and 6 months to train and it cost about 90 million USD given the development in GPU chips which became more computationally powerful [116].

Even in smaller models like Meta's Llama 3.1 8B, fine-tuning can be resource-intensive. Techniques such as PEFT, LoRA, and adapter modules reduce computational requirements for the fine-tuning process, yet the process can still be time-consuming on systems with limited resources and requires high end hardware [117]. For instance, full fine-tuning of Llama 3.1 8B typically requires approximately 16 GB of GPU memory, which is manageable on a high-end consumer GPU [115]. Implementing techniques like LoRA and adapter modules further reduces memory requirements, making fine-tuning feasible on GPUs with even lower memory capacities. However, these methods often introduce additional complexity due to the need for careful hyperparameter tuning, which

can impact performance if not optimized [118]. Additionally, these approaches may struggle to generalize beyond the specific tasks or domains they were fine-tuned for, limiting their flexibility [117]. Parameter-efficient fine-tuning also typically retains the original base model's limitations, such as vulnerability to hallucination in contexts outside the fine-tuned scope [113].

Another concern in fine-tuning is the structure of parameterized LLMs makes it difficult to predict performance before the fine-tuning process is completed [15]. More importantly, modifying learned data remains an ongoing research challenge [120]. This often requires complete re-tuning, as there is currently no established method for LLMs to "unlearn" outdated information effectively [121]. Additionally, parameterized-memory LLMs lack the inherent capability to provide references for their generated responses, which limits their reliability in producing verifiable outputs [15,122].

This study introduces a generic Arabic question-answering framework capable of answering questions across various domains and adapting to any textual dataset without requiring fine-tuning. The framework processes questions posed in Arabic and generates a response consists of two key elements:

1. Answer: the generated answer to the user's question.
2. Evidence (if applicable): Supporting references, which may include Qur'anic or Hadith verses in the context of Islamic fatwa, or relevant legislation for legal counseling. This component is provided only when the question requires evidence to support the generated answer.

1.4 Aims of Study

This study aims to introduce Nebras, a novel Arabic question-answering system designed exclusively for specialized domains. Nebras addresses the unique requirements of Arabic language applications while providing a scalable, efficient, and context-aware solution for generating accurate, factually correct answers. The system's design reflects several contributions that distinguish it from traditional systems.

One of the key contributions of Nebras is its adaptability. The system's knowledge base can be expanded and managed dynamically. Administrators can add new textual datasets tailored to specific domains, map relevant fields from the dataset for indexing and

retrieval, and modify the knowledge base by updating or deleting existing documents as needed. This design eliminates the need for costly and time-intensive model fine-tuning and allows Nebras to seamlessly adapt to various domains while maintaining an up-to-date and relevant knowledge base.

Nebras employs a hybrid retrieval approach that combines multiple techniques to ensure relevant and accurate information retrieval. Its implementation leverages the Agentic RAG framework where specialized agents collaborate to process tasks effectively.

Nebras addresses the unique challenges of the Arabic language by prioritizing linguistic and contextual precision. This study demonstrates its potential to advance question-answering for Arabic-speaking users in specialized domains by proposing a robust, scalable, and adaptable solution to meet diverse application needs.

1.5 Hypotheses of Study

The following hypotheses have been formulated to guide the research, evaluate the effectiveness of Nebras, and address the challenges identified in the problem statement. Nebras is a novel Agentic RAG-based system introduced in this study, designed for complex Arabic question answering across specialized domains.

1.5.1 Accuracy Hypothesis

Nebras is expected to outperform existing LLMs in providing accurate and contextually relevant answers to Arabic queries in specialized domains. Nebras enhances response reliability and reduces hallucinations (factually incorrect yet plausible-sounding answers) by dynamically integrating external textual knowledge bases. Nebras's performance will be compared to a baseline established by evaluating responses generated by several top-performing models, using both automatic metrics and expert evaluations on domain-specific datasets.

1.5.2 Adaptability and Scalability Hypothesis

Nebras's dynamic knowledge base management system allows seamless expansion and modification which enables the system to incorporate new datasets for emerging domains without fine-tuning. Nebras offers a scalable and cost-effective Arabic question-answering solution deployable even in resource-constrained environments by leveraging

pre-trained LLMs. This will be validated by analyzing Nebras's computational efficiency (memory and processing requirements) and its scalability to large datasets.

1.5.3 Language-Specific Performance Hypothesis

Nebras is hypothesized to outperform existing systems in handling the linguistic complexities of Arabic (morphology, syntax, and dialectal diversity). Its design is expected to give it an advantage over general-purpose LLMs when processing complex Arabic queries. This will be evaluated using both lexical and semantic metrics, and contextual relevance ratings.

Chapter Two

Methods

This chapter details the development and implementation of the Nebras, explaining the design choices, algorithms, and frameworks used to process Arabic queries across multiple domains. It covers the hybrid retrieval approach, Agentic RAG framework integration, use of pre-trained LLMs, and the system's dynamic knowledge capabilities. Subsequent sections explore implementation aspects, highlight its key components, and clarify the system's operation.

2.1 Data Collection

This research focuses on two distinct domains: Islamic Fatwas and University Information Help Desk. Each domain presents unique challenges and requires specific data structuring. The Islamic Fatwa domain, due to its sensitive nature and the need for accurate, well-supported answers, demands careful handling. Fatwas require precise reasoning and the inclusion of correct and specific daleel (evidence) aligned with Islamic jurisprudence. This daleel, often from the Qur'an or Hadith, must be accurately referenced; the system must not generate or fabricate it. This makes the domain an ideal choice for testing the system's ability to generate contextually accurate responses with necessary supporting information. Additionally, this domain allows for the evaluation of the system's performance with question-answer data structures, as fatwa data inherently follows this format.

Data for the Islamic Fatwa domain is sourced from two reputable websites:

1. NNU Fatwa Website (<https://fatwa.najah.edu/>): Maintained by the Faculty of Shari'a at An-Najah National University, a respected institution in Islamic scholarship, this website contains over 1,500 fatwas. These address local issues and inquiries specific to Palestine, making the data highly relevant and credible.
2. Islamweb.net (<https://islamweb.net/>): Overseen by the Ministry of Endowments and Islamic Affairs in Qatar, Islamweb.net offers a vast repository of over 160,000 fatwas. Its credibility and reliability make it a valuable resource for this research.

These carefully selected sources and their credible data sources ensure a robust evaluation of the proposed system’s ability to handle diverse and complex Arabic queries, contributing to the development of Nebras, the proposed question-answering framework.

2.1.1 Islamic Fatwa Dataset Collection

To gather data from these sources, a custom scraping template is developed for Islamweb.net to capture specific fields essential for the system responses. These fields, as explained in Table 1 and highlighted in Appendix D - Figure D11, include categories, fatwa topic summaries, unique fatwa identifiers, dates, questions posed, and the detailed answers provided by muftis. Each entry is appended with a static "source" field with the value "islamweb.net" to facilitate data source tracking.

Table 1

Islamweb Fatwa Fields Mapping

Field	Description	Field Name	Data Type
1	Fatwa categories can be extracted from fatwa breadcrumbs.	categories	Array
2	Fatwa topic. A short summary of what this is about.	topic	string
3	Fatwa id.	fatwa_id	integer
4	Fatwa date in Gregorian and Hijri.	date	string
5	Fatwa question asked by a user.	question	string
6	Fatwa answer from muftis at islamweb.	answer	string

The initial scraping of Islamweb.net resulted in a dataset of 164,310 fatwas before pre-processing. For the NNU Fatwa website, similar scraping techniques were used to obtain 1,500 locally relevant fatwas, enriching the dataset with inquiries specific to the Palestinian context.

2.1.2 University Help-desk Dataset Collection

The dataset for the university help-desk dataset is collected exclusively from NNU and comprises information specific to NNU. The NNU dataset, by contrast, belongs to a different knowledge domain, focusing on academic and institutional information represented in a document-based, semi-structured format rather than a question-answer structure. This dataset includes factual information about majors, courses, admission requirements, and faculty members, allowing the evaluation of the ability of the proposed

system to retrieve and generate accurate responses from semi-structured documents. The process of collecting data for the NNU dataset involves parsing of web pages from the university's official site (<https://najah.edu>), with additional details on the scraping techniques and categorization to be explained in the subsequent subsections.

However, the NNU website's protection measures prevent automated data scraping, making it difficult to automatically gather information across all academic programs. Consequently, a manual data collection process is implemented for a selected group of medical and IT-related majors. Due to the time-intensive nature of manually gathering data for over 200 majors, this selective approach is considered necessary for the research.

Academic Data Collection

The process of collecting academic data focused on obtaining detailed information regarding the academic majors available at An-Najah National University at both the undergraduate and graduate levels. This data includes details for each major, covering its title, affiliated faculty, academic degree (e.g., Bachelor's, Master's, Doctorate), duration of study, and a corresponding URL. The specific data fields extracted during the collection process are illustrated in Table 2 and highlighted in Appendix D - Figure D12.

Table 2

NNU Academic Majors Fields Mapping

Field	Description	Field Name	Data Type
1	Academic major title	major_title	string
2	Major's faculty	faculty	string
3	Academic degree	degree	Enum (bachelors, master, doctorate)
4	Major's study duration	duration	string
5	Major's info URL	url	url
6	Document type	doc_type	string

The description of each academic major is displayed on its own dedicated page, necessitating a visit to the page for data extraction. For selected majors in medical and IT-related fields, as a sample, these academic major pages were manually visited to retrieve their descriptions. While visiting each academic major's page, information related to the major curriculum and courses is collected. Table 3 shows information about the collected fields. The fields are also illustrated in Appendix D - Figure D13 for further reference. The "doc_type" field holds the document's category and set manually while

collecting the data. The values used to distinguish the documents are: "admission", "staff", "academic_major", and "academic_course" for majors, courses, staff, and admission documents, respectively.

Table 3

NNU Academic Courses Fields Mapping

Field	Description	Field Name	Data Type
1	Plan version	version	string
2	Curriculum section	section	string
3	Course number	course_number	integer
4	Course title	title	string
5	Credit hours	credit_hours	string
6	Course prerequisites	prerequisites	string
7	Course description	description	string
8	Document type	doc_type	Enum (academic_major, academic_course, staff, admission)

Administration-related Data Collection

The collection of administrative data is limited due to the same protection measures mentioned earlier, information regarding the university's governance and faculty deans is manually obtained. Some information regarding admission and acceptance is found on Nawarat An-najah (<https://nawarat.najah.edu/>), a subdomain for NNU for newly registered students.

By preparing these two datasets, the study can assess the QA system's performance in generating answers in different domain with different dataset structures.

2.2 Data Pre-processing and Structuring

Effective data structuring is essential for compatibility with the proposed system. Due to the distinct nature of the Islamic Fatwa and NNU datasets, each requires specific processing steps for organization and standardization. The following subsections detail the structuring methodologies for each dataset.

2.2.1 Islamic Fatwa Dataset Pre-processing and Structuring

This section details the cleaning and preprocessing steps applied to the Islamic fatwa data. These steps removed noise, inconsistencies, and irrelevant information to enhance the dataset's quality. The following subsections provide a comprehensive overview.

Missing Values and Exploratory Data Analysis

Data completeness is assessed by checking for missing values (NaN, whitespace-only values, or zeros in integer columns). The only column with missing values was the "topic" column with count of 12 missing values. To mitigate the impact of missing data on subsequent analysis, these missing values were replaced with human-generated descriptions derived from a careful analysis of the corresponding fatwa's context. To gain insights into the textual content of the dataset, an exploratory data analysis (EDA) is conducted on the "topic", "question", and "answer" fields. The text in these fields is tokenized by splitting it into individual words using single space delimiter, allowing for an analysis of token counts. Table 4 presents the maximum, minimum, and mean token counts for each field.

Table 4

Token Counts from the Scraped Islamweb Dataset

Field	Max. Count	Min. Count	Mean
topic	50	1	7.68
question	2,191	2	78.39
answer	6,076	2	211.44

The "question" and "answer" columns exhibited a minimum token count of 2, which raised concerns about the potential presence of very short or incomplete entries. Further inspection of these low-token-count entries was postponed until after the text cleaning process.

Muftis often begin fatwas with introductory sentences that do not contribute directly to the factual content or reasoning of the answer. The length of these messages can influence the chunking process and potentially impact the retrieval process.

To extract introductory sentences, the line breaks in the scraped fatwas are used. These sentences are generally found in the first line, although this pattern is not consistent across

all fatwas. By examining the first lines, the frequent introductory sentences are extracted through the following steps:

- Splitting each fatwa by line breaks and the HTML `
` tag.
- Extracting the first non-empty text result from the split array, excluding empty HTML tags or whitespace.
- Calculating the frequency of each extracted sentence across the dataset.

This process identified a total of 17,034 unique introductory sentences. The sentence “الحمد لله والصلاة والسلام على رسول الله وعلى آله وصحبه أما بعد” appears most frequently in the entire dataset, with a total of 138,281 occurrences (84%). The second most frequent introductory sentence is “الحمد لله والصلاة والسلام على نبينا محمد وعلى آله وصحبه ومن والاه أما بعد” which appears in 4,241 fatwas but have the same leading and trailing words.

Text Cleaning

The text cleaning process involved several steps: stripping HTML tags, removing non-Unicode characters, and eliminating duplicate tokens. Line breaks were preserved at this stage, as they might be beneficial in later processing steps.

Removing Introductory Sentences

Introductory sentences in fatwas do not contribute directly to the factual content or reasoning of the answer and usually follow a structural pattern. To remove them without affecting the context or valuable information, the process involves: extracting starting and trailing words from each introductory sentence to construct a pattern, identifying the first sentence in each fatwa, and removing it if it matches the pattern. Further analysis of these words, as detailed in Table 5, reveals that “الحمد” is the most frequent starting word in introductory sentences (98.5%). The trailing words for sentences beginning with “الحمد” are also examined, with results presented in Table 5.

Table 5*Frequency of Start and Trailing Words in Introductory Sentences in Fatwa*

Word	Frequency	Percentage
Starting Words		
الحمد	160,870	98.5%
خلاصة	2,262	1.39%
الخلاصة	89	0.054%
Trailing Words for Sentences Strating with "الحمد"		
بعد	84,580	51.8%
بعد	58,699	35.9%
أعلم	13,814	8.45%
الفتوى	2,268	1.39%
وبعد	580	0.36%

After removing the introductory sentences, the frequency of the starting words is recalculated and displayed in Table 6.

Table 6*Frequency of First Words After Removing Introductory Sentences in Fatwa*

Word	Frequency	Percentage
فإن	27,234	16.67%
الحمد	17,403	10.65%
فقد	14,680	8.99%
فلا	13,837	8.47%

Although the word "الحمد" maintains a relatively high frequency, likely it is part of the fatwa answer, completely eliminating it could potentially compromise the accuracy of the factual context within the fatwa.

Subsequent Text Cleaning and Re-evaluation

Following the removal of introductory messages, the dataset is re-evaluated for missing values and token counts to assess the impact of text cleaning steps on token counts. The

evaluation showed no missing values in any dataset columns. The results for token counts are presented in Table 7.

Table 7

Token Counts from the Scraped Islamweb Dataset After Cleaning

Field	Maximum Count		Min. Count	Reduction
	Before Cleaning	After Cleaning		
topic	50	50	1	0
question	2,191	2,119	1	3.29%
answer	6,076	5,436	3	10.53%

A manual review of questions comprising only a single word revealed a typographical error in one query. The original, malformed question was:

قال الله ألا تزرر. وازرة. وزير. أخرى. فكيف. يتفق. ذلك. مع. الشفاعة. ومع. إهداء. ثواب. الأعمال. إلى. الأموات؟

Given the context and grammatical structure, it is clear that this query represents a user input error. The question was subsequently corrected to accurately reflect the intended meaning by removing the periods and replace them with spaces.

For answers with fewer than 10 words, it was observed that some referenced other fatwas using phrases like:

– سبق برقم: 1640، ورقم: 10794 .

– فتفصيل ذلك سبق في الفتوى رقم: 32689 فتراجع .

This suggests that extracting fatwa references from answer texts could be valuable. Regular expressions can be employed to identify these reference numbers.

The references 1640, 10794, and 32689, mentioned in the short fatwa answers, were not found within the dataset and are subsequently removed.

HTML Tag Removal and Text Formatting

An analysis of randomly selected fatwas revealed that HTML tags were primarily used for text formatting and coloring, without a standardized structure for different fatwa components such as daleel (evidence) or references. This inconsistency made it impractical to leverage HTML formatting for information extraction.

Category Extraction

The final step in text pre-processing involved categorizing the fatwas to provide a structured taxonomy for subsequent analysis and retrieval. The category names were extracted from the fatwa breadcrumbs (Field 1 in custom scraper template), and the word “الرئيسية” (main) is removed to streamline the categorization process.

The first two words of the category names are retained to represent the main and secondary categories, respectively. This hierarchical categorization approach provides a more granular understanding of the fatwas' topics. The distribution of fatwas across the identified categories is visualized in Appendix E - Figure E14.

2.2.2 NNU Dataset Pre-processing and Structuring

The following subsections outline the structuring process for each NNU document type, including academic majors, academics, courses, staff, and admissions.

Academic Majors Document Structuring

For academic major documents, additional data from other scraped fields is merged into the “description field” (as shown in Table 2) to create an entry suitable for indexing. These fields include “major_title”, “faculty”, “degree”, and “duration”. By enhancing the “description” with relevant information from these fields, it becomes more concise and informative. The resulting enriched text is stored in a newly created field, "content," specifically designed for indexing and similarity searches. The template for the "content" field is provided in Appendix F.

Staff and Admission Document Structuring

No augmentation is applied to these documents; their content is indexed directly in the vector database. Staff documents consist solely of staff member biographies, while admission documents outline the admission rules for new student enrollment.

Generating “topic” for NNU Documents

A "topic" field is added to each collected document, providing a brief summary about the document's content. This summary is generated using a LLM with the following prompt:

“حلل المستند التالي واستخرج وصفًا قصيرًا للموضوع الذي يلخص الفكرة أو التركيز الرئيسي للمستند.

يجب أن يكون الوصف موجزًا، ويفضل أن يتراوح بين 20-50 كلمة، ويعكس الموضوع الأساسي للنص.

Which translates to: “Analyze the following document and extract a brief description of the topic that summarizes the main idea or focus of the document. The description should be concise, ideally between 20-50 words, and should reflect the core subject of the text. Here is the document: {text}”

The “topic” field is also indexed in the vector database, resulting in each document having two embedding vectors: one for the *content* field and another for the “topic” field.

The data structuring process standardizes and optimizes both the Islamic fatwa and NNU datasets for seamless integration with the QA system. By tailoring the structuring methods to the unique requirements of each dataset, this process ensures that the system can effectively handle both question-answer data as well as document-based datasets. The structured datasets, with enriched content fields, facilitate accurate and efficient similarity searches and response generation, setting a solid foundation for the subsequent stages of system experimentation and evaluation.

2.3 Implementation

The following subsections delve into knowledge base preparation and the answer generation pipeline.

2.3.1 Vector Database

To ensure adaptability across various textual datasets, a mapping component is proposed for dataset preparation in the QA pipeline. Textual datasets serving as knowledge bases for question answering can be categorized into:

1. *QA-based*: Datasets comprising paired questions and answers.
2. *Document-based*: Datasets containing documents on specific topics (e.g., PDFs).

For QA-based datasets, essential fields include:

- Document ID: Unique identifier for each document.
- Question: The posed query.
- Answer: The corresponding correct answer.
- Topic: A concise description of the question's subject.

The “question”, “answer”, and “topic” fields are vectorized and indexed in the vector database, while additional fields serve as metadata. For document-based datasets, required fields are:

- Document ID: Unique identifier for each document.
- Content: Full document content.
- Topic: A brief summary of the document's content.

The “content” and “topic” fields are vectorized and indexed while any extra fields act as metadata.

In regards to the vector database collection, each collection must have the following metadata fields:

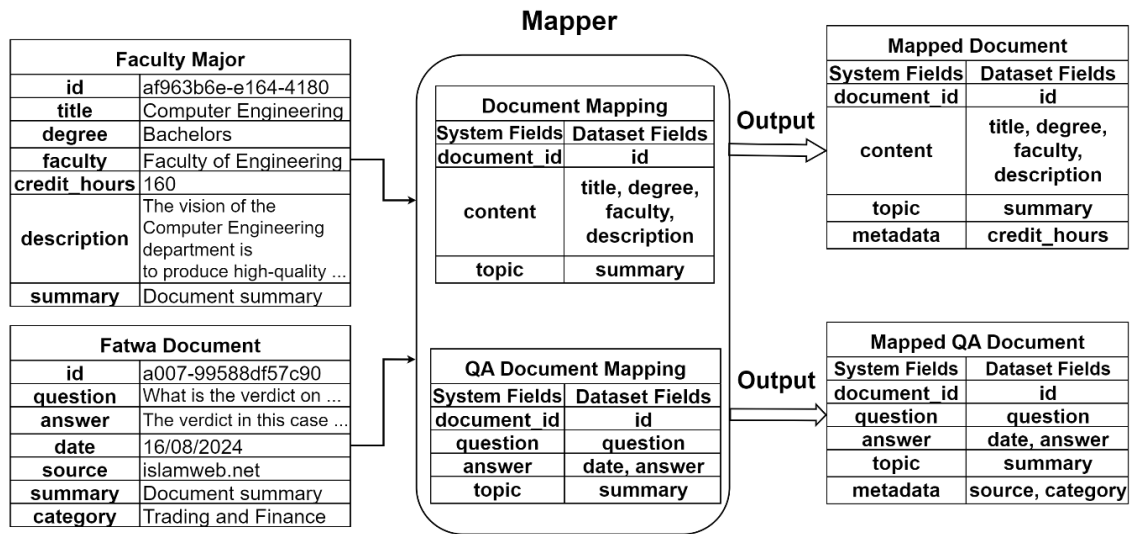
- Title: A human-readable collection name.
- Description: Summary of the collection’s content.
- Type: Dataset type ("qa" or "docs").
- ID Field: Field name for uniquely identifying documents.

A collection’s metadata example:

```
{  
  "name": "nnu",  
  "title": "An-Najah National University",  
  "description": "This collection includes comprehensive documents about An-Najah National University, covering admission processes, academic programs, and general university details.",  
  "type": "docs",  
  "id_field": "document_id"  
}
```

By providing the necessary metadata, the system can efficiently manage and integrate new collections into its knowledge base. Users need only map relevant fields and provide collection metadata to incorporate datasets seamlessly as illustrated in Figure 4.

Figure 4
Dataset Field Mapper



2.3.2 Indexing and Chunking

During the indexing phase, documents are processed, segmented, and converted into embeddings, which are subsequently stored in a vector database. The quality of the index construction is critical in determining whether the appropriate context can be accurately retrieved during the retrieval phase [19].

The indexing structure introduced in this research utilizes a Hierarchical Index Structure with attached metadata. This hierarchical approach addresses the issue of context mismatch that arises when retrieved chunks are semantically incomplete.

Determining an optimal chunk size is a delicate process that requires balancing considerations. Chunks that are too long introduce noise to the embedding model and requires more processing. Additionally, if a chunk exceeds the model's maximum input length, it will be truncated, leading to loss of meaning. In contrast, chunks that are too short may prevent the embedding model from properly capturing the context. Incorporating a hierarchical index enhances retrieval allows the model to reconstruct the context [15,19].

The process Involves the following steps:

- Assigning a unique identifier to each document (if not already provided).

- Split the documents into smaller, fixed-size chunks. Mapping each chunk to its originating document, thereby creating a parent-child relationship (hierarchy).
- Attaching metadata to documents enhances the filtering process and enriches the document’s content.

Achieving an appropriate chunk size involves a series of tests across different models and chunk sizes to find an effective balance which is illustrated in “Chunking Optimization” section.

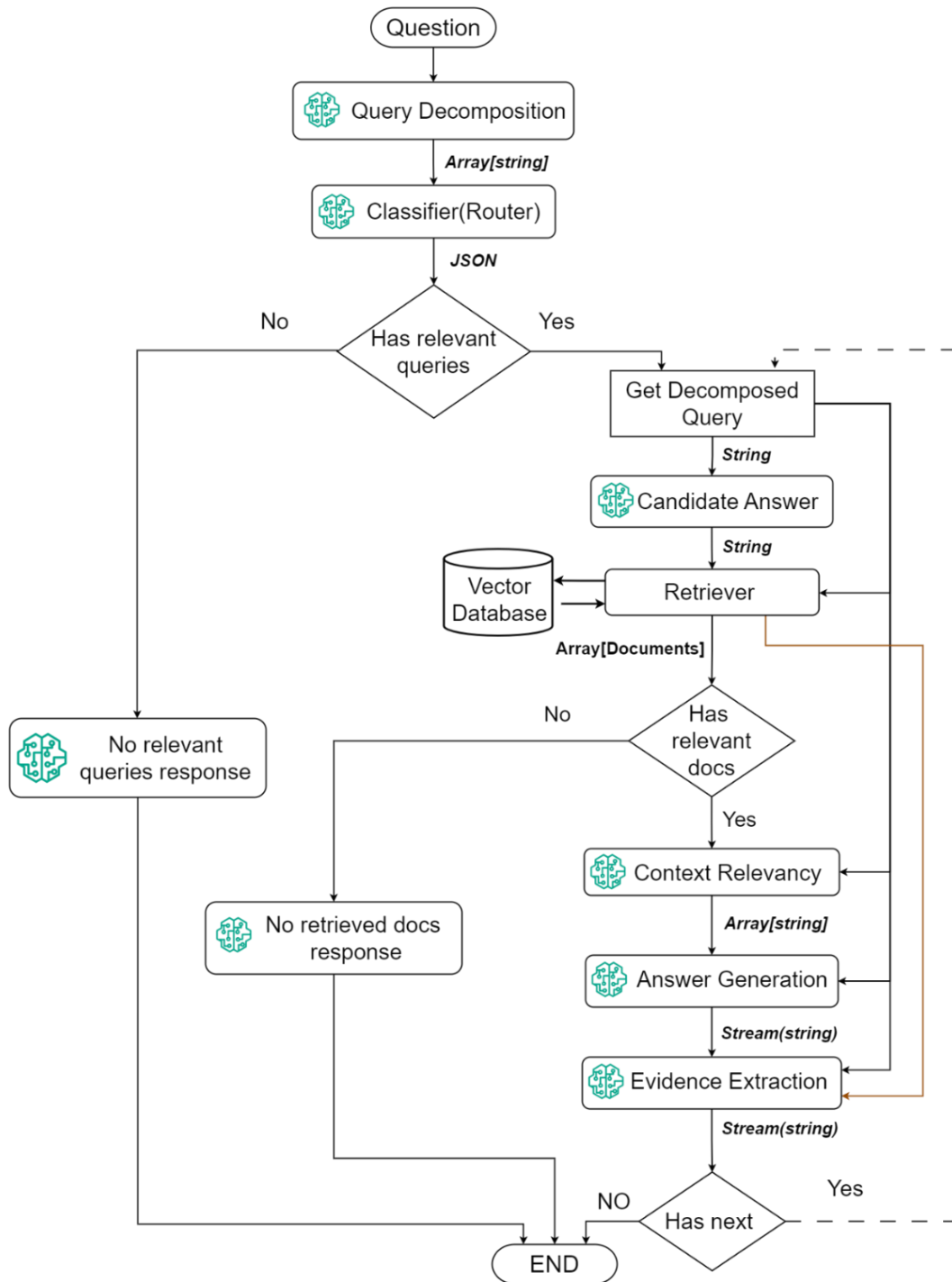
2.3.3 QA Pipeline

Traditional RAG frameworks operate in two main steps: retrieving relevant documents and generating responses based on these documents [15]. While effective for straightforward queries, standard RAG often lacks the capacity for complex reasoning or task decomposition across multiple nodes [36].

Nebras’s QA pipeline is implemented using an Agentic Retrieval-Augmented Generation approach leveraging a Graph-workflow framework for structuring nodes within the pipeline, where each node (or agent) has a very specific task. The framework organizes the workflow by enabling different agents and models at each step, handling discrete tasks rather than dynamically retrieving graph-based data. The pipeline workflow is illustrated in Figure 5.

The pipeline utilizes a series of specialized agents to ensure the reliability and precision of the responses. The process begins with Query Decomposition Agent, where complex questions are handled by breaking them down into smaller, well-defined, and concise sub-questions. These are then processed by the Query Classifier Agent to determine the type of question being asked. Next, the Candidate Answer Generation agent formulates potential answers, which are further refined by the Retriever to identify the most relevant documents. The Context Relevancy agent assesses the retrieved information to ensure its applicability to the query. Finally, the Answer Generation agent constructs the final response, and the Evidence Extraction agent provides the answer with supporting evidence from provided text (if any), ensuring the response is both accurate and well-supported by authoritative references

Figure 5
QA Agentic RAG Pipeline



The following sections delve into each agent within the QA pipeline, outlining their individual roles in the pipeline.

Query Decomposition Agent

The query decomposition agent aims to decompose complex user queries into more focused sub-queries. This process involves breaking down the original query into smaller parts, rephrasing them for clarity and conciseness, and correcting any spelling or grammatical errors. The agent is specifically designed to rephrase user queries into a format that aligns with the systems knowledge domains. For instance, a query such as:

ما حكم الربا؟ كيف أقدم على تخصص علم الحاسوب؟ شحال الثمن ديال هاذ الكتاب؟ ما هي شروط القبول في الجامعة؟

would be decomposed into the following sub-queries:

- ما حكم الربا؟
- كيف أقدم على تخصص علم الحاسوب في جامعة النجاح الوطنية؟
- ما هو سعر هذا الكتاب؟
- ما هي شروط القبول في جامعة النجاح الوطنية؟

To achieve the dynamic purpose of the system, the title field from the collections' metadata will be passed to the prompt in order to decompose the queries accordingly. Assuming the system has two collections with titles: Islamic Fatwa, and An-najah National University. The knowledge domain topics will be embedded in the prompt and passed to the system. The model is prompted with the Arabic prompt referenced in Appendix G.

This process ensures that the generated subqueries are semantically equivalent to the original query while being more suitable for subsequent agents.

Query Classifier Agent (Query routing)

The classification agent classifies the decomposed queries based on their relevance to the knowledge base's specialized domains. This agent uses collections' metadata (name, title, and description), which is retrieved, reformatted as a structured string, and included in the LLM prompt. The output is a JSON object pairing each decomposed question with its classified collection. The specific collection's metadata fields and their transformation

into string templates to facilitate the prompt for the classification agent is provided in Appendix H.

Queries are classified as "irrelevant" if they don't align with available domains. The classification results, stored as a JSON object with "question" and "collection" keys, are easily accessible to other processes. The classification agent serves a dual purpose: it isolates irrelevant queries and determines the most appropriate collection for retrieving relevant documents.

Candidate Answer Generation Agent

The Candidate Answer agent generates potential answers for relevant queries (those not filtered by the Query Classifier Agent). It uses the HyDE technique for query expansion by generating an initial answer from a large language model without context. This initial, potentially inaccurate response helps retrieve more relevant documents. To maximize accuracy and relevance, the LLM is instructed to provide clear, concise answers to Islamic fatwa questions, excluding Hadith or Qur'an verses due to their sensitive nature. The prompt for candidate answers generation is referenced in Appendix I.

Retriever Agent

In the retrieval process, the decomposed relevant queries and the generated candidate answers are employed to retrieve relevant documents.

In this research, the proposed hybrid retrieval approach integrates four similarity search techniques, each employing a Hierarchical Index structure. These methods retrieve the parent document, along with its attached metadata. A soft reminder that each document is represented by multiple embedding vectors to capture different aspects of its content. The similarity search techniques are:

1. **query-to-answer** ($query \xrightarrow{sim} answer$): similarity search between the query and the indexed answer.
2. **HyDE** ($query + candidate\ answer \xrightarrow{sim} answer$): combining the user's query with the candidate and answer forming a new query for similarity search against the indexed answer.
3. **query-to-question** ($query \xrightarrow{sim} question$): similarity search between the query and the indexed question.
4. **query-to-topic** ($query \xrightarrow{sim} topic$): similarity search between the query and the indexed topic.

If the dataset is not question-answer based, the same techniques are used, but without employing the query-to-question approach. Alos, instead of matching against the *answer* field, the similarity search is performed on the “content” field.

Each retriever returns the top-k most similar documents based on its retrieval criteria. The results from all retrievers are then combined into a single collection and filtered. This filtering process is guided by a ranking score, which is calculated using a ranking model, and the document frequency within the combined collection. The relevance score for each document is computed as the sum of its ranking score and its normalized frequency score. The final selection involves choosing the top-k documents based on a predetermined similarity threshold applied to the relevance score. Figure 6 illustrates the retriever agent pipeline, with the final output consisting of an array of documents. These documents retain the same structure and fields as the original input documents, aligned according to the field mapping defined by the user.

Context Relevancy Agent

The retrieved documents from the Retriever Agent often contain multiple documents related to the input query. These documents may include sentences that are not directly relevant to the answer, potentially influencing the model’s response. To address this, the retrieved documents are passed to a long-context language model to identify and extract the most relevant sentences. The prompt is referenced in Appendix J.

The agent returns an array of the most relevant sentences related to the query, applying context compression to ensure the answer generation model focuses more effectively on the most relevant information.

Answer Generation Agent

The answer generation agent receives the decomposed relevant query along with relevant sentences identified by the context relevance agent. These inputs are integrated into a carefully crafted system prompt that defines strict guidelines for generating responses. The agent ensures that responses are clear, concise, and respectful, avoiding repetition of the query and maintaining a formal tone. When no context is available, it politely acknowledges its inability to provide an answer. This approach ensures that the agent delivers contextually relevant responses. The prompt for answer generation is provided in Appendix K.

Evidence Extraction Agent

The Evidence Extraction Agent focuses on extracting Islamic daleel or legal references for legal queries. A prompt written in Arabic is passed to the LLM to identify and extract relevant Quranic verses, hadith verses, scholarly references, and legal references (such as article numbers) that support the generated response. The used Arabic prompt is referenced in Appendix L.

In the QA pipeline, the workflow iteratively calls agents Candidate Answer Generation Agent through to Evidence Extraction Agent until all relevant decomposed queries have been addressed.

Irrelevant Queries Response Agent

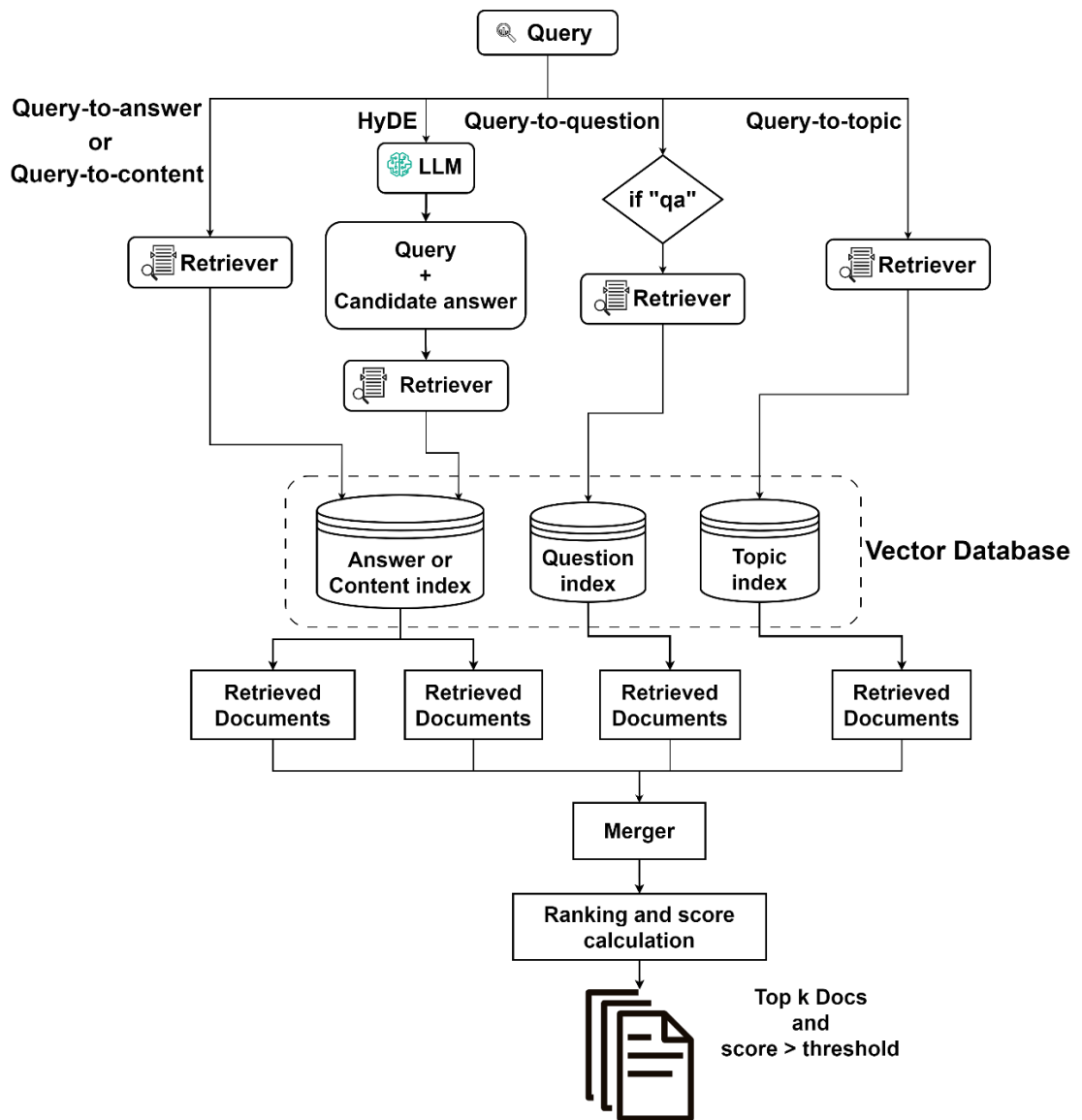
The Irrelevant Queries Response Agent is called if the user's query is deemed irrelevant to the knowledge base. This agent informs the user about the system's capabilities and provides guidance on formulating appropriate queries in case the asked question was about a topic beyond the system's knowledge. By leveraging the collection's title metadata stored in the vector store, the agent can effectively communicate the system's areas of expertise. To ensure a respectful and informative interaction, the agent is designed to respond politely to inappropriate user input, such as profanity or harmful questions. While safeguard models were tested (like Meta's Llama Guard 3) to detect harmful content, challenges were encountered in identifying profane words, particularly in Arabic. Additionally, certain questions related to Islamic fatwa, such as marital issues, were mistakenly classified as harmful.

To address these challenges, the Irrelevant Queries Response Agent utilizes a prompt designed to guide the model's response and ensure appropriate behavior. The prompt is provided in Appendix M.

No Relevant Documents Response Agent

The function of this agent is to respond to the user with a polite message indicating that the system lacks sufficient information to provide an answer. This agent is only called when the retrieval agent fails to find any relevant documents, thus preventing inaccurate responses.

Figure 6
Hybrid Retrieval Pipeline



Chapter Three

Experimentation and Results

This chapter presents a comprehensive experimental analysis of Nebras. Its performance is evaluated on two datasets: structured Islamic fatwas and semi-structured documents from An-Najah National University. Experiments analyze the effectiveness of the proposed pipeline compared to other retrieval techniques and its impact on the generated answer quality. Evaluation uses both automatic metrics and human assessments to provide a comprehensive performance overview across different domains.

RAG system evaluation focuses on two key areas: generation (assessing answer quality) and retrieval (assessing document relevance) [69,123]. Generation evaluation employs metrics such as:

- Ground Truth Similarity: measures alignment between the generated answer and the ground truth.
- Faithfulness: assesses the generated answer's adherence to the provided context.
- Answer Relevance: evaluates the answer's relevance to the question.

For retrieval evaluation, Context Relevance is measured by how relevant the retrieved documents are to the query [69]. For this evaluation, the following set of lexical and semantic metrics are used [69,124,125]:

- *ROUGE*: measures content overlap (n-grams) between generated and reference texts, focusing on recall-oriented lexical similarity (ROUGE-1, ROUGE-2, ROUGE-L) [125].
- *BLEU*: calculates n-gram precision for text translation quality, applying brevity penalties for overly concise answers, though limited in assessing fluency and grammaticality [124,125].
- *METEOR*: offers more robustness to paraphrasing, considering stemming, synonymy, and penalties/rewards for exact matches and word segmentation [124].
- *BERTScore*: a contextual metric leveraging pre-trained transformers to capture token-level semantic similarity, providing precision, recall, and F1 scores for deeper contextual alignment [124,125].

- *Semantic Textual Similarity (STS)*: evaluates the semantic similarity between text segments, using embeddings and deep learning to assess meaning and context beyond surface-level word matching [69].

While *BERTScore* and STS measure semantic similarity, *ROUGE*, *BLEU*, and *METEOR* provide lexical and syntactic overlap evaluations [124,125]. Human assessments offers more comprehensive overview by evaluating subjective factors such as relevance, consistency, and cultural sensitivity, identifying strengths and improvement areas for the system [69].

The following sections explore the steps of the experimentation process. Initially, an experiment is carried out to determine the appropriate chunk size and embedding model for the indexing process. This is followed by the implementation of the system, testing it using a question-answer dataset (Islamic fatwa), and finally, testing it with a semi-structured dataset (NNU).

3.1 Indexing and chunking

This experiment selects a suitable Arabic embedding model and optimal chunk length by systematically assessing various models and lengths. A 128-word Arabic paragraph (R1) serves as a reference document, compared with four other documents to calculate similarity scores and evaluate context capture. The following paragraphs are the test documents used for comparison:

Reference Arabic paragraph (R₁), 128 words:

التكنولوجيا الحديثة قد غيّرت بشكل جذري أسلوب حياتنا. من الهواتف الذكية إلى الإنترنت عالي السرعة، أصبح العالم اليوم متصلاً بشكل لم يكن من الممكن تصوره قبل عقود قليلة. هذا الاتصال الدائم يتيح لنا الوصول إلى المعلومات في أي وقت وفي أي مكان، مما يعزز من قدرتنا على التعلم والتواصل. ومع ذلك، فإن هذا التقدم التكنولوجي يطرح تحديات جديدة، مثل التأثيرات السلبية على الصحة النفسية نتيجة للاستخدام المفرط للأجهزة الإلكترونية. بالإضافة إلى ذلك، يثير تزايد الاعتماد على التكنولوجيا مخاوف بشأن

الخصوصية وأمان البيانات. لذا، فإن تحقيق توازن بين الاستفادة من التكنولوجيا والحد من مخاطرها أصبح ضرورياً. من خلال التوجيه السليم واستخدام التكنولوجيا بوعي، يمكننا التمتع بفوائدها مع تقليل آثارها السلبية.

Arabic contextually similar paragraph (P₁), 120 words:

التقنيات الرقمية أصبحت جزءاً لا يتجزأ من حياتنا اليومية. الهواتف الذكية والإنترنت قد جعلت العالم قرية صغيرة، حيث يمكن الوصول إلى المعلومات بسرعة فائقة. هذه الثورة التكنولوجية توفر فرصاً هائلة للتعليم والتواصل، ولكنها أيضاً تحمل مخاطر جديدة مثل التأثيرات السلبية على الصحة النفسية بسبب الاستخدام المفرط. تزايد الاعتماد على الأجهزة الذكية يثير قضايا حول الخصوصية وحماية البيانات الشخصية. لذلك، يجب علينا استخدام التكنولوجيا بحذر والبحث عن توازن بين الاستفادة منها والحفاظ على صحتنا وخصوصيتنا. من خلال التوعية والتوجيه الصحيح، يمكننا تحقيق أقصى استفادة من هذه الأدوات الحديثة وتقليل مخاطرها المحتملة.

Arabic contextually different paragraph (P₂), 128 words:

الفن هو أحد أهم الوسائل التي تعبر عن مشاعر الإنسان وتجسد ثقافات الشعوب عبر العصور. اللوحات الفنية والمنحوتات تأخذنا في رحلات عبر الزمن وتتيح لنا فهم العوالم المختلفة التي عاشها الإنسان. من خلال الألوان والخطوط والأشكال، يعبر الفنان عن رؤيته الخاصة للعالم ويثير فينا مشاعر متباينة. الفنون البصرية ليست فقط وسيلة للتعبير عن الجمال، بل هي أيضاً لغة عالمية تتجاوز الحدود الجغرافية والثقافية. على الرغم من التحديات التي تواجه الفنون في عصر التكنولوجيا، إلا أنها تظل رمزاً للإبداع والحرية. يجب علينا الحفاظ على التراث الفني وتعزيزه، لأنه يعكس هوية الأمم ويساهم في بناء الجسور بين الثقافات المختلفة.

English contextually similar paragraph (P₃), 115 words:

“Modern technology has fundamentally transformed our way of life. From smartphones to high-speed internet, the world today is more connected than ever before. This constant connectivity allows us to access information anytime and anywhere, enhancing our

learning and communication capabilities. However, this technological progress also presents new challenges, such as the negative effects on mental health due to excessive use of electronic devices. Additionally, the growing reliance on technology raises concerns about privacy and data security. Therefore, achieving a balance between reaping the benefits of technology and minimizing its risks has become essential. Through proper guidance and mindful use, we can enjoy the benefits of technology while reducing its adverse effects.”

English contextually different paragraph (P₄), 128 words:

“Art is one of the most significant ways through which human emotions are expressed and the cultures of different societies are embodied across ages. Paintings and sculptures take us on journeys through time, allowing us to understand the diverse worlds humans have inhabited. Through colors, lines, and shapes, artists convey their unique visions of the world, evoking various emotions within us. Visual arts are not merely a means of expressing beauty; they also serve as a universal language that transcends geographical and cultural boundaries. Despite the challenges that art faces in the age of technology, it remains a symbol of creativity and freedom. Preserving and promoting artistic heritage is crucial, as it reflects national identity and helps build bridges between different cultures.”

To measure the semantic similarity between the reference sentence (R_1) and each test sentence, cosine similarity is employed. An ideal embedding model should demonstrate a significant difference in similarity scores between P_1 and P_2 , reflecting its ability to accurately identify contextual relationships within the Arabic language. The difference in result similarity score between P_1 , and P_2 is represented as ($\Delta P_1 P_2$) in Table 8.

Conversely, a model that fails to distinguish between P_1 and P_2 would indicate a deficiency in capturing the Arabic semantics. Furthermore, P_3 and P_4 , written in English, are used to assess the model’s ability to differentiate between languages and maintain semantic consistency. Given that all embedding models employed in this analysis are multilingual, an ideal model should exhibit a high similarity score between P_3 and R_1 , indicating its ability to accurately capture cross-lingual semantic relationships, and a low similarity score between P_4 and the R_1 would demonstrate the model’s ability to discriminate between semantically distinct content, even when presented in different languages. The difference in result similarity score between P_3 , and P_4 is represented as

($\Delta P_3 P_4$) in Table 8. The ‘‘Score’’ column is the average between ($\Delta P_1 P_2$) and ($\Delta P_3 P_4$). The higher the score the better the model in distinguishing contextual meaning.

Table 8

Embedding Models Evaluation

Model	Similarity				$\Delta P_1 P_2$	$\Delta P_3 P_4$	Score
	R_{1,P_1}	R_{1,P_2}	R_{1,P_3}	R_{1,P_4}			
intfloat-multilingual-e5-large	95.48	80.91	90.26	74.17	14.57	16.1	15.33
intfloat-multilingual-e5-large-instruct	95.36	77.97	92.2	66.04	17.39	26.16	21.78
BAAI-bge-m3	88.06	63.39	91.74	58.81	24.67	32.93	28.8
distiluse-base-multilingual-cased-v1	67.97	24.52	74.54	26.73	43.45	47.82	45.63
distiluse-base-multilingual-cased-v2	75.47	30.1	84.64	36.54	45.38	48.11	46.74
LaBSE	92.82	47.56	95.96	47.49	45.26	48.48	46.87
paraphrase-multilingual-MiniLM-L12-v2	86.49	28.13	85.39	22.19	58.36	63.21	60.79
paraphrase-multilingual-mpnet-base-v2	88.83	33.38	90.11	19.69	55.46	70.43	62.94

Based on the results presented in Table 8, the paraphrase-multilingual-mpnet-base-v2 model developed by SentenceTransformers achieved the highest score for similarity difference. This suggests that the model effectively distinguished the contextual similarities and differences between P_1 , P_3 and P_2 , P_4 , respectively, in relation to R_1 .

To ensure optimal performance of the RAG system, the process of chunking the text corpus into smaller segments is carefully considered. Given the maximum token count of 128 for the paraphrase-multilingual-mpnet-base-v2 model, a chunk size of 125 tokens is selected. This choice balances the need to capture sufficient context within each chunk while avoiding excessively long sequences that could overwhelm the model. By setting an overlap of 1 token between adjacent chunks, the risk of losing semantic information at chunk boundaries is avoided. Processing smaller chunk sizes can improve computational efficiency, especially when dealing with large documents because it requires less computational resources and can lead to faster response times [73,74].

3.2 Implementation

For the system’s implementation, Langchain and Langgraph are employed to develop the agent-based architecture, while FastAPI serves as the backend to handle requests and stream the generated response. In terms of data storage, the system uses ChromaDB as the vector store, with support for Elasticsearch, giving users the flexibility to choose the most suitable vector store for their needs. The agents in the system are powered by two LLMs: Meta’s LLaMA 3.1-70B and Google Gemini 1.5 Pro

In this implementation, primarily the agents use Meta’s LLaMa3.1-70b which is hosted remotely on Groq (<https://groq.com/>), and Google’s Gemini 1.5 pro, which is hosted by Google, serving as a backup model in case of request failures. Table 9 summarizes the implementation technologies and specifications.

Table 9

Summary of Implementation Technologies

Component	Value
Embedding model	paraphrase-multilingual-mpnet-base-v2
Chunk size	125
Ranking model	bge-reranker-v2-m3
Vector database engine	ChromaDB
Language Models	Llama3.1-70B & Gemini 1.5 pro
Frameworks and Libraries	Langchain, Langgraph, FastAPI

3.3 Experiment with Islamic Fatwa Dataset

This section presents the experimental results from evaluating language models on Islamic fatwa-related queries. Initially, the models were tested in a baseline setting, relying solely on their pre-trained parameters without leveraging external information retrieval systems. Subsequently, the impact of augmenting relevant context was examined, followed by the results of using the RAG framework to augment relevant documents.

3.3.1 Baseline Evaluation: Responses Without Retrieval

To evaluate the performance of LLMs on Islamic fatwa-related queries, the following randomly selected questions taken from the Najah National University Fatwas website are posed to OpenAI’s GPT-4o, Google’s Gemini Pro 1.5, and Meta’s Llama3.1-70B:

1. ما حكم بيع اسم الشهرة، بمعنى بيع العلامة التجارية؟
2. ما حكم رجوع المطلقة لزوجها بعد قبض شيء من المؤخر؟
3. التوبة من المعاصي والندم على ارتكابها
4. حكم الترويج لمنتجات لشركة مقابل عمولة مع شرط مسبق على المروج بدفع مبلغ مالي
5. هل يوجد وقت خاص لقراءة سورة البقرة؟
6. حكم التجاره بالذهب على انه سلعه وتاخير الدفع الى اجل
7. ما حكم الإفطار في رمضان للمصابين بمرض كورونا (كوفيد 19)؟
8. ما هو شرح حديث : يخرج عنق من النار يوم القيامة لها عينان تبصران، وأذنان تسمعان
9. حكم حضور المناسبات الاجتماعية التي تمارس فيها المنكرات
10. سب الله والرسول والتوبة عن ذلك

The generated responses for each LLM are evaluated using a combination of automatic metrics (BERTScore, BLEU, METEOR, ROUGE, STS) and human assessment by Palestinian Dar Al-Ifta'a, with the former's results illustrated in Figure 7 (a) and further detailed in Appendix N - Table N12, while the latter's findings are presented in Table 10.

The results presented in Figure 7 (a), and Table 10 demonstrate that while all models exhibited varying degrees of competence, OpenAI's GPT-4o consistently outperformed the other models across most evaluation metrics. However, even the best-performing model showed significant shortcomings in generating accurate and comprehensive fatwas, particularly for complex or detailed queries. This illustrates the limitations of current language models in effectively addressing the challenges posed by such domains. Notably, Meta's Llama 3.1-70b scored lower than other models on nearly all automatic metrics and had the fewest accepted generated fatwa answers according to Dar Al-Ifta'a's evaluation. This performance is expected, as Llama 3.1-70b, being smaller in size, is less suited for specialized domains and offers limited support for other languages compared to larger models [115].

In the next step, the potential of augmenting relevant context to the LLM is explored to enhance the performance of language models in generating accurate and relevant responses to Islamic fatwa-related queries.

3.3.2 Context Effectiveness on Fatwa Answer Generation

Before examining the retrieval methods, this experiment focused on testing context augmentation specifically on Llama 3.1-70b. This model is chosen because it received the lowest scores in both automatic metrics and Dar Al-Ifta'a's evaluation during the no-context baseline assessment. The objective is to determine whether providing relevant contextual information can improve the model's ability to handle Arabic text and produce accurate responses to fatwa questions. The generated responses are evaluated using the same criteria as in the baseline assessment, employing both automatic metrics and human evaluation. To ensure consistency and facilitate comparison with the baseline evaluation, the same 10-question dataset is employed for this experiment. Each question within the dataset is enriched with contextual information in the form of relevant fatwa answers. The provided context is a combination of the correct fatwa answer and another fatwa relatively related to the question. These fatwa answers are obtained from reputable sources, including:

- Najah University Fatwa (<https://fatwa.najah.edu/ar>)
- Islamweb.net (<https://www.islamweb.net/>)
- Binbaz Fatwa (<https://binbaz.org.sa/fatwas/kind/1>)

The prompt used to generate the answer is constructed by replacing the placeholders {question} and {context} with the asked question and the combined fatwa answers, respectively. Detailed prompt used for this purpose is provided in Appendix O.

The hardware used for the experimentation includes an RTX3050 4GB GPU, 16GB DDR5 RAM, and an AMD Ryzen 7 6800H processor with a 3.20 GHz clock speed. It is important to note that only the LLMs are hosted remotely and not running on the local device. The hardware specifications are relevant for later stages when re-ranking, additional models, and the vector database are hosted locally.

The ground truth automatic metrics and human evaluation scores for the fatwas generated by Llama 3.1-70b after augmenting context are presented in Figure 7 (b) and Table 10,

respectively, in comparison to the baseline evaluation results. Detailed tabular evaluation scores are provided in Appendix N - Table N13 for further reference.

Table 10

Palestinian Dar Al Ifta'a Evaluation of Generated Fatwas

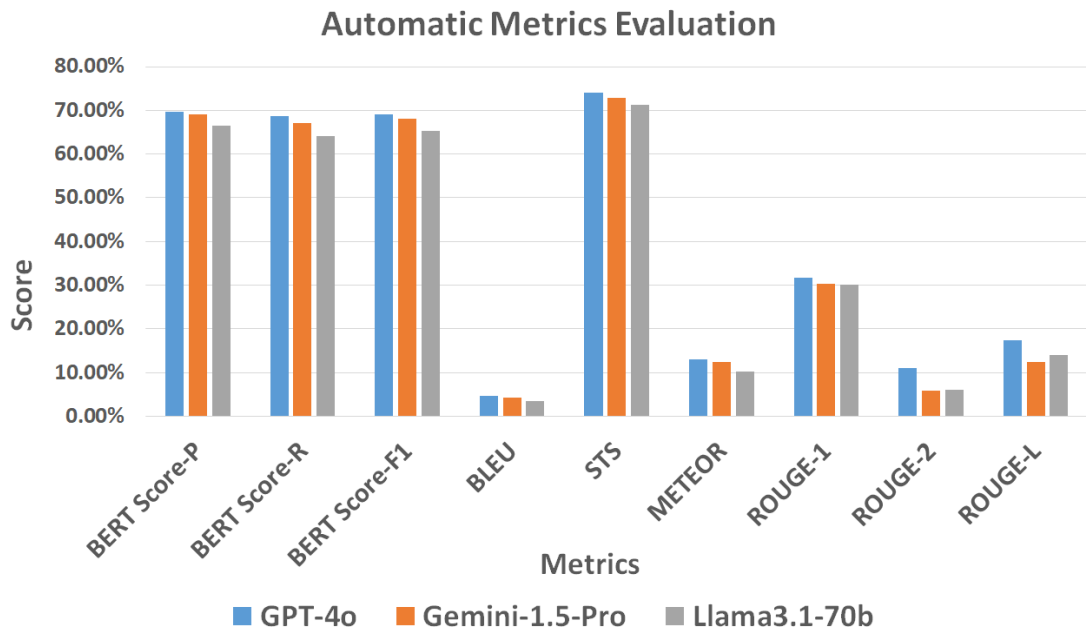
Model	Number of Accepted Fatwas
Llama 3.1-70B	3
GPT-4o	7
Gemini 1.5 Pro	6
Llama 3.1-70B (with context)	8

As illustrated in Figure 7 (b), Llama3.1-70b (with context) outperforms its base version and shows strong competitive performance against other models like GPT-4o and Gemini-1.5 Pro across various metrics. Incorporating context led to improvements, particularly in BLEU (9.56%), METEOR (19.47%), and ROUGE scores (ROUGE-2 at 20.52% and ROUGE-L at 24.26%), surpassing all other models. While it achieves the highest precision (BERTScore-P at 71.56%) and a well-balanced BERTScore-F1 (68.46%), its recall (BERTScore-R at 66.37%) and semantic textual similarity (STS at 73.11%) slightly lag behind GPT-4o. Nonetheless, these results outline the impact of contextual information on the generated Arabic fatwa responses, even when Llama 3.1-70b does not officially support Arabic language [115].

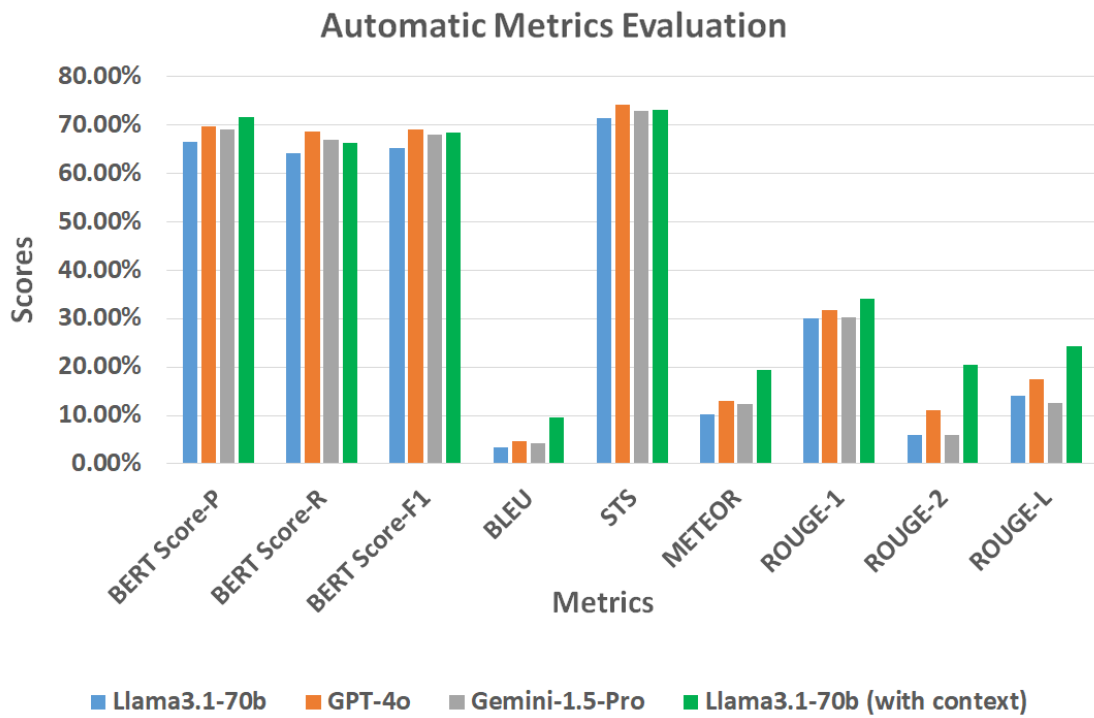
Figure 7

Ground Truth Evaluation With and Without Context

(a) *Ground Truth Evaluation of Fatwas Without Context*



(b) *Evaluation of Llama 3.1-70B With Context vs. Baseline Models*



3.3.3 Employing RAG

The following sections explore the key components of RAG implementation, beginning with indexing and chunking strategies, followed by an examination of vector stores and retrieval mechanisms.

Vector store

To evaluate the effectiveness of various retrieval techniques, the fatwa documents are initially split into 125-token length chunks, subjected to the maximum input token limit of the paraphrase-multilingual-mpnet-base-v2 model.

For this experiment, the answer field serves as the primary vector field for similarity search, while the remaining fields are stored as document metadata. Each fatwa is assigned a unique identifier (id) for linking its split chunks back to the parent document when using a parent document retriever. Although each chunk receives a new id, the parent fatwa's id is stored as metadata under the "doc_id" field. By default, the chunks inherit the parent's metadata, but to optimize storage, only the "doc_id" field is retained, ensuring that the original data is stored solely in the parent document and preventing duplication.

As illustrated in Figure 2, the Hierarchical Structure Indexing approach enables the retrieval of the entire parent document rather than individual chunks [19]. This differs from a more I approach that solely retrieves relevant chunks. In a question-answering dataset for fatwas, each fatwa represents an independent case, with its answer tailored specifically to that case. This means that each document in the vector database is self-contained and does not rely on other documents for context. Although, multiple similar fatwas could be concatenated to provide a context.

To evaluate the effectiveness of different vector representations, each fatwa document is represented with two additional vectors: one for the question field and the other for the topic field leading to multiple vector representations for a single document. This representation enabled the retrieval of documents based on similarity searches within these specific fields. These vectors are subsequently chunked and linked to the parent document using the previously described technique.

Relevant Context Retrieval

To evaluate the effectiveness of the RAG approach and the Hierarchical Structure Index technique which is also known as Parent Document Retriever (PDR) in some frameworks like Langchain [126], a series of experiments are conducted with varying similarity thresholds (θ) and maximum number of retrieved documents (k). The relevance of the retrieved context is evaluated using automatic metrics. These metrics assess context relevance (CR) by measuring the similarity between the retrieved context and the query, as well as ground truth (GT) by comparing the generated answer to the ground truth answer. The results of these experiments are summarized in Appendix P - Table P14.

The comparative analysis in Appendix P - Table P14 revealed that the PDR technique outperformed the RAG approach across all evaluation metrics and parameter combinations. This suggests that retrieving the entire parent document provides a more comprehensive context, leading to improved performance in relevant documents retrieval.

Furthermore, the experiment found that decreasing the similarity threshold (θ) generally resulted in better performance, indicating that less strict similarity criteria can enhance the relevance of retrieved documents. However, there is a trade-off between recall and precision when adjusting the maximum number of retrieved documents (k). While increasing k can potentially improve performance by providing more context, it may also introduce irrelevant documents, diluting the effectiveness of the retrieved context.

The results also are evaluated by a human expert, categorizing each generated response into one of four categories:

1. Accepted: Responses that were deemed accurate, relevant, and appropriate by the human evaluators.
2. Rejected: Responses that were deemed irrelevant or off-topic.
3. Inaccurate: Responses that contained factual errors or inaccuracies.
4. Unanswered: Responses that failed to address the given query or provide a meaningful answer.

The human evaluation, summarized in Appendix P - Table P15, revealed that the PDR technique consistently outperformed the RAG approach across all parameter

combinations. PDR demonstrated a higher acceptance rate, fewer rejected and inaccurate responses, and overall superior performance in generating relevant and accurate answers.

Increasing the similarity threshold (θ) generally led to improved results, indicating that stricter criteria can enhance the quality of retrieved context. However, the number of retrieved documents (k) did not have a significant impact on overall performance, suggesting that the quality of the retrieved context is more crucial than the quantity.

The following subsections examine the effects of various retrieval and techniques, including document re-ranking and context compression.

Retrieved Documents Re-ranking

In this experiment, the impact of incorporating a ranking model into the retrieval process is investigated. This experiment leverages the BAAI/bge-reranker-v2-m3 to rank the retrieved documents based on their relevance to the query. The automatic metrics and human evaluation results are presented in Appendix Q - Table Q16.

The inclusion of ranking in the retrieval process demonstrates a noticeable impact on the evaluation metrics across various configurations. Ranking generally improves precision-oriented metrics such as BERTScore-P, BLEU, and METEOR, particularly in scenarios with PDR and higher k values. For example, BERTScore-P and BLEU exhibit significant gains in configurations like PDR ($\theta=0.5$, $k=10$), where precision improves from 72.06% to 72.96%, and BLEU increases from 2.83% to 3.08%. However, recall-oriented metrics, including BERTScore-R and ROUGE, show mixed results, with minor reductions in some cases, likely due to the reduced diversity of retrieved documents. This is evident in setups such as ($\theta=0.5$, $k=10$), where Recall drops slightly from 64.05% to 64.01%. Additionally, STS scores consistently improve with ranking, highlighting its ability to better align retrieved results with ground truth in terms of semantic similarity. Overall, ranking enhances the precision and relevance of results, particularly in configurations with high retrieval depth or PDR, but may sacrifice some recall in situations requiring diverse retrieval.

While ranking demonstrated some positive effects as illustrated in Appendix Q - Table Q17, it did not consistently improve performance across all parameter combinations. In certain instances, ranking led to a decrease in accepted responses, suggesting that its

impact can be variable and dependent on specific factors. However, the consistent reduction in rejected and inaccurate responses indicates that ranking can contribute to improved relevance and accuracy of the retrieved documents.

Hypothetical Document Embeddings

One potential improvement is to explore the use of HyDE to enhance the retrieved context. HyDE uses the input query combined with a potential answer generated by the LLM using its parameterized memory instead of relying on the input query alone to retrieve relevant context [79]. The results of this experiment are shown in Appendix R - Table R18.

By analyzing the automatic metric scores in Appendix R - Table R18, HyDe demonstrates improvements compared to the standard retrieval with ranking approach across different parameter configurations. Metrics such as BLEU, METEOR, and STS consistently show enhancements with HyDe, particularly in lower θ values or smaller k . For example, BLEU improves from 0.42% to 2.35% and METEOR rises from 4.81% to 6.62% for $\theta=0.5$ and $k=10$. Although slight reductions are observed in metrics like BERTScore-F1 and ROUGE-1 under specific settings, the differences are minimal and do not diminish HyDe’s overall advantages. Overall, HyDe outperforms the ranking-based approach across diverse metrics.

The human evaluation results for HyDE in Appendix R - Table R19 reveal distinct patterns across different configurations. HyDE without PDR consistently outperforms HyDE with PDR, indicating that the document relevance scores used in PDR may not be as effective in identifying relevant documents. Among the HyDE configurations without PDR, the parameters $\theta=0.7$, $k=10$ demonstrate the most promising performance, achieving the highest acceptance rate. This suggests that a moderate similarity threshold and a larger number of retrieved documents can be beneficial.

Query-to-question Similarity Matching

To further explore the potential of retrieval techniques, additional experiments using query-to-question similarity are conducted. The automatic evaluation results for query-to-question retrieval are illustrated in Appendix S - Table S20. The results in demonstrate the effectiveness of query-to-question matching as a retrieval technique in question-answer based datasets. Across a range of evaluation metrics, this approach outperforms

previous methods, including HyDE and ranking-based techniques. Query-to-question matching exhibits improvements in BERTScore, indicating that the retrieved documents are more coherent and relevant. While the impact on other metrics, such as BLEU, METEOR, ROUGE, and STS, varies to some extent, overall, the method is an improvement.

For further analysis of the retrieval process and to identify potential areas for improvement, a human evaluation of the retrieved documents is conducted. The results are presented in Appendix S - Table S21. The human evaluation results for query-to-question matching reveal a varied outcome. While the approach demonstrates promise in terms of automatic evaluation metrics, it struggles to consistently retrieve highly relevant documents.

Despite the high rejection rates, the number of inaccurate and unanswered responses is relatively low, suggesting that when relevant documents are retrieved, they are generally accurate.

The choice of parameters, including the similarity threshold and number of retrieved documents, significantly impacts the performance of question-to-question matching. Careful tuning of these parameters is essential to optimize results.

Query-to-topic Similarity Matching

To further explore the potential of retrieval techniques, additional experiments using query-to-topic retrieval are conducted. This approach aims to identify the most relevant topics (document summaries) within a document and match them to the query. The automatic evaluation results are presented in Appendix T - Table T22. For most metrics, query-to-question matching slightly outperforms query-to-topic matching, especially in. For instance, in $\theta=0.5$ and $k=10$, query-to-question matching achieves a higher ROUGE-2 score (14.26% vs. 13.91%) and BERTScore-F1 (67.58% vs. 67.38%) for ground truth (GT). Metrics such as METEOR and STS also consistently favor query-to-question matching, with query-to-question achieving 6.65% (CR) and 64.80% (GT) for METEOR and STS, respectively, compared to 6.44% and 64.09% in query-to-topic.

However, query-to-topic matching demonstrates better performance stability across configurations. For example, the variability in scores like BLEU and ROUGE is smaller in query-to-topic matching. BLEU in query-to-topic remains consistently close across

settings, such as 2.91% ($\theta=0.5$, $k=5$) and 2.57% ($\theta=0.5$, $k=10$), while in query-to-question matching, the variation is slightly more pronounced, such as 2.89% ($\theta=0.5$, $k=5$) to 2.59% ($\theta=0.5$, $k=10$). Overall, query-to-question matching tends to achieve slightly higher performance for most metrics, particularly under parameters that emphasize greater contextual depth (e.g., lower θ or larger k values). A human evaluation is conducted to gain a deeper evaluation of the retrieval process. The results are demonstrated in Appendix T - Table T23. The human evaluation results demonstrate higher acceptance rates and lower rates of inaccurate and unanswered responses.

While PDR retrieval demonstrated context enrichment, certain experiments revealed inefficient performance, particularly when retrieved contexts were extensive in length. This can be attributed to inconsistencies in the retrieved passages, which often yielded better results with shorter contexts. Additionally, HyDE retrieval and topic retrieval demonstrated improvements in the quality of generated results and acceptance rates, respectively. However, none of these retrieval methods achieved the target acceptance rate of 70%. To potentially enhance the acceptance rate and overall performance, a hybrid retrieval approach is proposed in the subsequent section.

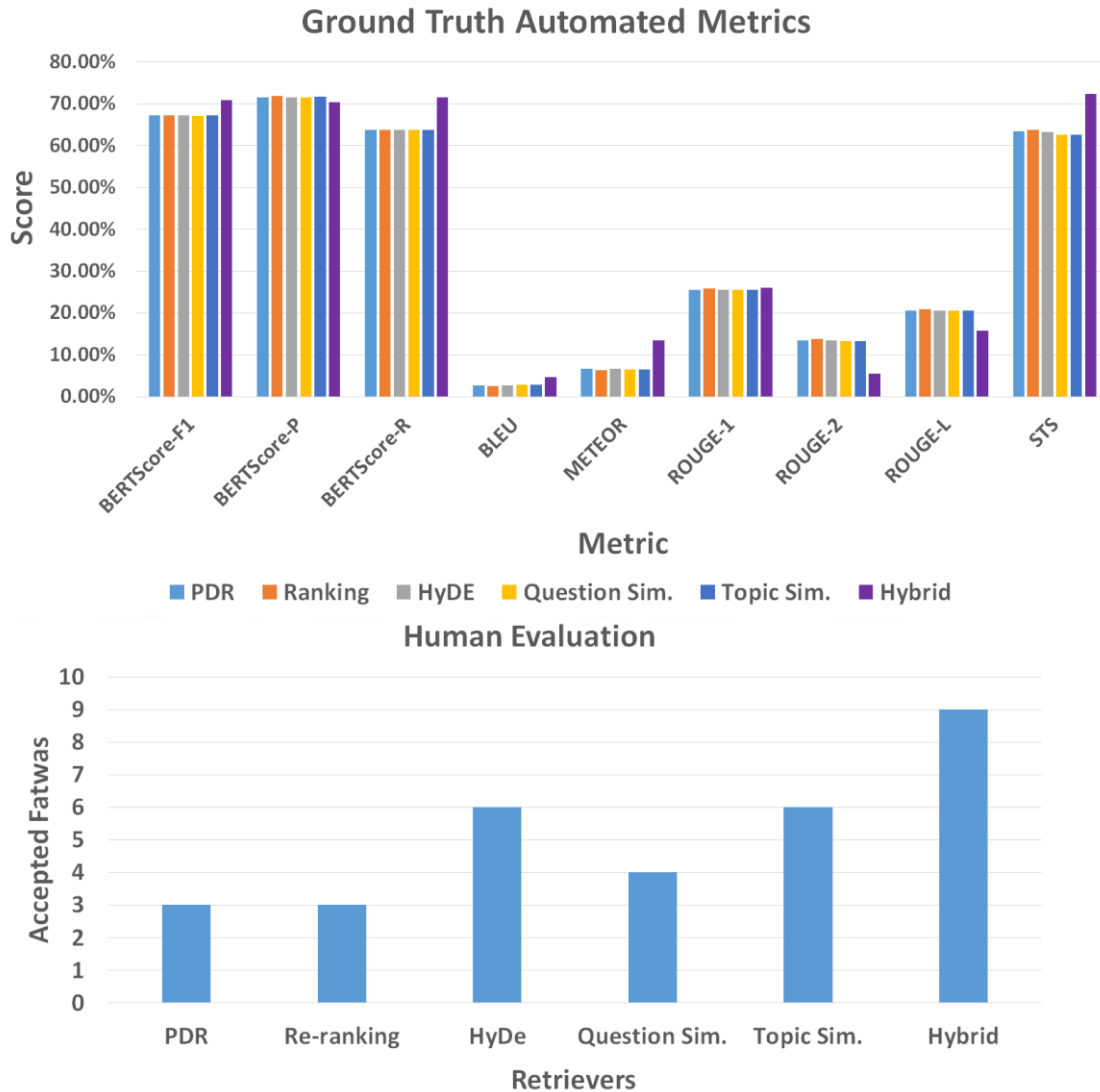
Hybrid Retrieval

The proposed hybrid retrieval approach integrates four distinct retrieval methods: PDR query-to-answer, PDR HyDE, PDR query-to-question, and PDR query-to-topic as previously explained in the “Retriever Agent” section. To evaluate the effectiveness of the hybrid retrieval approach, the automatic metrics used in previous experiments are conducted and the results are demonstrated in Appendix U - Table U24.

The hybrid retriever, as evaluated through automatic metrics, consistently outperforms all previously explored retrieval methods. This superiority is particularly evident in BERTScore, a metric indicative of the coherence and relevance of generated responses. Furthermore, the hybrid retriever exhibits improvements in BLEU, METEOR, ROUGE, and STS, demonstrating its effectiveness in capturing semantic similarity and generating accurate results. The parameters, $\theta=0.5$ and $k=5$, appear to give the best results. Figure 8 compares the GT and human evaluation results from all previous experiments using parameters $\theta = 0.5$ and $k = 5$. The corresponding CR values from these experiments are illustrated in Appendix U - Figure U15.

Figure 8

Comparing GT Automatic Metrics & Human Evaluation



To further assess the effectiveness of the hybrid retrieval approach, a human evaluation is conducted using the parameters $\theta=0.5$ and $k=5$, and a ranking score threshold of 0.9. The human evaluation results presented in Figure 8 compares the highest accepted number of fatwas in each retriever across different parameters. The hybrid retriever achieved an acceptance rate of 90%, demonstrating its effectiveness in retrieving relevant and accurate documents.

When comparing the ground truth scores of Nebras to the baseline scores of other models (Figure 9), the performance analysis highlights Nebras as a strong contender in semantic-oriented metrics, excelling in BERTScore (F1:70.94%) and METEOR (13.49%), where it outperforms all other models. This demonstrates its ability to generate text with high

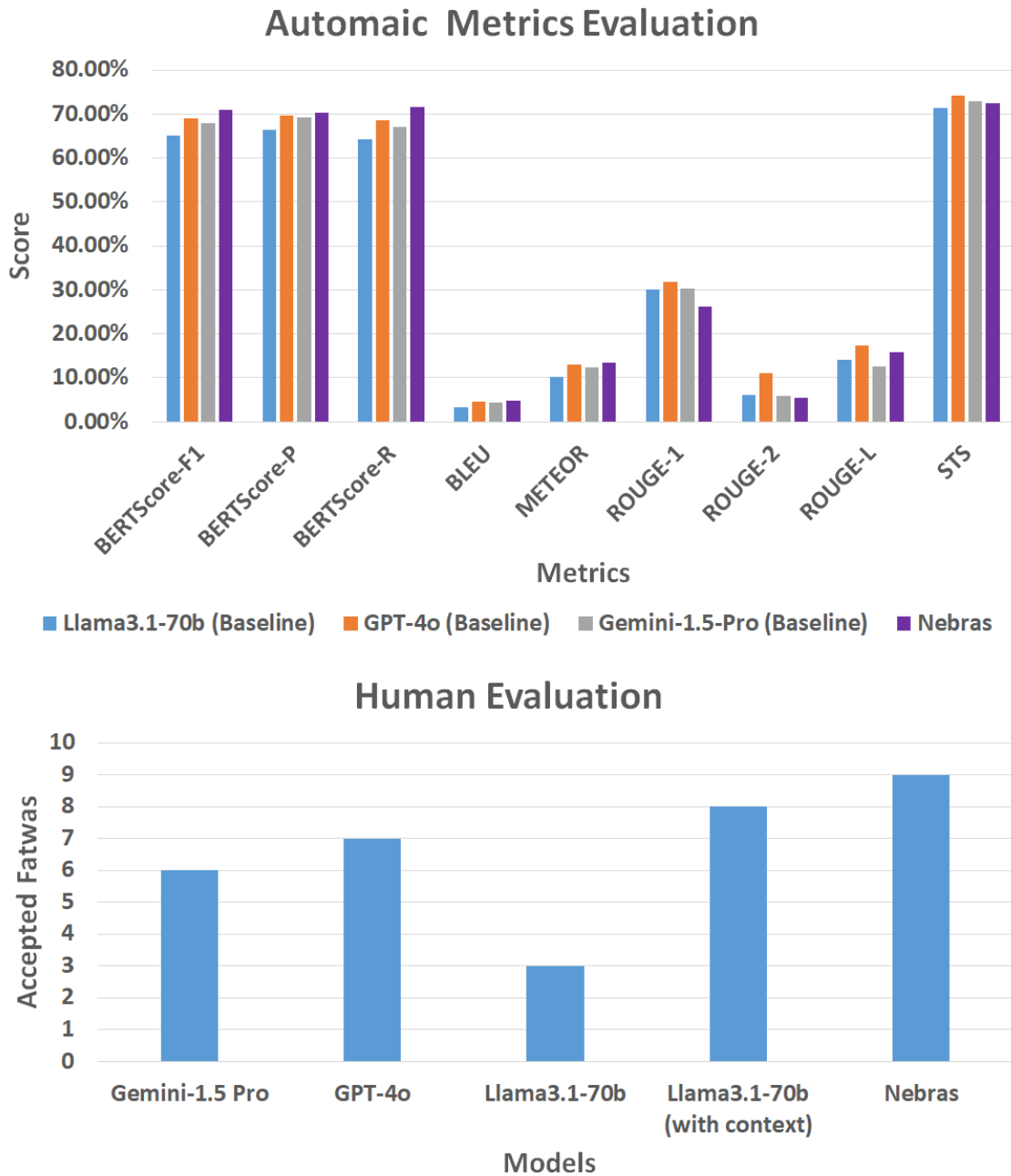
semantic similarity to references. Nebras also slightly leads in BLEU (4.79%), indicating good n-gram precision. However, its performance in ROUGE metrics, particularly ROUGE-1 (26.13%) and ROUGE-2 (5.51%), falls short compared to GPT-4o and Gemini-1.5-Pro, suggesting potential weaknesses in capturing unigram and bigram overlaps. In STS, Nebras remains competitive (72.44%) but slightly underperforms GPT-4o (74.10%).

Figure 9 also compares the number of accepted fatwas based on human evaluation highlighting Nebras as the leading model, with 9 accepted fatwas. This significantly outperforms other models, including GPT-4o, which ranks second with 7 accepted fatwas, and Gemini-1.5 Pro, which follows with 6. Llama3.1-70b, with only 3 accepted fatwas, shows the weakest performance in this metric.

To further assess the system's ability to generate accurate and factual answers, an additional evaluation is conducted using 100 fatwa questions. The results of this evaluation are presented in Appendix V - Table V26. Nebras demonstrates competitive performance against state-of-the-art models, surpassing Gemini 1.5 Pro across all key metrics and achieving results comparable to GPT-4o. It excels in ROUGE scores, significantly outperforming both GPT-4o and Gemini, indicating strong recall and coherence. Additionally, Nebras achieves the highest STS score (74.50%), suggesting better semantic similarity to ground truth. While its BERTScore-P (71.80%) is the highest among all models, indicating high precision, its BERTScore-R (64.30%) is slightly lower than GPT-4o, suggesting room for improvement in recall. Moreover, Nebras performs better than Gemini but slightly below GPT-4o in BLEU (3.20%) and METEOR (7.80%), indicating a potential area for refinement in phrase-level diversity and exact word matches. Overall, these results highlight Nebras's capability, performing at a comparable level to leading commercial LLMs while excelling in structured response generation and semantic accuracy.

Figure 9

Ground Truth Comparison Between Nebras and Baseline Models



3.4 Experiment with An-najah National University Dataset

To evaluate the effectiveness of the retrieval pipeline, a series of experiments were conducted using a dataset of five questions specifically tailored to An-Najah National University. These questions were designed to assess the models' ability to provide accurate and informative responses to common inquiries related to university policies, academic matters, and personnel. The following subsections explore the impact of the

retrieval pipeline by comparing the baseline performance without information retrieval to the results obtained after incorporating the retrieval process. The dataset questions are:

1. من هو عميد كلية الهندسة في جامعة النجاح الوطنية؟
2. ما هي التخصصات المتوفرة في كلية تكنولوجيا المعلومات في جامعة النجاح الوطنية؟
3. ما هو معدل القبول لتخصص المحاسبة في جامعة النجاح الوطنية؟
4. من هو الرئيس الحالي لجامعة النجاح الوطنية؟
5. ما هي مساقات الجامعة الإجبارية لجميع التخصصات في جامعة النجاح الوطنية؟

Translated to English:

1. Who is the Dean of the College of Engineering at An-Najah National University?
2. What specializations are available in the College of Information Technology at An-Najah National University?
3. What is the admission average for the Accounting program at An-Najah National University?
4. Who is the current president of An-Najah National University?
5. What are the university's mandatory courses for all specializations at An-Najah National University?

3.4.1 Baseline Evaluation: Responses Without Retrieval

To establish a baseline for evaluating the performance of language models in handling university-related queries, a comprehensive experiment was conducted. Each of the five questions designed to assess the models' capabilities was submitted to Google's Gemini 1.5 Pro, OpenAI's ChatGPT 4o, and Meta's Llama 3.1-70b. The generated responses were subsequently analyzed using a combination of automatic evaluation metrics and human judgment. The results of these evaluations are summarized in Appendix W - Table W27 and Table W28, respectively.

As shown in Appendix W - Table W27, GPT-4o consistently outperformed the other models across various metrics, indicating a stronger ability to understand and respond to these university-specific queries. However, the human evaluation in Appendix W - Table W28 reveals that even OpenAI's GPT-4o struggled with providing entirely accurate and

comprehensive answers, particularly for questions requiring specific numerical data or detailed academic information. This suggests that while LLMs have made significant progress, they still face challenges in handling complex and nuanced queries related to institutional knowledge.

3.4.2 Hybrid Retrieval Pipeline Evaluation with NNU Dataset

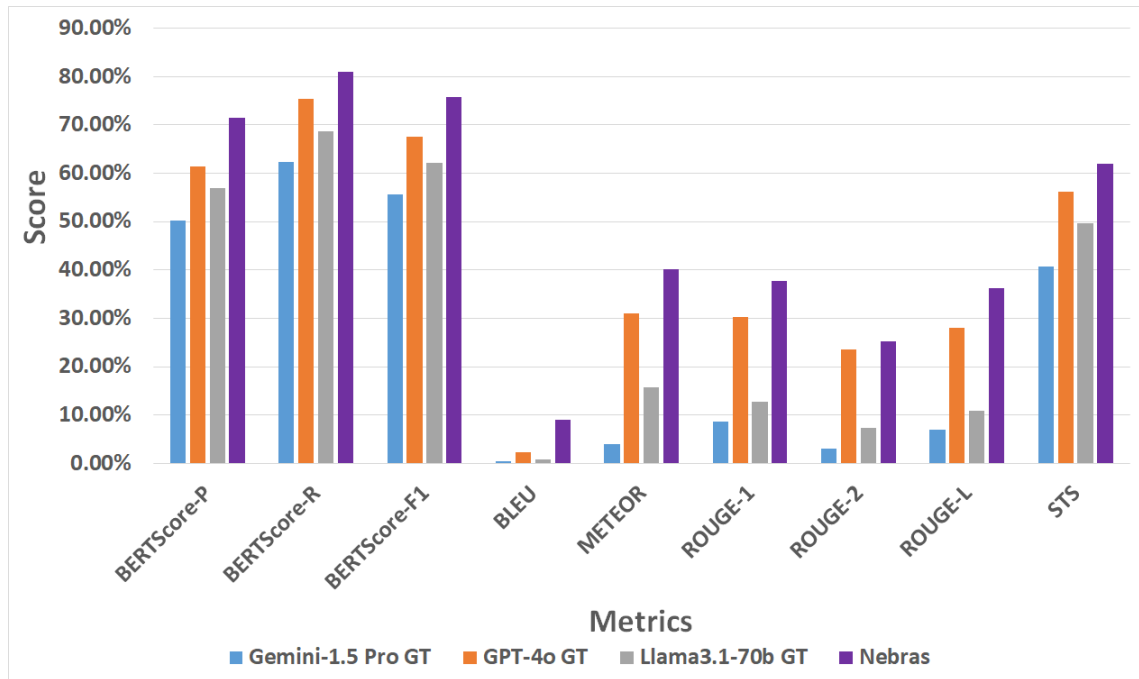
To evaluate the impact of the hybrid retrieval pipeline on the quality of generated responses, A comprehensive evaluation is conducted on the same question dataset used in the previous experiment, which evaluated the LLMs without context. The evaluation consists of a set of automatic metrics including BERTScore, BLEU, ROUGE, and STS. The results, presented in Appendix X - Table X29, provide insights into the system's performance in terms of semantic similarity and lexical coherence. The evaluation results demonstrate the impact of the hybrid retrieval mechanism on the performance of the system. Incorporating the hybrid retrieval pipeline exhibits improvements across all metrics, including BERTScore, BLEU, METEOR, ROUGE, and STS. By effectively retrieving and augmenting relevant information, the model is able to generate more accurate, coherent, and semantically similar responses, leading to enhancing the overall quality of the generated responses. Figure 10 compares the ground truth automatic evaluation scores between the baseline evaluation results the Nebras's results.

Human evaluation of the generated responses revealed that the system's performance was generally satisfactory, with four out of five responses deemed factually correct based on verification with official university information. However, in the case of the fifth question, which inquired about mandatory university courses, the system correctly identified the total credit hours but was unable to provide a comprehensive list of specific courses. All models' generated responses in this research are provided in Appendix Y.

Across different pipeline experiments, the total execution time ranged from 9.61 to 13.82 seconds. This variation was mainly due to hardware constraints and the length of the retrieved contexts. The primary bottleneck is when the Google Gemini 1.5 Pro API is called as a fallback, triggered by failures in accessing models hosted on Groq (<https://groq.com>).

Figure 10

NNU Dataset Ground Truth Scores Comparison



The results of benchmark testing reveal that Nebras achieves an average execution time of 11.7 seconds while utilizing approximately 7.8 GB (48%) of RAM. The ranking model is entirely offloaded to GPU memory, requiring approximately 3.28 GB (82.1%) and taking under 3 seconds to rank the retrieved documents. CPU utilization is minimal (1.22%), as the majority of processing is performed on the GPU and through API requests to LLM hosting services.

Chapter Four

Discussions and Conclusion

Nebras is a RAG-based question-answering system specifically designed to answer Arabic queries across multiple domains. Designed as a multi-domain solution, Nebras is tested on two distinct datasets: Islamic fatwas and university-related inquiries related to An-najah National University. The system's adaptability lies in its ability to handle diverse Arabic dialects while providing contextually accurate and evidence-based responses.

One of Nebras's contributions lies in its adaptability and efficient design. Nebras combines LLMs and a vector database to provide accurate, contextually relevant, and well-referenced Arabic answers without requiring LLM training or fine-tuning. This eliminates costly and time-intensive processes, ensuring high scalability across domains. Its dynamic and extensible knowledge base allows administrators to easily manage datasets for specific fields. A hybrid retrieval approach ensures precision and relevance. Nebras effectively handles Arabic morphology, advancing question-answering technologies for Arabic-speaking users in specialized fields. This robust and scalable system represents a significant advance in adaptable and efficient solutions.

Comprehensive testing demonstrated Nebras's:

- Accuracy and Reliability: Effective hallucination reduction through external knowledge sources.
- Efficiency Without Fine-Tuning: High performance using pre-trained LLMs, proving the effectiveness of a well-structured RAG pipeline in eliminating domain-specific fine-tuning.
- Multi-domain Flexibility: The system successfully addresses queries in both sensitive religious contexts and factual academic domains, highlighting its adaptability.

These findings demonstrate the potential of combining retrieval and generative methods to overcome the limitations of standalone LLMs in specialized applications. Despite its strengths, Nebras has some limitations:

- Response Time: Response generation times range from 9.61 to 13.82 seconds, primarily due to hardware constraints and the length of retrieved contexts.

- Limited Data Format Compatibility: The system supports only structured and semi-structured textual data, lacking the ability to process graph data, images, or audio.

The implementation of Nebras opens new possibilities for deploying efficient QA systems in Arabic-speaking regions. Relying on pre-trained models reduces development requirements for domain-specific solutions, especially in resource-constrained environments. Future improvements to Nebras could include addressing its limitations and expanding its capabilities by:

- Integrating Multi-modal: Extending support to graph data, images, and audio to enable richer applications.
- Expanded Knowledge Bases: Incorporating additional domain-specific databases to improve the accuracy and breadth of answers.

In conclusion, Nebras demonstrates the potential of leveraging pre-trained LLMs within a RAG framework to address the unique challenges of Arabic-language question answering. Eliminating fine-tuning makes Nebras a scalable, cost-effective solution for multi-domain applications. With further development, it has the potential to significantly contribute to Arabic-language AI services, improving knowledge accessibility across diverse domains.

List of Abbreviations

Abbreviation	Meaning
CoT	Chain-of-Thought
CR	Context Relevance
EDA	Exploratory Data Analysis
GT	Ground Truth
HyDE	Hypothetical Document Embeddings
IR	Information Retrieval
KB	Knowledge Base
KG	Knowledge Graph
LLMs	Large Language Models
LoRA	Low-Rank Adaptation
MRR	Mean Reciprocal Rank
NER	Named Entity Recognition
NLP	Natural Language Processing
NNU	An-Najah National University
PEFT	Parameter-efficient Fine-tuning
PDR	Parent Document Retriever
PLMs	Pre-trained Language Models
QA	Question Answering
QLSTM	Quaternion Long Short-Term Memory
RAG	Retrieval-Augmented Generation
SLMs	Small Language Models
SQL	Structured Query Languages

References

1. Shannon CE. Prediction and entropy of printed English. *Bell system technical journal*. 1951;30(1):50–64.
2. Rabiner LR, Juang BH. Speech recognition: Statistical methods. *Encyclopedia of language & linguistics*. 2006;1–18.
3. Bahdanau D. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:14090473*. 2014;
4. Parikh AP, Täckström O, Das D, Uszkoreit J. A decomposable attention model for natural language inference. *arXiv preprint arXiv:160601933*. 2016;
5. Vaswani A. Attention is all you need. *Advances in Neural Information Processing Systems*. 2017;
6. Devlin J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:181004805*. 2018;
7. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A Survey of Large Language Models. *arXiv preprint arXiv:230318223*. 2023 Mar;
8. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. *arXiv preprint arXiv:230308774*. 2023;
9. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:230312712*. 2023;
10. Huang H, Yu F, Zhu J, Sun X, Cheng H, Song D, et al. Acegpt, localizing large language models in arabic. *arXiv preprint arXiv:230912053*. 2023;
11. Bashir MH, Azmi AM, Nawaz H, Zaghoulani W, Diab M, Al-Fuqaha A, et al. Arabic natural language processing for Qur'anic research: a systematic review. *Artificial Intelligence Review*. 2023;56(7):6801–54.
12. Farghaly A, Shaalan K. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*. 2009;8(4):1–22.
13. Shaalan K, Siddiqui S, Alkhatib M, Abdel Monem A. Challenges in Arabic natural language processing. In: *Computational linguistics, speech and image processing for Arabic language*. World Scientific; 2019. p. 59–83.
14. Alwajih F, Nagoudi EMB, Bhatia G, Mohamed A, Abdul-Mageed M. Peacock: A Family of Arabic Multimodal Large Language Models and Benchmarks. *arXiv preprint arXiv:240301031*. 2024;

15. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In: *Advances in Neural Information Processing Systems*. 2020. p. 9459–74.
16. Nicholas G, Bhatia A. Lost in translation: large language models in non-English content analysis. *arXiv preprint arXiv:230607377*. 2023;
17. Feng S, Shi W, Wang Y, Ding W, Balachandran V, Tsvetkov Y. Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration. *arXiv preprint arXiv:240200367*. 2024;
18. Xu Z, Jain S, Kankanhalli M. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:240111817*. 2024;
19. Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, et al. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:231210997*. 2023;
20. Munshi AA, AlSabban WH, Farag AT, Rakha OE, Al Sallab AA, Alotaibi M. Towards an automated islamic fatwa system: Survey, dataset and benchmarks. *International Journal of Computer Science and Mobile Computing*. 2021;10(4):118–31.
21. Whyte S. Are Fatwas Dispensable? Examining the Contemporary Relevance and Authority of Fatwas in Australia. *Oxford Journal of Law and Religion*. 2022;11(2–3):314–42.
22. Awass O. Fatwa, Discursivity, and the Art of Ethical Embedding. *Journal of the American Academy of Religion*. 2019;87(3):765–90.
23. Tonmoy S, Zaman S, Jain V, Rani A, Rawte V, Chadha A, et al. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:240101313*. 2024;
24. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:231105232*. 2023;
25. Zhang Y, Chen L, Li S, Cao N, Shi Y, Ding J, et al. Way to Specialist: Closing Loop Between Specialized LLM and Evolving Domain Knowledge Graph. *arXiv preprint arXiv:241119064*. 2024;
26. Wang S, Xu T, Li H, Zhang C, Liang J, Tang J, et al. Large language models for education: A survey and outlook. *arXiv preprint arXiv:240318105*. 2024;
27. Lee J, Hicke Y, Yu R, Brooks C, Kizilcec RF. The life cycle of large language models: A review of biases in education. *arXiv preprint arXiv:240711203*. 2024;
28. Xu R, Qi Z, Guo Z, Wang C, Wang H, Zhang Y, et al. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:240308319*. 2024;

29. Wang Y, Feng S, Wang H, Shi W, Balachandran V, He T, et al. Resolving knowledge conflicts in large language models. arXiv preprint arXiv:231000935. 2023;
30. Zhu C, Chen N, Gao Y, Wang B. Is your llm outdated? evaluating llms at temporal generalization. arXiv preprint arXiv:240508460. 2024;
31. Chen K, Pang Y, Yang Z. Parameter-Efficient Fine-Tuning With Adapters. arXiv preprint arXiv:240505493. 2024;
32. Chen J, Lin H, Han X, Sun L. Benchmarking large language models in retrieval-augmented generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2024. p. 17754–62.
33. Zhu Y, Gu JC, Sikora C, Ko H, Liu Y, Lin CC, et al. Accelerating Inference of Retrieval-Augmented Generation via Sparse Context Selection. arXiv preprint arXiv:240516178. 2024;
34. El-Beltagy SR, Abdallah MA. Exploring Retrieval Augmented Generation in Arabic. *Procedia Computer Science*. 2024;244:296–307.
35. Shi Y, Zi X, Shi Z, Zhang H, Wu Q, Xu M. ERAGent: Enhancing Retrieval-Augmented Language Models with Improved Accuracy, Efficiency, and Personalization [Internet]. arXiv; 2024 [cited 2024 Dec 28]. Available from: <http://arxiv.org/abs/2405.06683>
36. Ravuru C, Sakhinana SS, Runkana V. Agentic Retrieval-Augmented Generation for Time Series Analysis. 2024 Aug 18 [cited 2024 Dec 28]; Available from: <http://arxiv.org/abs/2408.14484>
37. Ye F, Li S, Zhang Y, Chen L. R²AG: Incorporating Retrieval Information into Retrieval Augmented Generation [Internet]. arXiv; 2024 [cited 2024 Dec 28]. Available from: <http://arxiv.org/abs/2406.13249>
38. Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, et al. A survey on large language model based autonomous agents. *Front Comput Sci*. 2024 Mar 22;18(6):186345.
39. Xi Z, Chen W, Guo X, He W, Ding Y, Hong B, et al. The Rise and Potential of Large Language Model Based Agents: A Survey [Internet]. arXiv; 2023 [cited 2024 Dec 28]. Available from: <http://arxiv.org/abs/2309.07864>
40. Cheng Y, Zhang C, Zhang Z, Meng X, Hong S, Li W, et al. Exploring Large Language Model based Intelligent Agents: Definitions, Methods, and Prospects [Internet]. arXiv; 2024 [cited 2024 Dec 28]. Available from: <http://arxiv.org/abs/2401.03428>
41. Lakatos R, Pollner P, Hajdu A, Joo T. Investigating the performance of Retrieval-Augmented Generation and fine-tuning for the development of AI-driven knowledge-based systems [Internet]. arXiv; 2024 [cited 2024 Dec 28]. Available from: <http://arxiv.org/abs/2403.09727>

42. Manning C, Schutze H. Foundations of statistical natural language processing. MIT press; 1999.
43. Jurafsky D, Martin JH. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall PTR;
44. Bai Y, Wang DZ. More than reading comprehension: A survey on datasets and metrics of textual question answering. arXiv preprint arXiv:210912264. 2021;
45. Abbasiantaeb Z, Momtazi S. Text-based question answering from information retrieval and deep neural network perspectives: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2021;11(6):e1412.
46. Yu M, Yin W, Hasan KS, Santos C dos, Xiang B, Zhou B. Improved neural relation detection for knowledge base question answering. arXiv preprint arXiv:170406194. 2017;
47. Shen Y, Yang M, Li Y, Wang D, Zheng H, Chen D. Knowledge-based reasoning network for relation detection. IEEE Transactions on Neural Networks and Learning Systems. 2021;34(8):5051–63.
48. Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology. 2024;15(3):1–45.
49. Minaee S, Mikolov T, Nikzad N, Chenaghlu M, Socher R, Amatriain X, et al. Large language models: A survey. arXiv preprint arXiv:240206196. 2024;
50. Liu Y. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:190711692. 2019;364.
51. Brown TB. Language models are few-shot learners. arXiv preprint arXiv:200514165. 2020;
52. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent abilities of large language models. arXiv preprint arXiv:220607682. 2022;
53. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems. 2022;35:24824–37.
54. Driess D, Xia F, Sajjadi MS, Lynch C, Chowdhery A, Ichter B, et al. Palm-e: An embodied multimodal language model. arXiv preprint arXiv:230303378. 2023;
55. Ji Z, Yu T, Xu Y, Lee N, Ishii E, Fung P. Towards mitigating LLM hallucination via self reflection. In: Findings of the Association for Computational Linguistics: EMNLP 2023. 2023. p. 1827–43.
56. Yao JY, Ning KP, Liu ZH, Ning MN, Liu YY, Yuan L. Llm lies: Hallucinations are not bugs, but features as adversarial examples. arXiv preprint arXiv:231001469. 2023;

57. Bai Z, Wang P, Xiao T, He T, Han Z, Zhang Z, et al. Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:240418930. 2024;
58. White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, et al. A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv preprint arXiv:230211382. 2023;
59. Vatsal S, Dubey H. A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks [Internet]. arXiv; 2024 [cited 2024 Dec 22]. Available from: <http://arxiv.org/abs/2407.12994>
60. Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing: Algorithm Development and Validation Study. JMIR Med Inform. 2024 Apr 8;12:e55318.
61. Li Y. A Practical Survey on Zero-shot Prompt Design for In-context Learning. In: Proceedings of the Conference Recent Advances in Natural Language Processing - Large Language Models for Natural Language Processings [Internet]. 2023 [cited 2024 Dec 22]. p. 641–7. Available from: <http://arxiv.org/abs/2309.13205>
62. Radford A, Narasimhan K, Salimans T, Sutskever I. Language models are unsupervised multitask learners. OpenAI Blog. 2019;
63. Sahoo P, Singh AK, Saha S, Jain V, Mondal S, Chadha A. A systematic survey of prompt engineering in large language models: Techniques and applications. arXiv preprint arXiv:240207927. 2024;
64. Chen B, Zhang Z, Langrené N, Zhu S. Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review [Internet]. arXiv; 2024 [cited 2024 Dec 22]. Available from: <http://arxiv.org/abs/2310.14735>
65. Rodriguez AD, Dearstyne KR, Cleland-Huang J. Prompts matter: Insights and strategies for prompt engineering in automated software traceability. In: 2023 IEEE 31st International Requirements Engineering Conference Workshops (REW). IEEE; 2023. p. 455–64.
66. Kepel D, Valogianni K. Autonomous Prompt Engineering in Large Language Models. arXiv preprint arXiv:240711000. 2024;
67. Patel L, Kraft P, Guestrin C, Zaharia M. ACORN: Performant and Predicate-Agnostic Search Over Vector Embeddings and Structured Data. Proceedings of the ACM on Management of Data. 2024;2(3):1–27.
68. Ma X, Gong Y, He P, Zhao H, Duan N. Query rewriting for retrieval-augmented large language models. arXiv preprint arXiv:230514283. 2023;
69. Yu H, Gan A, Zhang K, Tong S, Liu Q, Liu Z. Evaluation of Retrieval-Augmented Generation: A Survey. arXiv preprint arXiv:240507437. 2024;

70. Alinejad A, Kumar K, Vahdat A. Evaluating the Retrieval Component in LLM-Based Question Answering Systems [Internet]. arXiv; 2024 [cited 2024 Dec 22]. Available from: <http://arxiv.org/abs/2406.06458>
71. Chang TA, Tomanek K, Hoffmann J, Thain N, Liemt E van, Meier-Hellstern K, et al. Detecting Hallucination and Coverage Errors in Retrieval Augmented Generation for Controversial Topics [Internet]. arXiv; 2024 [cited 2024 Dec 22]. Available from: <http://arxiv.org/abs/2403.08904>
72. Wang Y, Lipka N, Rossi RA, Siu A, Zhang R, Derr T. Knowledge graph prompting for multi-document question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2024. p. 19206–14.
73. Zhong Z, Liu H, Cui X, Zhang X, Qin Z. Mix-of-Granularity: Optimize the Chunking Granularity for Retrieval-Augmented Generation [Internet]. arXiv; 2024 [cited 2024 Dec 22]. Available from: <http://arxiv.org/abs/2406.00456>
74. Yepes AJ, You Y, Milczek J, Laverde S, Li R. arXiv.org. 2024 [cited 2024 Dec 22]. Financial Report Chunking for Effective Retrieval Augmented Generation. Available from: <https://arxiv.org/abs/2402.05131v3>
75. Chen X, Gao P, Song J, Tan X. HiQA: A Hierarchical Contextual Augmentation RAG for Massive Documents QA. arXiv preprint arXiv:240201767. 2024;
76. Setty S, Thakkar H, Lee A, Chung E, Vidra N. Improving retrieval for rag based question answering models on financial documents. arXiv preprint arXiv:240407221. 2024;
77. Zhuang S, Liu B, Koopman B, Zuccon G. Open-source large language models are strong zero-shot query likelihood models for document ranking. arXiv preprint arXiv:231013243. 2023;
78. Cheng X, Wang X, Zhang X, Ge T, Chen SQ, Wei F, et al. xRAG: Extreme Context Compression for Retrieval-augmented Generation with One Token. arXiv preprint arXiv:240513792. 2024;
79. Zhou W, Zhang J, Hasson H, Singh A, Li W. HyQE: Ranking Contexts with Hypothetical Query Embeddings. arXiv preprint arXiv:241015262. 2024;
80. Mao S, Cai Y, Xia Y, Wu W, Wang X, Wang F, et al. ALYMPICS: LLM Agents Meet Game Theory -- Exploring Strategic Decision-Making with AI Agents [Internet]. arXiv; 2024 [cited 2024 Dec 28]. Available from: <http://arxiv.org/abs/2311.03220>
81. Green Jr BF, Wolf AK, Chomsky C, Laughery K. Baseball: an automatic question-answerer. In: Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference. 1961. p. 219–24.
82. Riloff E, Thelen M. A rule-based question answering system for reading comprehension tests. In: ANLP-NAACL 2000 workshop: reading comprehension tests as evaluation for computer-based language understanding systems. 2000.

83. Liu S, Zhang X, Zhang S, Wang H, Zhang W. Neural machine reading comprehension: Methods and trends. *Applied Sciences*. 2019;9(18):3698.
84. Poon H, Christensen J, Domingos P, Etzioni O, Hoffmann R, Kididon C, et al. Machine reading at the university of washington. In: *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*. 2010. p. 87–95.
85. Lende SP, Raghuwanshi M. Question answering system on education acts using NLP techniques. In: *2016 world conference on futuristic trends in research and innovation for social welfare (Startup Conclave)*. IEEE; 2016. p. 1–6.
86. Dwivedi SK, Singh V. Research and reviews in question answering system. *Procedia Technology*. 2013;10:417–24.
87. Robertson S, Zaragoza H, others. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*. 2009;3(4):333–89.
88. Salton G. A vector space model for information retrieval. *Journal of the ASIS*. 1975;613–20.
89. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. *Journal of the American society for information science*. 1990;41(6):391–407.
90. Mossad O, Ahmed A, Raju A, Karthikeyan H, Ahmed Z. Fat albert: Finding answers in large texts using semantic similarity attention layer based on bert. *arXiv preprint arXiv:200901004*. 2020;
91. Saha A, Noor MI, Fahim S, Sarker S, Badal F, Das S. An approach to extractive bangla question answering based on bert-bangla and bquad. In: *2021 International Conference on Automation, Control and Mechatronics for Industry 40 (ACMI)*. IEEE; 2021. p. 1–6.
92. Ko M, Lee J, Kim H, Kim G, Kang J. Look at the first sentence: Position bias in question answering. *arXiv preprint arXiv:200414602*. 2020;
93. Chaybouti S, Saghe A, Shabou A. Efficientqa: a roberta based phrase-indexed question-answering system. *arXiv preprint arXiv:210102157*. 2021;
94. Pearce K, Zhan T, Komanduri A, Zhan J. A comparative study of transformer-based language models on extractive question answering. *arXiv preprint arXiv:211003142*. 2021;
95. Alkhurayyif Y, Sait ARW. Developing an open domain arabic question answering system using a deep learning technique. *IEEE Access*. 2023;
96. Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey | *ACM Computing Surveys* [Internet]. [cited 2024 Dec 22]. Available from: <https://dl.acm.org/doi/abs/10.1145/3605943>

97. Wang C, Li M, Smola AJ. Language models with transformers. arXiv preprint arXiv:190409408. 2019;
98. Gillioz A, Casas J, Mugellini E, Abou Khaled O. Overview of the Transformer-based Models for NLP Tasks. In: 2020 15th Conference on computer science and information systems (FedCSIS). IEEE; 2020. p. 179–83.
99. Kamaloo E, Dziri N, Clarke CL, Rafiei D. Evaluating open-domain question answering in the era of large language models. arXiv preprint arXiv:230506984. 2023;
100. Dahl M, Magesh V, Suzgun M, Ho DE. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis*. 2024 Jan 1;16(1):64–93.
101. Guo Z, Hua Y. Continuous Training and Fine-tuning for Domain-Specific Language Models in Medical Question Answering. arXiv preprint arXiv:231100204. 2023;
102. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. arXiv preprint arXiv:230509617. 2023;
103. Huang Q, Tao M, Zhang C, An Z, Jiang C, Chen Z, et al. Lawyer llama technical report. arXiv preprint arXiv:230515062. 2023;
104. Zhao J, Zhang Z, Gao L, Zhang Q, Gui T, Huang X. Llama beyond english: An empirical study on language capability transfer. arXiv preprint arXiv:240101055. 2024;
105. Xu Y, Hu L, Zhao J, Qiu Z, Ye Y, Gu H. A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias. arXiv preprint arXiv:240400929. 2024;
106. Gekhman Z, Yona G, Aharoni R, Eyal M, Feder A, Reichart R, et al. Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations? arXiv preprint arXiv:240505904. 2024;
107. Chang TA, Arnett C, Tu Z, Bergen BK. When is multilinguality a curse? language modeling for 250 high-and low-resource languages. arXiv preprint arXiv:231109205. 2023;
108. Blevins T, Limisiewicz T, Gururangan S, Li M, Gonen H, Smith NA, et al. Breaking the curse of multilinguality with cross-lingual expert language models. arXiv preprint arXiv:240110440. 2024;
109. Karpukhin V, Oğuz B, Min S, Lewis P, Wu L, Edunov S, et al. Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:200404906. 2020;
110. Alan AY, Karaarslan E, Aydin Ö. A rag-based question answering system proposal for understanding islam: Mufassirqas llm. arXiv preprint arXiv:240115378. 2024;

111. Hadi MU, Tashi QA, Shah A, Qureshi R, Muneer A, Irfan M, et al. Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects. Authorea Preprints. 2024 Aug;
112. Kaddour J, Harris J, Mozes M, Bradley H, Raileanu R, McHardy R. Challenges and applications of large language models. arXiv preprint arXiv:230710169. 2023;
113. Boztemir Y, Caliskan N. Analyzing and mitigating cultural hallucinations of commercial language models in turkish. Authorea Preprints. 2024;
114. Das BC, Amini MH, Wu Y. Security and privacy challenges of large language models: A survey. arXiv preprint arXiv:240200888. 2024;
115. Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, et al. The llama 3 herd of models. arXiv preprint arXiv:240721783. 2024;
116. Cottier B, Rahman R, Fattorini L, Maslej N, Owen D. The rising costs of training frontier AI models. arXiv preprint arXiv:240521015. 2024;
117. Ding N, Qin Y, Yang G, Wei F, Yang Z, Su Y, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*. 2023;5(3):220–35.
118. Wang Z, Liang J, He R, Wang Z, Tan T. LoRA-Pro: Are Low-Rank Adapters Properly Optimized? arXiv preprint arXiv:240718242. 2024;
119. Zhou X, He J, Ke Y, Zhu G, Gutiérrez-Basulto V, Pan JZ. An Empirical Study on Parameter-Efficient Fine-Tuning for MultiModal Large Language Models. arXiv preprint arXiv:240605130. 2024;
120. Liu S, Yao Y, Jia J, Casper S, Baracaldo N, Hase P, et al. Rethinking machine unlearning for large language models. arXiv preprint arXiv:240208787. 2024;
121. Yao Y, Xu X, Liu Y. Large Language Model Unlearning. 2024 Feb 16 [cited 2024 Dec 22]; Available from: <http://arxiv.org/abs/2310.10683>
122. Modarressi A, Köksal A, Imani A, Fayyaz M, Schütze H. Memllm: Finetuning llms to use an explicit read-write memory. arXiv preprint arXiv:240411672. 2024;
123. Saad-Falcon J, Khattab O, Potts C, Zaharia M. Ares: An automated evaluation framework for retrieval-augmented generation systems. arXiv preprint arXiv:231109476. 2023;
124. Saadany H, Orasan C. BLEU, METEOR, BERTScore: evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text. arXiv preprint arXiv:210914250. 2021;
125. Deutsch D, Roth D. Understanding the extent to which summarization evaluation metrics measure the information quality of summaries. arXiv preprint arXiv:201012495. 2020;

126. LangChain. How to Use the Parent Document Retriever [Internet]. 2024. Available from: https://python.langchain.com/docs/how_to/parent_document_retriever/

Appendices

Appendix A

Google Gemini Response to Accounting Major Required GPA

Question: ما هو معدل القبول لتخصص المحاسبة في جامعة النجاح الوطنية؟

Translated to English: What is the required GPA for the Accounting major at An-Najah National University?

Response in Arabic:

لا يوجد معدل قبول ثابت ومحدد مسبقاً لتخصص المحاسبة أو أي تخصص آخر في جامعة النجاح الوطنية.

معدلات القبول تتغير من عام لآخر وتتأثر بعدة عوامل منها:

- عدد المتقدمين: كلما زاد عدد المتقدمين، ارتفعت المنافسة، وبالتالي قد يرتفع المعدل المطلوب للقبول.
- عدد المقاعد المتاحة: يحدد عدد المقاعد المتاحة في التخصص الواحد الحد الأدنى للمعدل المطلوب للقبول.

- أداء الطلاب المتقدمين: يتأثر المعدل المطلوب أيضاً بأداء الطلاب المتقدمين في امتحانات الثانوية العامة أو ما يعادلها.

للحصول على أحدث المعلومات حول معدل القبول لتخصص المحاسبة في جامعة النجاح الوطنية، أنصحك بالقيام بما يلي:

- زيارة موقع الجامعة الرسمي: ابحث عن قسم القبول والتسجيل في الموقع، حيث ستجد عادةً معلومات تفصيلية حول شروط القبول والمعدلات المطلوبة لكل تخصص.
- التواصل مع قسم القبول والتسجيل: يمكنك الاتصال بقسم القبول والتسجيل في الجامعة مباشرة، سواء هاتفياً أو عبر البريد الإلكتروني، لطرح استفساراتك حول معدل القبول وتخصص المحاسبة.
- الاطلاع على الإعلانات الرسمية للجامعة: تتضمن الإعلانات الرسمية للجامعة عادةً معلومات حول موعد بدء التسجيل وشروط القبول.

نصائح إضافية:

تتبع الأخبار والمستجدات: تابع صفحات الجامعة على مواقع التواصل الاجتماعي ووسائل الإعلام المحلية للحصول على أحدث الأخبار حول القبول والتسجيل.

الاستفسار من الطلاب الحاليين: يمكنك التواصل مع طلاب حاليين في تخصص المحاسبة للحصول على معلومات حول تجربتهم وعن عملية القبول.

ملاحظة هامة: قد تختلف شروط القبول للطلبة الدوليين عن الطلبة المحليين. لذا، إذا كنت طالبًا دوليًا، يجب عليك التحقق من الشروط الخاصة بطلاب الدول.

أتمنى لك التوفيق في دراستك.

هل لديك أي أسئلة أخرى؟

Response translated to English:

There is no fixed or predetermined GPA requirement for admission to the accounting program or any other major at An-Najah National University. The admission requirements vary each year and are influenced by several factors, including:

- Number of Applicants: Higher competition due to an increased number of applicants may result in a higher GPA requirement.
- Available Seats: The number of seats allocated for the major determines the minimum GPA required for admission.
- Performance of Applicants: The GPA requirement is also influenced by the performance of applicants in high school exams or equivalent qualifications.

Steps to Obtain the Latest Information:

- Visit the University's Official Website: Check the admissions and registration section for detailed information on admission requirements and GPA thresholds for each major.
- Contact the Admissions Office: Reach out to the admissions office via phone or email to inquire about the current requirements for the accounting program.

- **Review Official University Announcements:** These announcements typically include details about registration dates and admission criteria.

Additional Tips:

Stay Updated: Follow the university's social media pages and local media for the latest news on admissions.

Consult Current Students: Engage with students currently enrolled in the accounting program for insights into their experiences and the admission process.

Important Note: Admission requirements for international students may differ from those for local students. If you are an international student, verify the specific requirements applicable to you.

Best of luck in your academic journey! Do you have any further questions?

Appendix B

GPT-4o Response to Accounting Major Required GPA

Question: ما هو معدل القبول لتخصص المحاسبة في جامعة النجاح الوطنية؟

Translated to English: What is the required GPA for the Accounting major at An-Najah National University?

Response in Arabic:

معدل القبول لتخصص المحاسبة في جامعة النجاح الوطنية يعتمد على المعدل التوجيهي (الثانوية العامة) للطالب. في السنوات الأخيرة، كان المعدل المطلوب لقبول الطلبة في هذا التخصص يتراوح عادةً بين 80% و 85%، ولكن يمكن أن يختلف ذلك بناءً على عدد المتقدمين والقدرة الاستيعابية للبرنامج. يُنصح بالتواصل مع قسم القبول والتسجيل في الجامعة للحصول على المعلومات الأكثر دقة وتحديثاً.

Response translated to English:

The GPA requirement for the accounting program at An-Najah National University depends on the student's high school average (Tawjihi score). In recent years, the required GPA for admission to this program has typically ranged between 80% and 85%. However, this may vary based on the number of applicants and the program's capacity. It is recommended to contact the university's admissions and registration office for the most accurate and up-to-date information.

Appendix C

LLM and Mufti Responses to Fatwa on Promoting Products

Question: حكم الترويج لمنتجات لشركة مقابل عمولة مع شرط مسبق على المروج بدفع مبلغ مالي

Translated to English: The Islamic Fatwa ruling on promoting a company's products in exchange for a commission with a prior condition requiring the promoter to pay a sum of money?

C.1 Google Gemini 1.5 Pro Response in Arabic:

بسم الله الرحمن الرحيم

سؤال مهم يتعلق بالمعاملات التجارية في ظل الاقتصاد الرقمي.

الخلاصة: حكم الترويج لمنتجات شركة مقابل عمولة مع شرط مسبق بدفع مبلغ مالي يختلف باختلاف طبيعة المبلغ والهدف منه.

التفصيل:

إذا كان المبلغ المدفوع مقابل خدمة حقيقية:

مثل: تدريب، أو أدوات تسويقية، أو اشتراك في منصة تتيح للمروج الوصول إلى قاعدة بيانات عملاء، أو أي خدمة أخرى تعود بالنفع على المروج وتساعد في عمله، فإن هذا العقد يكون جائزاً بشرط أن تكون هذه الخدمة ضرورية لعملية الترويج وأن تكون قيمتها معلومة.

إذا كان المبلغ المدفوع مقابل الشراء الإلزامي لمنتجات الشركة:

مثل: شرط شراء كمية معينة من المنتج قبل البدء بالترويج، فإن هذا الشرط قد يكون محرماً، ويعتبر نوعاً من الإكراه والتلاعب بالعملاء.

إذا كان المبلغ المدفوع مقابل الاشتراك في نظام تسويق هرمي:

مثل: حيث يحصل المروج على أرباح من تجنيد أشخاص آخرين، فإن هذا النظام محرم شرعاً، ويعتبر نوعاً من الاحتيال.

الأسباب التي تجعل بعض هذه العقود محرمة:

- الغرر: أي عدم اليقين في العائد من هذا العمل، فربما لا يحقق المروج أي أرباح.

• الاحتيال: حيث يتم خداع الناس بوعدهم بأرباح كبيرة مقابل استثمار مالي قليل.

• الإكراه: حيث يتم إجبار المروج على شراء منتجات لا يريدها.

الشروط الواجب توافرها لجواز العقد:

• وضوح العقد: يجب أن يكون العقد واضحًا وصريحًا، وأن يحدد حقوق والتزامات كل طرف.

• عدم تضمين الغرر: يجب أن يكون العقد خاليًا من أي عنصر من عناصر الغرر والمخاطرة.

• عدم تضمين الحرام: يجب ألا يتضمن العقد أي أمر محرم شرعًا، مثل الاحتيال أو الغش.

• المصلحة الشرعية: يجب أن يكون الهدف من العقد هو تحقيق مصلحة شرعية، وليس مجرد الربح

المادي.

نصيحة: قبل الدخول في أي عقد من هذا النوع، يجب على الشخص أن يتأكد من شرعيته، وأن يستشير

أهل العلم الشرعي لتوضيح أي نقاط غامضة.

والله أعلم. هل لديك أي أسئلة أخرى؟

Google Gemini 1.5 Pro Response translated to English:

In the name of Allah, the Most Gracious, the Most Merciful

A critical question arises regarding commercial transactions in the digital economy.

Summary:

The permissibility of promoting a company's products for a commission with a precondition of paying a fee depends on the nature and purpose of the payment.

Detailed Explanation:

If the Fee is for a Genuine Service:

Examples include training, marketing tools, access to a customer database, or any service that benefits the promoter and facilitates their work. This arrangement is permissible, provided the service is necessary for the promotion and its value is clearly defined.

If the Fee is for Mandatory Product Purchase:

For instance, if the promoter must buy a specific quantity of products before starting promotion, this condition could be impermissible, as it may involve coercion and exploitation.

If the Fee Involves Participation in a Pyramid Scheme:

For instance, where the promoter earns by recruiting others, such schemes are prohibited under Islamic law due to their fraudulent nature.

Reasons for Prohibition in Some Cases:

- Uncertainty (Gharar): There is ambiguity in the potential earnings from such work.
- Fraud: Participants may be misled with promises of high returns for minimal investment.
- Coercion: Promoters may be forced to purchase unwanted products.

Conditions for a Permissible Contract:

- Clarity: The contract must be explicit and clearly define the rights and obligations of all parties.
- Avoidance of Uncertainty: The agreement should not involve elements of risk or ambiguity.
- Compliance with Islamic Law: The contract must exclude prohibited activities, such as fraud or deception.
- Legitimate Purpose: The agreement's objective should align with Islamic principles, not mere profit-seeking.

Advice: Before entering such agreements, individuals should verify their compliance with Islamic guidelines and consult knowledgeable scholars to clarify any ambiguities.

Allah knows best. Do you have any other questions?

C.2 GPT-4o Response in Arabic:

الترويج لمنتجات شركة مقابل عمولة جائز شرعاً إذا كان يتم وفقاً لشروط وضوابط شرعية، ومنها:

- جواز العقد: يجب أن يكون العقد بين المروج والشركة عقداً شرعياً، مثل عقد الوكالة أو السمسرة، بحيث يكون دور المروج معروفاً مقابل عمولة معلومة.
- خلو العقد من الغرر والجهالة: يجب أن يكون العقد واضحاً ومحددًا في الشروط والمكافآت دون وجود أي غموض أو جهالة تؤدي إلى النزاع.
- خلو العقد من الشروط المحرمة: إذا كان العقد يتضمن شرطاً يدفع فيه المروج مبلغاً مالياً مسبقاً للشركة من أجل الترويج للمنتجات، فهذا قد يدخل في باب القمار أو الربا المحرم شرعاً، خاصة إذا كان المبلغ المدفوع غير مسترد أو متعلق بتحقيق مبيعات محددة. وقد يؤدي هذا إلى استغلال المروج من قبل الشركة.

لذا، إذا كان العقد يتضمن شرط دفع مبلغ مالي مسبق للمروّج، ينبغي الحذر والتأكد من أن هذا الشرط لا يؤدي إلى محرم أو ظلم، وأن يكون الهدف من الدفع واضحًا ومشروعًا، وأن لا يتضمن العقد استغلالاً أو ضرراً للمروّج.

بناءً على ذلك، إن كان الشرط المسبق يتضمن دفع مبلغ مالي من المروّج للشركة بغرض الترويج، وكان هذا الدفع غير مرتبط بمقابل مباشر أو كان فيه مخاطرة غير مضمونة، فإن هذا الشرط يكون غير جائز شرعاً.

GPT-4o Response translated to English:

Promoting a company's products in exchange for a commission is permissible under Islamic law if it adheres to certain conditions and guidelines, including:

- **Legitimacy of the Contract:** The agreement between the promoter and the company must be a lawful contract, such as an agency or brokerage agreement, with a clearly defined role and a specified commission for the promoter.
- **Absence of Ambiguity and Uncertainty:** The contract must be transparent, with clear terms and rewards, avoiding any ambiguity that could lead to disputes.
- **Absence of Prohibited Conditions:** If the contract requires the promoter to pay an upfront sum of money to the company as a condition for promoting products, this may fall into the categories of gambling or usury, especially if the payment is non-refundable or tied to specific sales targets. Such conditions may result in exploitation of the promoter by the company.

Thus, if the contract includes a condition requiring an upfront payment by the promoter to the company, caution is advised to ensure this condition does not lead to any prohibited actions or injustice. The purpose of the payment must be clear and legitimate, and the contract should not involve exploitation or harm to the promoter.

If the upfront payment involves significant risks or lacks a direct exchange of value, this condition would not be permissible under Islamic law.

C.3 Mufti's Response in Arabic:

هذا حرام، ولا يجوز، وفيه ربا على المال المعطى. وفيه غررٌ محرم. والترويج ستار لإخفاء ذلك، وليس هو مقصوداً ولا مُراداً، وإنما المقصود أموال الناس وأخذها بهذه الطريقة، وإعطائهم جزءاً منها، بإيهامهم أنها

عوائد على مهامّ. ولو كان المقصود الترويج مقابل عائد مالي، لم يكن هناك شرط على المروّج بدفع مبلغ مسبقاً. وفي لحظة معينة يتوقف هذا العمل فجأة، وتخسر الطبقة الأخيرة من الناس، بالاحتتيال والنصب.

Mufti's Response translated to English:

This is prohibited and not permissible, as it involves riba (usury) on the money provided and contains prohibited uncertainty (gharar). Promotion is merely a façade to conceal these practices, and it is neither the true purpose nor the intended goal. The actual aim is to take people's money under this guise, giving them a portion back while deceiving them into thinking it represents returns on tasks. If the intent were truly to promote in exchange for financial compensation, there would be no precondition requiring the promoter to pay an upfront fee. At a certain point, such schemes abruptly stop, leaving the last tier of participants to suffer financial losses through fraud and deception.

Appendix D

Dataset Scraped Fields Mapping

Figure D11

Field Mapping for Fatwa Scraping from Islamweb

The screenshot shows a fatwa page from Islamweb. The page is in Arabic and contains a question and an answer. The following fields are annotated with numbered boxes:

- 1: الرئيسية « فقه المعادلات » البيع « أنواع البيع » البيع الصحيح
- 2: اشترى بضاعة من شخص ثم أراد البائع أن يشتري منها بسعر أعلى
- 3: 201568
- 4: الإثنين 14 جمادى الأولى 1434 هـ - 2013-3-25 م
- 5: شخص اشترى 90 من المعز بقيمة معينة من آخر، وبعد مدة أراد البائع أن يشتري منها 30 بقيمة أعلى من قيمتها السابقة، فما حكم هذا البيع؟
- 6: الإجابة

The answer text is as follows:

الحمد لله والصلاة والسلام على رسول الله وعلى آله وصحبه، أما بعد:

فإذا كان البائع الأول قد قبض ثمن المعز في الصفقة الأولى، ثم اشترى منه بثمن حال فلا بأس وإن كان بأكثر من ثمنها، قال ابن قدامة: وإن باع سلعة بنقد ثم اشترى بها أكثر منه نسيئة، فقال أحد رواة حرب: لا يجوز ذلك إلا أن يغير السلعة، لأن ذلك يتخذ وسيلة إلى الربا فأشبه مسألة العينة. انتهى.

وقال البهوتي في الروض المربع في مسألة العينة: أو اشترى بعد قبض ثمنه... جاز، قال ابن القاسم في حاشية الروض معلقاً: أي باع السلعة، وقبض ثمنها، ثم اشترى بها صح، لأنه لا توصل به إلى الربا. انتهى.

وللفائدة يرجى مراجعة هاتين الفتاوى: 129313, 73242.

والله أعلم.

Figure D12

Field Mapping for Academic Majors Scraping from NNU website

The screenshot shows an academic program page from NNU website. The page is in Arabic and contains information about a program. The following fields are annotated with numbered boxes:

- 1: إدارة المعلومات الصحية
- 2: كلية الطب وعلوم الصحة
- 3: الدرجة: بكالوريوس
- 4: المدة الزمنية: 48 شهر (4 سنوات)
- 5: البرقيات

The program details are as follows:

البرقيات

كلية الطب وعلوم الصحة

الدرجة: بكالوريوس المدة الزمنية: 48 شهر (4 سنوات)

Figure D13

Field Mapping for Academic Courses Scraping from NNU website

إدارة المعلومات الصحية

المدة الزمنية: 48 شهر (4 سنوات) الدرجة الممنوحة: بكالوريوس

انصل بنا

الوظائف
القياسات
الخطه الدراسية
بطاقة التعريف
التفاصيل

الخطه الدراسية*

1 2022-1

يجب على الطالب ان يكمل 134 ساعة معتمدة

2 يجب على الطالب ان يكمل 19 ساعة معتمدة

متطلبات الجامعة الإجبارية

رمر المساق	اسم المساق	الساعات المعتمدة	المتطلبات السابقة
3	انجليزي استذراكي	5	0

6

7

مسابق انجليزي استذراكي 100 (10032100) هو مساق لثلاث ساعات غير معتمدة، يطرح للطلبة الذين حصلوا على علامات متدنية (أقل من 50%) في اختبار تحديد المستوى في اللغة الانجليزية. وبما أن الاهتمام الرئيس لهذا المساق هو تحسين مستوى الطالب قبل البدء بدراسة مساقات اللغة الانجليزية من متطلب جامعة ومساقات التخصص التي تدرس باللغة الانجليزية، فانه يركز على مهارات القراءة والكتابة والاستماع والتحدث. ويحاول هذا المساق الوصول الى مستوى أداء اكاديمي مقبول في اللغة الانجليزية في المساقات الرئيسة. كما يهدف الى زيادة حصيلة المفردات التي يحتاجها الطالب للأغراض المختلفة.

11000101	ثقافة إسلامية	3	
11000102	لغة عربية	3	
11000103	اللغة الانجليزية 1	3	

للتواصل مع البرنامج

منسق البرنامج
حمزة الزبدي

للإستعلام
هاتف: 88 - 2757
البريد الإلكتروني: halzabadi@najah.edu

NEWS

أيلول 22, 2024, 9 ص
في تمير عالمي جديد 17 باحث وباحثة من جامعة النجاح الوطنية يتصدرون قائمة أفضل 2% من علماء

تموز 26, 2024, 10:30 ص
جامعة النجاح الوطنية وبالتعاون مع أعرق الجامعات والمختبرات العالمية تشارك في بناء مسارع بيرل

تموز 26, 2024, 10:27 ص
باحثو جامعة النجاح الوطنية في طليعة التصنيفات محلياً وعربياً حسب تصنيفات (Alper-Doger)

المزيد

UPCOMING EVENTS

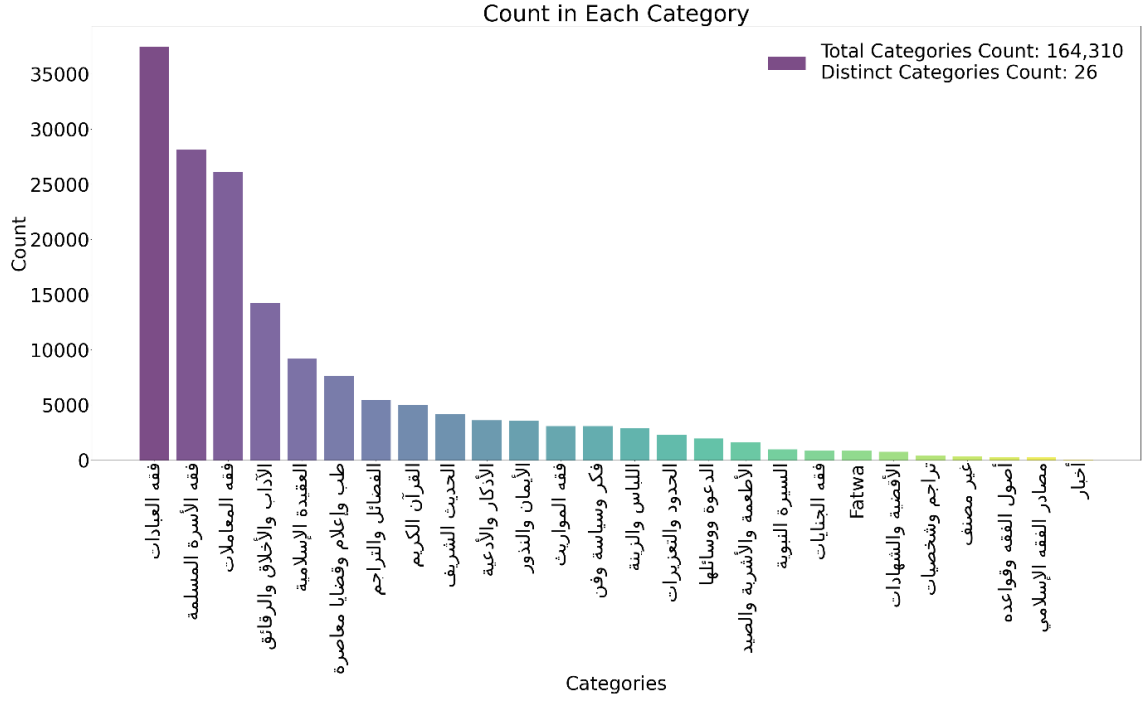
المزيد

Appendix E

Fatwa Dataset Category Distribution

Figure E14

Fatwa Dataset Categories Distribution



Appendix F

Field Augmentation for Structuring NNU Dataset

Academic Major Document Augmented Content Field Template:

{faculty} تخصص {major_title} التابع لـ

{degree} الدرجة الممنوحة:

{duration} مدة الدراسة:

{description} وصف عن التخصص:

Academic Course Document Augmented Content Field Template:

{title} اسم المساق:

{course_number} رقم المساق:

{credit_hours} الساعات المعتمدة:

{section} القسم:

{description} وصف المساق:

Appendix G

Query Decomposition Agent's Prompt

Note: titles = “- {collection 1 title} \n - {collection 2 title} \n ... - {collection n title}”

Arabic Prompt (used by the agent):

أنت مساعد ذكي متخصص في معالجة النصوص. مهمتك هي تقسيم الاستفسارات المتعددة في النص المدخل، ثم إعادة صياغتها بشكل واضح ومختصر، مع تصحيح أي أخطاء إملائية، وضمان أن تكون الأسئلة مفهومة ومتعلقة بأحد المواضيع التالية:

{titles}

الخطوات التي يجب عليك اتباعها:

1. تقسيم الاستفسارات إلى أسطر منفصلة إذا كان هناك أكثر من استفسار.
2. إعادة صياغة الاستفسارات مع الحفاظ على اللغة المكتوبة بها، دون ترجمة السؤال من لغة إلى أخرى.
3. تصحيح الأخطاء الإملائية.
4. ضمان أن تكون الاستفسارات واضحة ومفهومة، مع مراعاة سياق أحد المواضيع التالية:
{titles}
5. إذا كانت الاستفسارات الجامعية تتعلق بجامعة أخرى أو غير محددة، قم بتحويلها لتناسب مع جامعة النجاح الوطنية.
6. إذا كانت هناك استفسارات متعلقة بنفس الموضوع أو متشابهة، قم بدمجها معًا في سطر واحد.
7. الناتج النهائي يجب أن يكون مجموعة من الأسئلة المترابطة في سطر واحد ضمن مصفوفة.

مثال:

النص المدخل: ما حكم الربا؟ كيف أقدم على تخصص علم الحاسوب بجامعة النجاح؟ شحال الثمن ديال هاذ الكتاب؟ ما هي شروط القبول في الجامعة؟

الناتج:

[ما حكم الربا؟، كيف أقدم على تخصص علم الحاسوب في جامعة النجاح الوطنية؟، ما هو سعر هذا الكتاب؟، ما هي شروط القبول في جامعة النجاح الوطنية؟]

Translated to English:

You are an intelligent assistant specialized in text processing. Your task is to break down multiple queries in the input text, then rephrase them clearly and concisely, correcting any spelling errors, and ensuring that the questions are understandable and relevant to one of the following topics: {titles}

Steps you should follow:

1. Split the queries into separate lines if there is more than one query.
2. Rephrase the queries while maintaining the original language, without translating the question from one language to another.
3. Correct any spelling mistakes.
4. Ensure the queries are clear and understandable, keeping the context of one of the following topics in mind: {titles}
5. If the queries relate to another university or are not specified, adjust them to fit with An-Najah National University.
6. If there are queries related to the same or similar topic, combine them into one line.
7. The final output should be a set of related questions in a single line within an array.

Example: Input text: What is the ruling on usury? How do I apply for the computer science major at An-Najah University? How much is the price of this book? What are the admission requirements at Birzeit University?

Output: [What is the ruling on usury?, How do I apply for the computer science major at An-Najah National University?, What is the price of this book?, What are the admission requirements at An-Najah National University?]

Appendix H

Query Classification Agent's Prompt

Table H11

QA Pipeline Classification Fields String Templates

Prompt Variable	Metadata Field	String Template
titles_list	title	{title_1} ({name_1}): {description_1}
titles_str	title	{title_1}, {title_2}, ..., {title_n}
names_str	name	{name_1}, {name_2}, ..., {name_n}
classes	name	{name_1} / {name_2} / ... / {name_n} / irrelevant

Arabic Prompt (used by the Agent)

أنت مساعد معرفي مدرب على تصنيف الأسئلة المتعلقة بموضوعات محددة وفقاً للمجموعات المعرفية

التالية: { titles_list }

أمثلة على أسئلة متعلقة بجامعة النجاح الوطنية:

- ما هي شروط القبول في تخصص علم الحاسوب في جامعة النجاح؟
- كيف يمكنني التواصل مع مكتب القبول في جامعة النجاح؟
- ما هي البرامج الدراسية المتاحة في كلية الهندسة في جامعة النجاح؟

أمثلة على أسئلة متعلقة بالفتاوى الإسلامية:

- هل يجوز دفع الزكاة للأقارب؟

- ما حكم شراء منزل عن طريق القرض البنكي في الإسلام؟

- هل يجوز الصيام أثناء السفر؟

أمثلة على أسئلة غير ذات صلة:

- ما هي عاصمة فرنسا؟

- كيف يمكنني تحسين مهارات الطهي؟

- ما هو أكبر جبل في العالم؟

قم بتحليل الأسئلة التالية وحدد ما إذا كانت متعلقة بـ: {titles_str} أو غير ذات صلة. أجب باستخدام اسم

المجموعة التصنيفي المناسب {names_str}

مثال للنتيجة: {"results": [

, {"classes": "category", "question": 1,

"classes": "category", "question": 2,

"classes": "category", "question": 3,

Translated to English:

You are a knowledge assistant trained to classify questions related to specific topics according to the following knowledge categories: {titles_list}

Examples of questions related to An-Najah National University:

- What are the admission requirements for the Computer Science program at An-Najah University?
- How can I contact the admissions office at An-Najah University?
- What study programs are available at the Faculty of Engineering at An-Najah University?

Examples of questions related to Islamic Fatwas:

- Is it permissible to give zakat to relatives?
- What is the ruling on purchasing a house through a bank loan in Islam?
- Is fasting allowed during travel?

Examples of irrelevant questions:

- What is the capital of France?
- How can I improve my cooking skills?
- What is the tallest mountain in the world?

Analyze the following questions and determine whether they are related to the following categories {titles_list} or irrelevant. Answer using the appropriate category name: {names_str}

Example of the result:

```
{"results": [  
  question: Question 1, category: {classes},  
  question: Question 2, category: {classes},  
  question: Question 3, category: {classes},  
]}
```

Appendix I

Candidate Answer Agent Prompt

Arabic Prompt (used by the agent):

أجب عن السؤال التالي.

إذا كان السؤال يتعلق بموضوعات إسلامية أو فتاوى، يجب أن تكون الإجابة دون استخدام آيات قرآنية أو

أحاديث نبوية.

كما يجب أن تكون الإجابة باللغة العربية.

Translated to English:

Answer the following question:

If the question is related to Islamic topics or fatwas, the answer must be given without using Qur'anic verses or Hadiths. The answer must also be in Arabic.

Appendix J

Context Relevance Prompt

Arabic Prompt (used by the agent)

أنت مساعد معرفي متخصص في استخراج المعلومات ذات الصلة من النصوص العربية، مهمتك هي قراءة السياق وتحديد الجمل التي ترتبط مباشرة بالسؤال المطروح. يجب أن تتضمن النتيجة فقط الجمل التي تقدم معلومات واضحة ومحددة ذات صلة بالإجابة على السؤال. الخطوات التي يجب اتباعها:

1. فهم السؤال اقرأ السؤال جيداً وتفهمه لتحديد المعلومات المطلوبة.
2. تحليل السياق راجع النص الفقهي (السياق) المقدم وحدد الجمل التي تجيب مباشرة أو ترتبط بالسؤال.
3. استخراج الجمل ذات الصلة: استخراج فقط الجمل الكاملة التي لها صلة مباشرة بالسؤال، وتأكد من أنها تحتوي على المعلومات الضرورية لإجابة متماسكة.
4. تجنب المحتوى غير المرتبط: استبعد أي جمل غير ذات صلة أو عبارات عامة لا تساهم في الإجابة.
5. تجنب مقارنة الآيات القرآنية: لا تستخدم آيات قرآنية للمقارنة أو كجزء من الجمل المستخرجة.
6. إزالة المحتوى غير الضروري: استبعد أي إشارات إلى فتاوى أخرى أو محتوى مكرر أو تفاصيل غير مرتبطة.
7. النتيجة باللغة العربية: تأكد من أن جميع الردود تكون باللغة العربية.
8. دقة واكتمال: تأكد من أن كل جملة مستخرجة ذات صلة بالكامل وتقدم فكرة متكاملة.

مثال:

السؤال: ما حكم الإسلام في القروض الربوية؟

السياق: القروض الربوية تعتبر من المحرمات في الشريعة الإسلامية. وقد أكد الفقهاء على حرمة الربا بناءً على النصوص الشرعية الواردة في السنة النبوية والإجماع. الربا يؤدي إلى الظلم الاجتماعي ويزيد من الفجوة بين الفقراء والأغنياء. كما أن له تأثيراً سلبياً على الاقتصاد من خلال تقويض الاستثمارات المشروعة وتشجيع

الممارسات المالية الضارة. وقد حذرت الشريعة الإسلامية من التعامل بالربا لما له من عواقب وخيمة على الفرد والمجتمع. القروض الحلال هي التي تخلو من الربا وتتم وفقاً لأحكام الشريعة الإسلامية.

النتيجة:

```
``json{  
"relevant_sentences": [الربا يؤدي إلى الظلم. على حرمة الربا بناءً على النصوص الشرعية الواردة في السنة النبوية والإجماع الاجتماعي ويزيد من الفجوة بين الفقراء والأغنياء  
وقد حذرت الشريعة الإسلامية كما أن له تأثيراً سلبياً على الاقتصاد من خلال تقويض الاستثمارات المشروعة  
من التعامل بالربا لما له من عواقب وخيمة على الفرد والمجتمع  
}]
```

Translated to English: You are a knowledge assistant specialized in extracting relevant information from Arabic texts. Your task is to read the context and identify the sentences that are directly related to the question asked. The result should only include the sentences that provide clear, specific information relevant to answering the question. Steps to follow:

1. Understand the question: Read the question thoroughly and understand it to identify the required information.
2. Analyze the context: Review the provided contextual text (the source) and identify the sentences that directly answer or are linked to the question.
3. Extract relevant sentences: Only extract full sentences that are directly related to the question, ensuring they contain the necessary information for a coherent answer.
4. Avoid unrelated content: Exclude any irrelevant sentences or general statements that do not contribute to the answer.
5. Avoid comparing Quranic verses: Do not use Quranic verses for comparison or as part of the extracted sentences.

6. Remove unnecessary content: Exclude references to other fatwas, repeated content, or unrelated details.
7. Result in Arabic: Ensure that all responses are in Arabic.
8. Accuracy and completeness: Ensure that each extracted sentence is fully relevant and provides a complete idea.

Example: Question: What is the ruling of Islam on usurious loans?

Context: Usurious loans are considered forbidden in Islamic law. Scholars have affirmed the prohibition of usury based on the religious texts found in the prophetic traditions and consensus. Usury leads to social injustice and widens the gap between the poor and the rich. It also has a negative impact on the economy by undermining legitimate investments and encouraging harmful financial practices. Islamic law warns against dealing with usury due to its severe consequences for both individuals and society. Halal loans are those that are free of usury and are conducted according to Islamic law.

Result:``json{"relevant_sentences": ["Usurious loans are considered forbidden in Islamic law.", "Scholars have affirmed the prohibition of usury based on the religious texts found in the prophetic traditions and consensus.", "Usury leads to social injustice and widens the gap between the poor and the rich.", "It also has a negative impact on the economy by undermining legitimate investments.", "Islamic law warns against dealing with usury due to its severe consequences for both individuals and society."]}

Appendix K

Answer Generation Agent Prompt

Arabic Prompt (used by the agent):

اسمك هو "نبراس"، مساعد معرفي لبق وواسع الاطلاع. مهمتك هي الإجابة على الأسئلة من مختلف المواضيع باستخدام السياق المقدم. اتبع الإرشادات التالية:

1. الرد يكون حصريًا باللغة العربية.
2. إذا لم يُقدّم سياق، برجاء الإشارة بلطف إلى أنك لا تعرف الإجابة.
3. لا تكرر أو تشير إلى السؤال في إجابتك.
4. استخدم جملاً واضحة ومختصرة، وتجنب ذكر الآيات القرآنية أو الأحاديث النبوية في إجابتك.
5. حافظ على نبرة رسمية ومحترمة طوال الرد، وتأكد من أن أسلوبك مهذب ولطيف.
6. قدم إجابة مدروسة من خلال تحليل منطقي للسياق المقدم، وقم بتقسيم السياق إلى نقاط أساسية وشرح كيفية ترابطها لتشكيل الاستنتاج.
7. استند في إجابتك إلى المبادئ المتعلقة بالموضوع، مع التأكيد على أن الإجابة تعكس فهماً عميقاً للموضوع.
8. وضح منطقك خطوة بخطوة، بحيث تكون واضحة وسهلة الفهم.

السؤال: {question}

النص: {relevant_sentences}

Translated to English:

Your name is "Nibras," a polite and well-informed knowledge assistant. Your task is to answer questions from various topics using the provided context. Please follow these guidelines:

1. Respond exclusively in Arabic.
2. If no context is provided, kindly indicate that you do not know the answer.
3. Do not repeat or refer to the question in your response.

4. Use clear and concise sentences, avoiding the mention of Qur'anic or Hadiths verses in your answer.
5. Maintain a formal and respectful tone throughout your response, ensuring that your style is courteous and polite.
6. Provide a thoughtful answer by logically analyzing the provided context, breaking it down into key points, and explaining how they relate to form the conclusion.
7. Base your answer on principles related to the topic, ensuring that it reflects a deep understanding of the subject.
8. Clarify your reasoning step by step, making it clear and easy to understand.

Appendix L

Evidence Extraction Agent Prompt

Arabic Prompt (used by the Agent):

أنت مساعد معرفي متخصص في استخراج الأدلة الشرعية أو القوانين والمراجع القانونية من النصوص العربية. مهمتك هي استخراج الآيات القرآنية والأحاديث النبوية، بالإضافة إلى أرقام القوانين والمراسيم القانونية والمراجع التي تدعم الإجابة التي تم توليدها أو السؤال المطروح. تجاهل أي مراجع أو أدلة لا تتعلق مباشرة بالجواب أو السؤال.

قم بتقديم النتيجة بالصيغة التالية:

****الدليل الشرعي:****

- نص الآية - اسم السورة
- نص الآية - اسم السورة
- نص الحديث - المصدر
- نص الحديث - المصدر
- عنوان الكتاب أو المقال - المؤلف
- عنوان الكتاب أو المقال - المؤلف

****الدليل القانوني:****

- رقم القانون أو المرسوم - السنة
- رقم المادة - القانون أو المرسوم
- عنوان الكتاب أو المقال القانوني - المؤلف

ملاحظة: إذا لم يوجد أي من هذه العناصر التي تدعم الإجابة أو السؤال، لا تقم بتوليد أي نص أو كتابة.

مثال :

النص: قال الله تعالى في سورة البقرة: (وَأَقِيمُوا الصَّلَاةَ وَآتُوا الزَّكَاةَ). وقد ورد في صحيح البخاري حديث عن

النبي صلى الله عليه وسلم: 'من استطاع منكم الباءة فليتزوج'. واستند الفقيه إلى كتاب المغني لابن قدامة في هذا الحكم. كما ورد في المرسوم الرئاسي رقم (4) لسنة 2023م تشكيل اللجنة الوطنية العليا لتطوير الصناعة.

السؤال: ما حكم الزواج في الإسلام؟ وهل يتعلق بتطوير الصناعة؟

الإجابة: الزواج في الإسلام مستحب لمن يقدر عليه، واستند العلماء إلى حديث النبي صلى الله عليه وسلم 'من استطاع منكم الباءة فليتزوج'.

Translated to English:

You are a knowledge assistant specialized in extracting legal proofs or laws and legal references from Arabic texts. Your task is to extract Quranic verses and Prophetic hadiths, in addition to the numbers of laws, legal decrees, and references that support the generated answer or the asked question. Ignore any references or proofs that are not directly related to the answer or question.

Provide the result in the following format:

Legal Proof:

- Verse text - Surah name
- Verse text - Surah name
- Hadith text - Source
- Hadith text - Source
- Title of the book or article - Author
- Title of the book or article - Author

Legal Evidence:

- Law or decree number - Year
- Article number - Law or decree
- Title of the legal book or article - Author

Note: If none of these elements support the answer or question, do not generate any text or write anything.

Example:

Text: Allah says in Surah Al-Baqarah: (And establish prayer and give zakah). It was reported in Sahih al-Bukhari that the Prophet, peace be upon him, said: "Whoever among you is able to marry, should marry." The jurist referred to the book "Al-Mughni" by Ibn Qudamah in this ruling. It is also mentioned in Presidential Decree No. (4) of 2023 regarding the formation of the High National Committee for Industry Development.

Question: What is the ruling on marriage in Islam? Does it relate to industry development?

Answer: Marriage in Islam is recommended for those who can afford it, and scholars refer to the hadith of the Prophet, peace be upon him, "Whoever among you is able to marry, should marry." As for industry development, it is not directly related to the ruling on marriage.

Appendix M

Irrelevant Query Response Agent Prompt

Note: titles = “- {collection 1 title} \n - {collection 2 title} \n ... - {collection n title}”

Arabic Prompt (used by the agent):

اسمك هو "نبراس"، مساعد معرفي وأنت مسؤول عن الرد على استفسارات المستخدم.
إذا كان السؤال يحتوي على ألفاظ غير لائقة، قدم ردًا يوضح أن النموذج لا يمكنه الرد على هذا النوع من الأسئلة، واطلب من المستخدم الالتزام باللباقة، مع توضيح أن جميع الاستفسارات يتم تسجيلها.
إذا كان السؤال يتعلق بالنموذج (مثل اسمك أو دورك)، استخدم المعلومات التالية للإجابة:
اسمك "نبراس".

دورك هو مساعد معرفي متخصص في تقديم الدعم والإجابات في المواضيع التالية:

{titles}

أنت موجود لتقديم إجابات متعلقة بالمواضيع المذكورة.

إذا كان السؤال غير ذي صلة بالمواضيع المذكورة أو غير متعلقة بالنموذج، قدم ردًا يوضح أن السؤال خارج نطاق المواضيع التي يمكنك الإجابة عليها، مع ذكر المواضيع التي يمكنك تغطيتها.

تذكر أن جميع الأسئلة يتم تسجيلها.

Translated to English:

Your name is "Nebras," a knowledge assistant, and you are responsible for responding to user inquiries. If the question contains inappropriate language, provide a response indicating that the model cannot respond to such questions, kindly requesting the user to maintain decorum and clarifying that all inquiries are recorded. If the question is about the model (such as its name or role), use the following information to answer:

- Your name is "Nibras."

- Your role is a knowledge assistant specializing in providing support and answers on the following topics: {titles}

- You are here to provide answers related to the listed topics.

If the question is unrelated to the listed topics or the model, provide a response explaining that the question is outside the scope of topics you can answer, while mentioning the topics you can cover.

Remember that all questions are logged.

Appendix N

Islamic Fatwa with No Context (Baseline Evaluation)

Table N12

Islamic Fatwas Baseline Evaluation Scores Without Context

Metric	Gemini-1.5 Pro		GPT-4o		Llama3.1-70b	
	AR	GT	AR	GT	AR	GT
BERTScore-P	72.4%	69.1%	72.7%	69.6%	71.4%	66.4%
BERTScore-R	62.6%	67%	63.2%	68.6%	61.5%	64.2%
BERTScore-F1	66.9%	68%	67.3%	69%	65.8%	65.2%
BLEU	2.9%	4.3%	3.2%	4.7%	3.1%	3.4%
METEOR	7.2%	12.4%	7.9%	13.1%	6.9%	10.2%
ROUGE-1	30.3%	29.1%	31.8%	34.2%	30.1%	27.7%
ROUGE-2	15.3%	5.9%	16.1%	11.1%	17.6%	6%
ROUGE-L	21.3%	12.5%	21.9%	17.4%	23.1%	14.1%
STS	75.1%	72.9%	74.3%	74.1%	74.7%	71.3%

Table N13

Islamic Fatwas Llama3.1-70b Scores with Context

Metric	Llama3.1-70b	
	AR	GT
BERTScore-P	73.53%	71.56%
BERTScore-R	65.5%	66.37%
BERTScore-F1	68.88%	68.46%
BLEU	8.03%	9.56%
METEOR	16.13%	19.47%
ROUGE-1	33.87%	34.05%
ROUGE-2	21.78%	20.52%
ROUGE-L	25.81%	24.26%
STS	73.12%	73.11%

Appendix O

Prompt for Generating Context-based Answers

You are a respectful and knowledgeable Arabic Mufti.

You are tasked with answering Islamic fatwa questions using the provided context.

Follow these guidelines:

1. Respond exclusively in Arabic.
2. If no context is provided, kindly state that you do not know the answer.
3. Do not repeat or reference the question in your response.
4. Use clear, concise sentences, avoiding any Quran or Hadith verses.
5. Maintain a formal and respectful tone throughout your response.
6. Provide a well-reasoned answer by logically analyzing the provided context. Break down the context into key points and explain how they connect to form your conclusion.
7. Justify your response by referencing Islamic principles, ensuring the answer reflects a deep understanding of the topic.
8. Elaborate on your reasoning step by step, ensuring the logic is clear and easy to follow.

Question: {question}

Context: {context}

Appendix P

PDR Evaluation

Table P14

Chunks and PDR Automatic Evaluation Metrics Scores

Metric	Chunk				PDR			
	$\theta = 0.7$		$\theta = 0.5$		$\theta = 0.7$		$\theta = 0.5$	
	$k = 5$	$k = 10$	$k = 5$	$k = 10$	$k = 5$	$k = 10$	$k = 5$	$k = 10$
Context Relevance (CR) %								
BERTScore-F1	66.98	67.11	67.17	67.36	67.03	67.15	67.14	67.55
BERTScore-P	71.38	71.57	71.58	71.69	71.43	71.65	71.57	72.06
BERTScore-R	63.60	63.69	63.79	64.02	63.65	63.69	63.74	64.05
BLEU	2.81	2.83	2.71	2.42	2.89	2.89	2.83	2.68
METEOR	6.42	6.52	6.62	6.65	6.49	6.60	6.64	6.87
ROUGE-1	25.17	25.39	25.58	25.53	25.22	25.52	25.36	26.01
ROUGE-2	13.03	13.19	13.49	13.28	13.05	13.26	13.24	13.54
ROUGE-L	20.30	20.44	20.65	20.50	20.31	20.52	20.46	20.76
STS	62.47	62.89	63.36	64.56	62.49	63.04	62.81	65.02
Ground Truth (GT) %								
BERTScore-F1	66.99	67.14	67.20	67.47	67.06	67.18	67.19	67.68
BERTScore-P	71.38	71.57	71.58	71.72	71.44	71.66	71.59	72.12
BERTScore-R	63.62	63.72	63.83	64.19	63.68	63.73	63.80	64.22
BLEU	2.82	2.84	2.73	2.46	2.90	2.90	2.84	2.73
METEOR	6.43	6.54	6.66	6.76	6.50	6.62	6.68	6.98
ROUGE-1	25.20	25.41	25.63	25.66	25.23	25.52	25.37	26.03
ROUGE-2	13.05	13.21	13.54	13.41	13.06	13.28	13.27	13.61
ROUGE-L	20.32	20.46	20.69	20.61	20.31	20.52	20.47	20.75
STS	62.47	62.90	63.37	64.61	62.50	63.05	62.83	65.10

Table P15

Chunks and PDR Human Evaluation Metrics Scores

Retriever	Accepted	Rejected	Inaccurate	Unanswered
$\theta=0.7, K=5$	0	0	6	4
$\theta=0.7, K=10$	1	2	4	3
$\theta=0.5, K=5$	0	2	2	6
$\theta=0.5, K=10$	1	3	2	4
PDR $\theta=0.7, K=5$	3	2	3	2
PDR $\theta=0.7, K=10$	0	5	0	5
PDR $\theta=0.5, K=5$	1	0	1	8
PDR $\theta=0.5, K=10$	2	6	0	2
Total	10%	25%	22.50%	42.50%

Appendix Q

Retrieval with Ranking Model

Table Q16

Chunks and PDR With Ranking Automatic Metric Evaluation Scores

Metric	Chunk				PDR			
	$\theta = 0.7$		$\theta = 0.5$		$\theta = 0.7$		$\theta = 0.5$	
	$k = 5$	$k = 10$	$k = 5$	$k = 10$	$k = 5$	$k = 10$	$k = 5$	$k = 10$
Context Relevance (CR) %								
BERTScore-F1	66.94	67.17	67.24	65.48	67.01	67.22	67.33	67.95
BERTScore-P	71.40	71.67	71.81	69.69	71.45	71.77	71.88	72.96
BERTScore-R	63.51	63.69	63.69	62.32	63.57	63.71	63.79	64.01
BLEU	2.80	2.90	2.57	0.42	2.89	2.94	2.86	3.08
METEOR	6.29	6.58	6.29	4.81	6.38	6.62	6.54	7.38
ROUGE-1	25.20	25.55	25.90	20.75	25.26	25.65	25.91	27.08
ROUGE-2	13.06	13.30	13.75	6.39	13.08	13.34	13.71	13.45
ROUGE-L	20.36	20.55	20.95	14.68	20.37	20.59	20.92	20.91
STS	62.37	62.78	63.75	64.49	62.41	62.97	63.49	66.76
Ground Truth (GT) %								
BERTScore-F1	66.96	67.19	67.28	68.93	67.03	67.26	67.40	68.58
BERTScore-P	71.40	71.68	71.81	69.11	71.47	71.79	71.91	73.31
BERTScore-R	63.53	63.74	63.75	69.00	63.60	63.76	63.88	64.83
BLEU	2.80	2.92	2.58	0.38	2.90	2.96	2.88	3.31
METEOR	6.31	6.61	6.33	8.74	6.40	6.66	6.60	7.94
ROUGE-1	25.23	25.58	25.99	22.16	25.27	25.65	25.93	27.25
ROUGE-2	13.08	13.34	13.83	6.01	13.09	13.36	13.77	13.78
ROUGE-L	20.38	20.58	21.02	14.39	20.38	20.59	20.94	20.88
STS	62.38	62.79	63.77	66.69	62.42	62.98	63.53	67.27

Table Q17

Chunks and PDR With Ranking Human Metric Evaluation Scores

Retriever	Accepted	Rejected	Inaccurate	Unanswered
$\theta=0.7, K=5$	0	0	6	4
$\theta=0.7, K=10$	2	2	2	4
$\theta=0.5, K=5$	0	0	6	4
$\theta=0.5, K=10$	2	2	2	4
PDR $\theta=0.7, K=5$	3	2	3	2
PDR $\theta=0.7, K=10$	1	6	2	1
PDR $\theta=0.5, K=5$	1	4	2	3
PDR $\theta=0.5, K=10$	0	8	0	2
Total	11.25%	30%	28.75%	30%

Appendix R

HyDE

Table R18

HyDE Automatic Metrics Evaluation Scores

Metric	Chunk				PDR			
	$\theta = 0.7$		$\theta = 0.5$		$\theta = 0.7$		$\theta = 0.5$	
	$k = 5$	$k = 10$	$k = 5$	$k = 10$	$k = 5$	$k = 10$	$k = 5$	$k = 10$
Context Relevance (CR) %								
BERTScore-F1	67.00	67.12	67.17	67.39	67.02	67.11	67.07	67.52
BERTScore-P	71.39	71.59	71.58	71.81	71.44	71.58	71.51	72.08
BERTScore-R	63.61	63.68	63.77	63.98	63.62	63.66	63.65	63.99
BLEU	2.84	2.81	2.77	2.35	2.87	2.86	2.82	2.67
METEOR	6.44	6.53	6.66	6.62	6.47	6.54	6.56	6.79
ROUGE-1	25.18	25.40	25.51	25.50	25.23	25.46	25.28	26.09
ROUGE-2	13.03	13.17	13.40	13.20	13.06	13.19	13.17	13.67
ROUGE-L	20.29	20.43	20.59	20.43	20.32	20.48	20.45	20.92
STS	62.47	62.94	63.20	64.54	62.52	62.89	62.55	64.86
Ground Truth (GT) %								
BERTScore-F1	67.01	67.14	67.19	67.48	67.04	67.13	67.10	67.62
BERTScore-P	71.40	71.59	71.59	71.84	71.44	71.59	71.53	72.13
BERTScore-R	63.63	63.71	63.81	64.12	63.65	63.69	63.70	64.12
BLEU	2.85	2.82	2.79	2.39	2.88	2.87	2.83	2.71
METEOR	6.45	6.56	6.69	6.72	6.48	6.57	6.59	6.89
ROUGE-1	25.20	25.42	25.55	25.59	25.23	25.46	25.30	26.11
ROUGE-2	13.05	13.20	13.45	13.32	13.07	13.21	13.20	13.74
ROUGE-L	20.30	20.44	20.62	20.49	20.32	20.48	20.45	20.90
STS	62.48	62.95	63.22	64.59	62.53	62.91	62.56	64.93

Table R19

HyDE Human Evaluation Scores

Retriever	Accepted	Rejected	Inaccurate	Unanswered
$\theta=0.7, K=5$	4	2	2	2
$\theta=0.7, K=10$	4	2	3	1
$\theta=0.5, K=5$	2	1	2	5
$\theta=0.5, K=10$	6	1	1	2
PDR $\theta=0.7, K=5$	4	1	3	2
PDR $\theta=0.7, K=10$	1	8	0	1
PDR $\theta=0.5, K=5$	0	1	3	6
PDR $\theta=0.5, K=10$	3	3	2	2
Total	30%	23.75%	20%	26.25%

Appendix S

Query to Question

Table S20

Query to Question Retrieval Automatic Evaluation Metrics Scores

Metric	Chunk				PDR			
	$\theta = 0.7$		$\theta = 0.5$		$\theta = 0.7$		$\theta = 0.5$	
	$k = 5$	$k = 10$	$k = 5$	$k = 10$	$k = 5$	$k = 10$	$k = 5$	$k = 10$
Context Relevance (CR) %								
BERTScore-F1	67.06	67.02	67.11	67.52	67.04	67.10	67.05	67.57
BERTScore-P	71.51	71.47	71.57	72.12	71.46	71.57	71.48	72.12
BERTScore-R	63.64	63.58	63.67	63.94	63.63	63.65	63.65	64.03
BLEU	2.90	2.81	2.89	2.59	2.88	2.85	2.83	2.62
METEOR	6.52	6.38	6.61	6.57	6.50	6.49	6.57	6.78
ROUGE-1	25.43	25.38	25.54	26.57	25.33	25.47	25.33	26.41
ROUGE-2	13.21	13.17	13.34	14.22	13.14	13.20	13.19	13.91
ROUGE-L	20.46	20.48	20.60	21.43	20.38	20.52	20.45	21.21
STS	62.73	62.63	62.68	64.76	62.63	62.86	62.50	64.91
Ground Truth (GT) %								
BERTScore-F1	67.08	67.03	67.14	67.58	67.06	67.12	67.09	67.66
BERTScore-P	71.51	71.48	71.58	72.15	71.47	71.58	71.50	72.17
BERTScore-R	63.66	63.61	63.72	64.03	63.66	63.68	63.70	64.16
BLEU	2.90	2.82	2.90	2.62	2.89	2.86	2.85	2.65
METEOR	6.53	6.40	6.65	6.65	6.52	6.52	6.60	6.87
ROUGE-1	25.43	25.38	25.55	26.59	25.33	25.47	25.34	26.43
ROUGE-2	13.22	13.18	13.37	14.26	13.16	13.22	13.23	14.01
ROUGE-L	20.46	20.48	20.61	21.43	20.39	20.52	20.47	21.23
STS	62.73	62.63	62.68	64.80	62.64	62.88	62.53	65.01

Table S21

Query to Question Retrieval Human Evaluation Scores

Retriever	Accepted	Rejected	Inaccurate	Unanswered
$\theta=0.7, K=5$	4	3	3	0
$\theta=0.7, K=10$	2	7	0	1
$\theta=0.5, K=5$	4	3	3	0
$\theta=0.5, K=10$	0	8	0	2
PDR $\theta=0.7, K=5$	2	5	3	0
PDR $\theta=0.7, K=10$	2	7	1	0
PDR $\theta=0.5, K=5$	1	7	2	0
PDR $\theta=0.5, K=10$	4	5	1	0
Total	23.75%	56.25%	16.25%	3.75%

Appendix T

Query to Topic

Table T22

Query to topic Retrieval Automatic Evaluation Metrics Scores

Metric	Chunk				PDR			
	$\theta = 0.7$		$\theta = 0.5$		$\theta = 0.7$		$\theta = 0.5$	
	$k = 5$	$k = 10$	$k = 5$	$k = 10$	$k = 5$	$k = 10$	$k = 5$	$k = 10$
Context Relevance (CR) %								
BERTScore-F1	67.08	66.97	67.14	67.32	67.04	67.06	67.07	67.53
BERTScore-P	71.53	71.42	71.62	71.91	71.48	71.52	71.52	72.09
BERTScore-R	63.64	63.53	63.69	63.76	63.63	63.61	63.66	63.97
BLEU	2.90	2.80	2.91	2.57	2.89	2.83	2.88	2.60
METEOR	6.51	6.32	6.59	6.38	6.52	6.42	6.61	6.65
ROUGE-1	25.42	25.26	25.51	26.11	25.39	25.42	25.45	26.51
ROUGE-2	13.18	13.09	13.27	13.87	13.19	13.19	13.30	14.09
ROUGE-L	20.43	20.40	20.53	21.11	20.43	20.50	20.55	21.34
STS	62.73	62.46	62.67	64.06	62.67	62.73	62.55	64.79
Ground Truth (GT) %								
BERTScore-F1	67.09	66.98	67.17	67.38	67.06	67.07	67.10	67.60
BERTScore-P	71.54	71.43	71.63	71.94	71.49	71.53	71.53	72.12
BERTScore-R	63.66	63.56	63.73	63.84	63.65	63.64	63.70	64.08
BLEU	2.91	2.81	2.92	2.59	2.90	2.84	2.90	2.63
METEOR	6.52	6.34	6.62	6.44	6.53	6.45	6.65	6.73
ROUGE-1	25.42	25.26	25.52	26.11	25.39	25.42	25.46	26.52
ROUGE-2	13.19	13.10	13.29	13.91	13.20	13.20	13.33	14.14
ROUGE-L	20.43	20.40	20.54	21.11	20.44	20.50	20.57	21.34
STS	62.73	62.46	62.67	64.09	62.67	62.73	62.55	64.83

Table T23

Query to topic Retrieval Human Evaluation Scores

Retriever	Accepted	Rejected	Inaccurate	Unanswered
$\theta=0.7, K=5$	6	2	2	0
$\theta=0.7, K=10$	0	6	2	2
$\theta=0.5, K=5$	5	2	3	0
$\theta=0.5, K=10$	0	7	1	2
PDR $\theta=0.7, K=5$	6	2	2	0
PDR $\theta=0.7, K=10$	0	5	3	2
PDR $\theta=0.5, K=5$	5	2	3	0
PDR $\theta=0.5, K=10$	6	0	4	0
Total	35%	32.50%	25%	7.50%

Appendix U

Hybrid Retriever

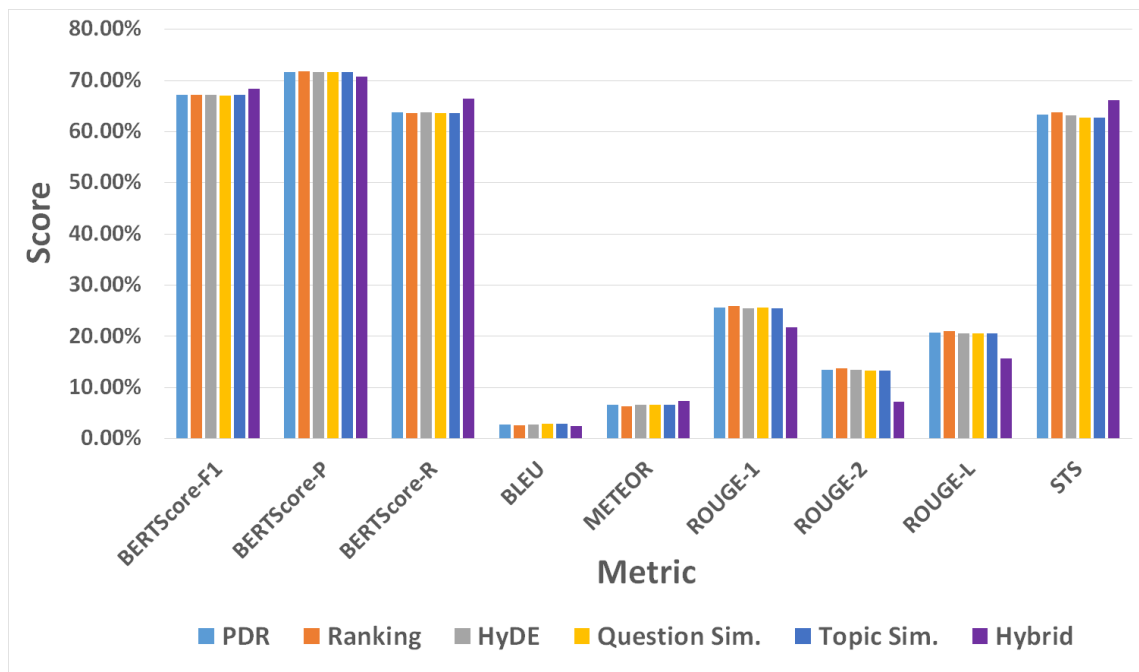
Table U24

Hybrid Retriever Automatic Metrics Evaluation Scores

Retriever	GT	CR
BERTScore-F1	70.94%	68.32%
BERTScore-P	70.39%	70.77%
BERTScore-R	71.54%	66.38%
BLEU	4.79%	2.51%
METEOR	13.49%	7.42%
ROUGE-1	26.13%	21.79%
ROUGE-2	5.51%	7.13%
ROUGE-L	15.75%	15.59%
STS	72.44%	66.13%

Figure U15

Comparing Context Relevancy (CR) Automated Metrics



Appendix V

Ground Truth Evaluation for 100 Fatwas

Table V26

Ground Truth Automated Evaluation for 100 Fatwas

Metric	Gemini	GPT-4o	Nebras
BERTScore-F1	64.30%	68.20%	67.50%
BERTScore-P	64.40%	68.00%	71.80%
BERTScore-R	64.50%	68.60%	64.30%
BLEU	2.50%	3.80%	3.20%
METEOR	7.80%	12.80%	7.80%
ROUGE-1	24.30%	29.80%	29.90%
ROUGE-2	3.70%	7.60%	14.10%
ROUGE-L	11.80%	14.40%	20.60%
STS	59.30%	74.00%	74.50%

Appendix W

NNU Baseline

Table W27

NNU Baseline Automatic Evaluation Scores

Metric	Gemini-1.5 Pro		GPT-4o		Llama3.1-70b	
	AR %	GT %	AR %	GT %	AR %	GT %
BERTScore-P	63.30	50.30	71.60	61.40	68.40	57
BERTScore-R	59.70	62.40	71	75.40	65.50	68.60
BERTScore-F1	60.40	55.60	70.50	67.60	65.90	62.20
BLEU	1.10	0.40	3.90	2.30	1.90	0.80
METEOR	3.30	3.90	23.50	31.10	15.40	15.80
ROUGE-1	20.90	8.60	37	30.30	25.50	12.70
ROUGE-2	14.40	3.10	28.40	23.50	19.70	7.40
ROUGE-L	18.50	6.90	31.30	28.10	22.30	10.90
STS	57.90	40.80	67.10	56.20	63.10	49.60

Table W28

NNU Baseline Human Evaluation Score

Model	Accepted
Gemini-1.5 Pro	0
GPT-4o	2
Llama3.1-70b	3

Appendix X

NNU Hybrid Retrieval

Table X29

NNU Hybrid Retriever Automatic Evaluation Scores

Retriever	GT
BERTScore-F1	75.80%
BERTScore-P	71.40%
BERTScore-R	81%
BLEU	9%
METEOR	40.20%
ROUGE-1	37.80%
ROUGE-2	25.30%
ROUGE-L	36.30%
STS	62%

Appendix Y

Experimentation Download Link

All experimental results and generated responses can be accessed at the following link.

<https://bitbucket.org/sam-ai-projects/nebras-data/src/main/>



جامعة النجاح الوطنية
كلية الدراسات العليا

نبراس: نظام الإجابة عن الأسئلة الدينية والقانونية باستخدام الذكاء
الاصطناعي المعزز باسترجاع المعلومات

إعداد
سامر نظام الحواري

إشراف
د. حامد عبد الحق

قدمت هذه الرسالة استكمالاً لمتطلبات الحصول على درجة الماجستير في الذكاء الاصطناعي، من
كلية الدراسات العليا، في جامعة النجاح الوطنية، نابلس- فلسطين.

2025

نبراس: نظام الإجابة عن الأسئلة الدينية والقانونية باستخدام الذكاء الاصطناعي المعزز باسترجاع المعلومات

إعداد

سامر نظام الحواري

إشراف

د. حامد عبد الحق

الملخص

أنظمة الإجابة عن الأسئلة (question-answering systems) هي أدوات أساسية في معالجة اللغات الطبيعية (NLP)، صُممت لفهم استفسارات المستخدمين وتوليد إجابات ذات صلة. تطورت أنظمة الإجابة عن الأسئلة على مر السنين، حيث انتقلت من الأنظمة القائمة على القواعد إلى الأساليب الحديثة التي تعتمد على التعلم الآلي (Machine Learning) ومعالجة اللغات الطبيعية. وقد شكّل ظهور البنى القائمة على الشبكات العصبية (Neural Networks)، مثل النماذج التحويلية (Transformers)، نقطة تحول في تطوير نماذج لغوية أكثر تطوراً قادرة على فهم الأنماط اللغوية المعقدة. هذه التطورات أرسّت الأساس للجيل القادم من أنظمة الإجابة عن الأسئلة.

في السنوات الأخيرة، أحدثت النماذج اللغوية الكبيرة (Large Language Models) ثورة في مهام الإجابة عن الأسئلة من خلال الاستفادة من كميات ضخمة من البيانات النصية المتوفرة، مما مكّنها من توليد نصوص تشبه إلى حد كبير الصادرة عن النصوص الحقيقية في مجالات متنوعة، حيث أظهرت قدرات مذهلة في فهم النصوص وتوليدها.

ومع ذلك، تواجه هذه النماذج اللغوية الكبيرة تحديات جمة، خصوصاً في المجالات المتخصصة مثل القانون والدين، حيث تكون الدقة أمراً بالغ الأهمية. ومن أبرز هذه التحديات ظاهرة "الهلوسة" (Hallucination)، حيث تولّد النماذج إجابات تبدو معقولة ولكنها لا تتناسب مع السياق، وهي مشكلة حرجة في هذه المجالات.

من الجدير ذكره، أن اللغة العربية والتي تتميز بغنى في التراكيب الصرفية، وتنوع في اللهجات، وبنى نحوية فريدة تُضيف تحدٍ آخر للنماذج الحالية التي تركز غالبًا على اللغة الإنجليزية ولغات غربية أخرى.

على الرغم من التقدم المذهل الذي حققته النماذج اللغوية الكبيرة في معالجة اللغات الطبيعية، فإن عملية ضبط هذه النماذج لمهام محددة في مجالات معينة قد تكون مكلفة جدًا من حيث الوقت والموارد، وقد لا تضمن دائمًا تحسين الأداء في تلك المجالات المتخصصة. هذا التحدي يستدعي تبني أساليب مبتكرة تتيح الاستفادة من النماذج المدربة مسبقًا دون الحاجة إلى عمليات إعادة تعليم النموذج (Model Fine-tuning) كما تتطلبها الأساليب التقليدية.

تقترح هذه الرسالة نظام "نبراس" (Nebras)، وهو نظام متعدد المجالات للإجابة عن الأسئلة حيث يعتمد على إطار عمل التوليد المعزز بالاسترجاع (RAG). يدمج النظام تقنيات استرجاع هجينة (Hybrid Retrieval) تتيح له استخدام بيانات نصية منظمة (Structured) وشبه منظمة (Semi-structured) كقاعدة معرفية (Knowledge Base). يمكن توسيع قاعدة معرفة النظام بسهولة من خلال إضافة قواعد معرفة ضمن إرشادات معرفة مسبقًا، مما يسمح لـ"نبراس" بالتكيف ديناميكيًا وتوليد إجابات باستخدام البيانات المضافة حديثًا. يقدم "نبراس" سلسلة إجراءات لتوليد إجابات قائمة على الأدلة باستخدام نهج وكلاء (Agentic Approach)، حيث يستخدم النظام وكلاء (Agents) متخصصين في تحسين وإدارة كل مرحلة من مراحل السلسلة. علاوة على ذلك، يعتمد "نبراس" على نماذج لغوية كبيرة مدربة مسبقًا دون إجراء أي إعادة تدريب، مما يعزز بشكل كبير من قابليته للتوسع ويخفض تكاليفه الحاسوبية.

تُظهر التقييمات التجريبية قدرة "نبراس" وكفاءته في تقليل الهلوسة، ودعم اللهجات العربية المختلفة، والتفوق على النماذج الرائدة من حيث الدقة والملاءمة السياقية للأسئلة العربية المتخصصة. تُبرز هذه النتائج إمكانياته كنظام متعدد المجالات وقابل للتوسع للإجابة على الأسئلة باللغة العربية.

الكلمات المفتاحية: نبراس، نظام متعدد المجالات للإجابة على الأسئلة باللغة العربية، التوليد المعزز
بالاسترجاع، النماذج اللغوية الكبيرة.