BACHELOR OF SCIENCE IN SOFTWARE ENGINEERING

ADAB: A Culturally-Aligned Automated Response Generation Framework for Islamic App Reviews by Integrating ABSA and Hybrid RAG

K.M.Tahlil Mahfuz Faruk
200042158
Mushfiqur Rahman Talha
200042124
H.M.Kawsar Ahamad
200042108

Department of Computer Science and Engineering
Islamic University of Technology
September, 2025

ADAB: A Culturally-Aligned Automated Response Generation Framework for Islamic App Reviews by Integrating ABSA and Hybrid RAG

K.M.Tahlil Mahfuz Faruk
200042158
Mushfiqur Rahman Talha
200042124
H.M.Kawsar Ahamad
200042108

Department of Computer Science and Engineering
Islamic University of Technology
September, 2025

Declaration of Candidate

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by **K.M.Tahlil Mahfuz Faruk**, **Mushfiqur Rahman Talha**, and **H.M.Kawsar Ahamad** under the supervision of **Syed Rifat Raiyan**, Lecturer, Department of Computer Science and Engineering and co-supervision of **Dr. Kamrul Hasan**, Professor, Department of Computer Science and Engineering and **Dr. Hasan Mahmud**, Professor, Department of Computer Science and Engineering, Islamic University of Technology, Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others have been acknowledged in the text and a list of references is given.

Syed Rifat Raiyan
Lecturer
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Date: September 30, 2025

Dr. Kamrul Hasan
Professor
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Date: September 30, 2025

Dr. Hasan Mahmud
Professor
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Date: September 30, 2025

K.M.Tahlil Mahfuz Faruk
Student ID: 200042158
Date: September 30, 2025

Mushfiqur Rahman Talha
Student ID: 200042124
Date: September 30, 2025

H.M.Kawsar Ahamad
Student ID: 200042108
Date: September 30, 2025

Contents

List of Figures

List of Tables

List of Abbreviations

ABSA	Aspect-Based Sentiment Analysis
AI	Artificial Intelligence
CUDA	Compute Unified Device Architecture
GPU	Graphics Processing Unit
LLM	Large Language Model
NLP	Natural Language Processing
RAG	Retrieval-Augmented Generation
ABSA	Aspect-Based Sentiment Analysis
ANN	Approximate Nearest Neighbor
BM25S	Best Match 25 with Sparse Scoring
FAISS	Facebook AI Similarity Search
HCAI	Human-Centered Artificial Intelligence
IR	Information Retrieval
LLM	Large Language Model
MPNet	Masked and Permuted Pre-trained Network
QA	Question Answering
RAG	Retrieval-Augmented Generation

Acknowledgement

We would like to express our deepest gratitude to our supervisor, Syed Rifat Raiyan, for his unwavering patience, insightful feedback, and continuous guidance throughout the course of this thesis. His expertise and encouragement have been invaluable to our academic and professional development.

We are also sincerely grateful to our co-supervisors, Dr. Kamrul Hasan, Dr. Hasan Mahmud, and Md. Mohsinul Kabir, for their valuable support, constructive suggestions, and consistent motivation, which greatly contributed to the completion of this work.

We would like to extend our special thanks to Dr. Riasat Islam, Trustee and Head of R&D at GreenTech Apps Foundation, for his supervision and for teaching us how to conduct research effectively. We are also thankful to Greentech Apps Foundation for providing the resources and technical support that greatly facilitated the evaluation of our research.

Finally, we extend our heartfelt thanks to our parents for their unconditional love, encouragement, and unwavering support throughout our academic journey. Their belief in us has been a constant source of strength and inspiration.

Abstract

Automated review response systems have advanced considerably, yet most fail to incorporate Islamic etiquette, values, and cultural norms, essential for meaningful engagement with Muslim users. Prior research has shown that timely and thoughtful engagement with user reviews can improve user perception. However, managing responses at scale remains a significant challenge for developers, particularly when cultural and religious considerations must be upheld. This research proposes ADAB, a framework for generating review responses that are culturally congruent with Islamic application contexts. The approach integrates a hybrid Retrieval-Augmented Generation (RAG) pipeline that employs agentic chunking and FAISS HNSW indexing to preserve contexts, with aspect-based sentiment analysis (ABSA) for fine-grained understanding of user feedback, and etiquette-aware prompt engineering to imbue responses with appropriate Islamic decorum. We also introduce a new open-source dataset of Islamic app reviews that supports the system's development and evaluation. Direct pairwise comparisons showed that ADAB's responses were preferred in 40% of cases, compared to 15.3% for the baseline, with 44.7% ties. On average, our approach achieves an overall improvement of 9.9%, with the largest gain in application specificity (+30.39%). Wilcoxon signed-rank test confirmed significant improvements in accuracy ($p = 0.0004$), relevancy ($p = 0.0417$), and specificity ($p = 8 \times 10^{-9}$), while grammatical correctness showed negligible change ($p = 0.453$). These results demonstrate that embedding cultural alignment in AI systems can foster trust and empathy, charting a path toward more respectful and human-centered response generation.

Chapter 1

Introduction

1.1 Background and Research Overview

The rapid advancement of artificial intelligence and natural language processing (NLP) has transformed how applications interact with users, with review response generation systems—especially those augmented by Large Language Models (LLMs) becoming vital for managing feedback on platforms like the Google Play Store. While widely used for general-purpose apps, Islamic applications remain underserved despite their growing importance in the digital lives of Muslim users.

This thesis introduces a specialized review response generation system for Islamic applications through a multi-stage pipeline that combines sentiment analysis, aspect extraction, retrieval, and generative modeling. A user review is analyzed for sentiment at both global and aspect levels, then enriched with supporting context retrieved from a curated knowledge base using a Retrieval-Augmented Generation (RAG) framework. These components guide an LLM to generate responses that are respectful, informative, and aligned with Islamic etiquette and language norms.

The work aligns with the Human-Centered Artificial Intelligence (HCAI) paradigm by emphasizing empathy, cultural sensitivity, and human values. Since reviews often mix technical issues with ethical expectations, the system is designed not to replace developers but to support them in crafting responses that foster trust, humility, and meaningful interaction—key principles of HCAI.

1.2 Motivation and Scope

The motivation for this research arises from the increasing role of Islamic mobile applications—such as Quran and prayer apps—in the daily lives of millions of users. Reviews on these platforms often go beyond technical issues to express cultural and ethical expectations, making them more sensitive than those of general-purpose applications [16], [18]. Yet, existing review response generation systems rarely address this context, resulting in replies that may be accurate but lack cultural appropriateness [25], [42]. This gap highlights the need for a system capable of generating responses that are both informative and respectful of Islamic etiquette.

Aligned with the principles of Human-Centered Artificial Intelligence (HCAI), this research seeks to design a RAG-based pipeline that integrates retrieval, aspect-based sentiment analysis, and prompt engineering to produce contextually grounded and culturally aligned responses. By combining dense and sparse retrieval techniques [10], [24], [29], reranking, and aspect-based sentiment analysis [23], [40], the system ensures relevance and empathy. Furthermore, prompt engineering techniques guide Large Language Models to reflect humility, gratitude, and respect in line with Islamic values [3], [7]. In doing so, this work not only advances technical methods in response generation [17], [36] but also demonstrates how AI can be adapted to serve culturally sensitive and human-centered applications.

1.3 Problem Statement and Research Objectives

1.3.1 Problem Statement

The sheer volume of user feedback makes manual handling nearly impossible. A recent dataset indicates that 39.48% of collected apps receive over 10,000 reviews, while 20.00% surpass 100,000 reviews. Yet, according to Hassan et al. [20], developers respond to only 2.8% of reviews. Most responses are limited to long or low-rated reviews and 97.5% of user-developer dialogues end after just one interaction. This highlights a pressing need for intelligent, automated systems that can scale to the demands of user engagement while maintaining accuracy and sensitivity.

This thesis aims to develop an automated reply system that not only addresses the user's query accurately but also adheres to Islamic cultural values, enhancing the interaction experience on public review platforms.

Research Objectives:

- Develop a Retrieval-Augmented Generation (RAG) pipeline that integrates dense, sparse, and hybrid retrieval methods for accurate context retrieval.
- Employ **Agentic Chunking** to create context-preserving document chunks that capture both app-specific knowledge and Islamic etiquette without information loss.
- Incorporate aspect-based sentiment analysis to ensure responses address specific user concerns with empathy and precision.
- Apply prompt engineering techniques to align generated responses with Islamic etiquette, emphasizing humility, gratitude, and cultural sensitivity.
- Evaluate the system's effectiveness compared to baseline response generation models in terms of relevance, coherence, and cultural appropriateness.

1.3.2 Research Challenges

The development of a culturally aligned review response generator for Islamic applications involves several nuanced challenges:

- **Context-preserving chunking:** Traditional approaches such as fixed-size or semantic chunking often lead to information loss, which reduces retrieval accuracy. Designing an agentic chunking mechanism that preserves contextual meaning while remaining efficient is a core challenge.
- **Cultural alignment:** Responses must embody humility, gratitude, and respect in accordance with Islamic etiquette. Embedding such values into a generative pipeline requires careful integration of cultural knowledge and prompt design, which goes beyond standard NLP practices.
- **Aspect-based sentiment analysis integration:** Identifying sentiments not only at the document level but also for specific application aspects (e.g., Quran recitation, ads, or interface design) introduces additional complexity while being crucial for nuanced and empathetic responses.
- **Retrieval accuracy:** Balancing sparse and dense retrieval methods to ensure precise and semantically relevant chunk selection remains a technical challenge, particularly when dealing with heterogeneous knowledge sources.
- **System evaluation:** As no established benchmarks exist for Islamic app review response generation, evaluation requires a hybrid approach, combining human

assessments of cultural appropriateness with automated methods such as LLM-based evaluators.

1.3.3 Contribution

This thesis makes several novel contributions such as:

- **HCAI-based etiquette-aligned framework:** We propose a Retrieval-Augmented Generation (RAG) pipeline specifically tailored for Islamic applications. Unlike existing systems, the framework incorporates cultural and ethical alignment by embedding Islamic etiquette into the response generation process. This ensures that outputs are not only factually accurate but also respectful, empathetic, and consistent with human-centered design principles.
- **Domain-specific dataset:** An open-source dataset of Islamic app reviews was curated and provided by the GreenTech App Foundation. This dataset is the first of its kind, enabling the training, evaluation, and benchmarking of response generation systems in a domain that has been underserved in prior research.
- **Evaluation methodology:** The system is evaluated using a hybrid approach that combines human evaluation with Large Language Models (LLMs) acting as judges. This dual evaluation strategy addresses the lack of existing benchmarks for Islamic app response generation, ensuring both reliability and scalability while highlighting cultural appropriateness and linguistic quality.

1.3.4 Organization

The remainder of this thesis is organized as follows. Chapter 1 introduces the research problem, highlights its significance, and outlines the motivation and scope of the study. Chapter 2 reviews related work on retrieval-augmented generation, sentiment analysis, and automated response generation, situating this research within the broader academic landscape. Chapter 3 presents the proposed methodology in detail, including the integration of agentic chunking, hybrid retrieval, aspect-based sentiment analysis, and etiquette-aligned prompt engineering within the Human-Centered AI framework. Chapter 4 discusses the experimental setup, datasets, evaluation strategies, and results, supported by both human evaluation and LLM-as-a-judge assessments. Finally, Chapter 5 concludes the thesis by summarizing key findings, reflecting on limitations, and suggesting directions for future research. Together, these chapters form a coherent narrative that progresses from motivation and background to methodology, results, and conclusions.

Chapter 2

Related Works

2.1 Literature Review

The research landscape on automated app review response generation and user feedback management has undergone a remarkable transformation over the past decade. What once began as simple, rule-based systems designed to filter or categorize feedback has now evolved into highly sophisticated pipelines powered by *deep neural networks*, *retrieval-augmented methods*, and most recently, *large language models (LLMs)*. This evolution can be broadly divided into four waves: (i) early sentiment analysis and preprocessing approaches, (ii) neural response generation, (iii) retrieval-augmented generation (RAG) frameworks with hybrid retrieval strategies, and (iv) self-improving, prompt-engineered LLM-based systems. Each wave reflects the broader progress in natural language processing (NLP), from hand-crafted heuristics to adaptive, knowledge grounded, and human-aligned systems.

Early sentiment analysis and preprocessing: The earliest approaches focused heavily on **sentiment classification** and **linguistic preprocessing**, which were necessary to make sense of noisy and diverse user feedback. Classical machine learning models such as Support Vector Machines (SVM), Naive Bayes (NB), and Random Forests (RF) were the workhorses of this era [39]. Their effectiveness, however, often depended more on the quality of preprocessing than on the model itself. For example, Alturayeif et al. [4] demonstrated that apps reviews could be effectively classified into aspects and sentiments using logistic regression and SVMs, provided that the text was carefully tokenized and normalized. Similarly, Mustofa and Idris [30] showed that ensemble models could outperform single classifiers on noisy Google Play data, especially after steps like stopword removal and normalization.

Another crucial thread was multilingual and low-resource environments. Al Asad et al. [2] explored GPT-based sentiment detection on social media, emphasizing the role of pretrained models in handling informal or dialect-rich text. These works collectively established that *data preprocessing and feature design were as important as the models themselves*. As computational resources improved, deep learning architectures began to replace hand-crafted features, enabling automatic extraction of semantic and syntactic patterns. Models like MPNet [35] and later transformer-based encoders allowed much richer contextual understanding, marking a turning point in sentiment and aspect-based analysis. Surveys such as those by Jayakody et al. [23] and Hua et al. [43] charted this shift, showing a steady trajectory from lexicon-based methods to CNNs, RNNs, and ultimately transformers. Toolkits like PyABSA [40] then lowered the barrier to entry, allowing researchers and practitioners to replicate and extend these advances with ease.

Neural response generation for app reviews: As app stores grew and developers were flooded with reviews, researchers moved from mere classification to the harder problem of generating responses automatically. This was the birth of response generation systems tailored to app reviews. Gao et al. introduced RRGen, one of the first attempts to treat review response generation as a machine translation task [16]. By conditioning on metadata like sentiment, review rating, and app category, RRGen was able to produce replies that were far more natural than rule-based templates. However, the model struggled with generalization and often produced generic outputs.

To overcome these issues, Zhang et al. developed TRRGen, which used a Transformer architecture instead of RNNs [42]. This shift was critical: the Transformer's self-attention mechanism enabled the model to capture long-range dependencies and adapt better to the emotional tone of reviews. With features like app category embeddings, TRRGen could generate replies that felt both more empathetic and context-specific. Still, both RRGen and TRRGen required large amounts of annotated training data, which is expensive to obtain, and tended to falter when confronted with reviews that contained multiple or ambiguous intents.

Parallel to these generative approaches, some researchers opted for safer, retrieval-based systems. Greenheld et al. [18] built a socially-informed, template-driven system that retrieved and adapted pre-existing developer responses, ensuring politeness and tone control. Unlike generative systems that risk hallucinating facts, these systems prioritized *consistency, safety, and brand voice*. Hassan et al. [20] also studied the actual dialogues between users and developers, revealing patterns such as the frequent use of apologies and thanks, which later informed system design. These early neural

and retrieval systems highlighted a recurring tension: the trade-off between flexibility and safety.

Retrieval-Augmented Generation (RAG) and hybrid retrieval: As the limitations of purely generative models became apparent—particularly their tendency to hallucinate or drift off-topic—researchers turned to RAG. This approach combines retrieval (to ground responses in external documents) with generation (to phrase the reply fluently). While early "naive RAG" methods simply appended retrieved passages to the prompt, they often suffered from irrelevant or redundant content [17]. Surveys by Gao et al. [17] and Zhu et al. [44] documented the emergence of **advanced RAG**, which introduced query rewriting, reranking, and post-retrieval filtering, and **modular RAG**, where components like retrievers, classifiers, and memory modules could be independently optimized.

The BEIR benchmark [36] played a pivotal role in this wave. By evaluating diverse systems across 18 datasets, BEIR showed that while dense retrievers like DPR [24] performed strongly in-domain, they often failed under domain shifts. Surprisingly, simple lexical baselines like BM25 remained highly competitive out-of-domain. Meanwhile, late-interaction models like ColBERT and re-rankers offered the best generalization but came at the cost of higher latency. Supporting tools such as FAISS [14], MPNet encoders [35], and training datasets like MS MARCO [31] became the backbone of retrieval pipelines.

More recent work has focused on **improving the retrieval layer itself**. BM25S[29] drastically reduced the latency of lexical retrieval, while Zhang et al. [41] proposed graph-based hybrid retrieval that aligns dense and sparse representations for efficiency. Hsu & Tzeng's DAT framework [21] introduced dynamic weighting between dense and sparse retrieval, adapting to each query. Similarly, Arabzadeh et al. [6] and Wang et al. [38] explored query-level prediction of the best retrieval strategy to balance effectiveness and cost. In applied contexts, Ganesh et al. [15] introduced *Context-Augmented Retrieval (CAR)*, a modular pipeline that classifies queries before hybrid retrieval, thus reducing noise and improving response speed. Lee et al. [28] extended these ideas to hybrid knowledge bases that combine text and relations, while query expansion methods like Query2Doc [37], AGR [11], and prompt-based expansion [22] demonstrated how LLMs can actively steer retrieval for better coverage.

LLMs, prompt engineering, and self-improving systems: The most recent wave is defined by the integration of **large language models**. Unlike earlier models that required heavy fine-tuning and static datasets, LLMs bring generalization, fluency,

and reasoning ability. Researchers now treat them not only as generators but also as *evaluators* and *search agents*. Surveys by Sahoo et al. [33] and Amatriain [5] consolidate an array of techniques: Chain-of-Thought reasoning, role-play prompting [27], reflection, self-consistency, and automatic prompt engineering. These methods show that even without fine-tuning, careful prompt design can significantly improve reasoning and alignment.

In applied response generation, Azov et al. [7] introduced *SCRABLE*, a framework that integrates GPT-4 into a retrieval loop with an "LLM-as-a-Judge." This means the system not only generates replies but also evaluates them, iteratively refining prompts to improve factuality, tone, and empathy. This feedback-aware architecture represents a paradigm shift: systems can now *learn from their own outputs* and self-improve in real time.

At the infrastructure level, open-weight LLMs such as GPT-OSS [1] are redefining the research landscape by providing powerful models under permissive licenses, though with associated safety risks. In contrast, Albuquerque et al. [3] demonstrated that smaller open-source LLMs, when fine-tuned, can achieve performance comparable to proprietary systems on customer feedback tasks, offering a cost-efficient alternative. Meanwhile, comparative studies show that cross-encoder rerankers [12], [32] remain highly competitive in retrieval tasks, outperforming even GPT-4 on some datasets, though at greater computational expense.

2.2 Critical Analysis and Synthesis

From static generators to modular, grounded pipelines. The early wave of neural generators, such as TRRGen, clearly demonstrated the benefit of **attention mechanisms** and the use of contextual embeddings (e.g., review ratings and app categories) for producing more empathetic and fluent replies [42]. However, despite these advances, TRRGen and its predecessors often fell into the trap of generating *generic, one-size-fits-all replies*, which risked alienating users by appearing insincere or irrelevant. The challenge was not fluency—these models could write well—but grounding: they lacked access to dynamic, up-to-date knowledge and were unable to adapt to new contexts in real time. Pre-trained transformers (e.g., BERT, RoBERTa) provided a leap in human-judged quality, especially in low-resource settings, but this improvement came at the cost of significantly higher inference latency and compute requirements, making them difficult to deploy in production-scale environments [9].

In contrast, template- and retrieval-based systems [18] ensured outputs were always

safe, polite, and on-brand, since they drew from pre-validated developer responses. Yet, these approaches traded away personalization and adaptability: while they excelled in tone consistency, they struggled with addressing novel complaints or complex, multi-intent reviews. This tension between **creativity vs. control** remains a recurring theme in the field.

Retrieval-Augmented Generation (RAG) emerged as a middle ground. By combining retrieval with generation, RAG systems reduced hallucinations and grounded responses in external documents. Surveys such as Gao et al. [17] and Zhu et al. [44] trace how the field evolved from naive "retrieve-and-append" pipelines, which often suffered from irrelevant passages, to **advanced RAG** systems that add query rewriting and reranking, and finally to **modular RAG**, where different components (retrievers, classifiers, rerankers, memory modules) can be tuned or swapped independently. Frameworks like CAR [15] exemplify this modular approach, integrating classification-based routing and hybrid retrieval to improve both efficiency and domain alignment. Meanwhile, community benchmarks and libraries such as BEIR [36], FAISS [14], DPR [24], MPNet [35], and MS MARCO [31] have become essential infrastructure providing the shared evaluation settings needed to measure progress and stress-test generalization.

Improving retrieval as the lever for better RAG. A consistent theme across the literature is that **retrieval quality is the bottleneck for RAG**. No matter how advanced the generator, if the retrieved context is irrelevant, the output is likely to be misleading or hallucinated. Recent innovations therefore focus heavily on strengthening the retrieval layer. BM25S[29] introduced a high-speed lexical search method that achieves orders-of-magnitude improvements in efficiency without sacrificing exactness, making it more practical for large-scale, real-time applications. Hybrid dense sparse methods [41] take this further by aligning representations across modalities, while still benefiting from the precision of sparse methods like BM25.

Dynamic strategies such as DAT [21] push this adaptability further, allowing the balance between dense and sparse retrieval to shift on a per-query basis, guided by LLM judgments. Similarly, Arabzadeh et al. [6] and Wang et al. [38] argue that no single retrieval strategy works best for all queries, and propose query-level meta-predictors that dynamically select the most cost-effective approach.

Alongside these retrieval innovations, LLM-based query expansion techniques such as AGR [11] and Query2Doc [37] have proven particularly effective. Instead of relying solely on the user's original wording, these systems rewrite or expand the query into richer, semantically faithful alternatives, ensuring that the retrieval step captures

more relevant passages. Together, these innovations make it clear that improving retrieval—through speed, adaptivity, or expansion—is the most powerful lever for improving the overall effectiveness of RAG.

Self-improvement and evaluation. Another critical theme in the literature is how systems are **evaluated and refined over time**. Traditional metrics such as BLEU, ROUGE, and METEOR, which measure surface-level word overlap, fall short in capturing the qualities that matter most to users—such as empathy, politeness, tone alignment, and factual helpfulness. This gap has motivated the development of new evaluation paradigms.

SCRABLE [7] exemplifies this shift by introducing an "LLM-as-a-Judge" component. Here, GPT-4 not only generates responses but also evaluates them, scoring each reply on dimensions like relevance, specificity, and tone. These scores are then used to iteratively refine prompts, creating a **feedback loop that continuously improves system performance**. This approach points beyond static metrics to a world where evaluation itself is automated, dynamic, and aligned with human preferences.

This self-improving paradigm is echoed in the rise of prompt engineering. Surveys [5], [33] and methods like role-play prompting [27] show how carefully designed instructions can trigger models to reason more systematically, explain their decisions, or adopt personas that improve tone alignment. Together, these trends suggest that future systems will be not only grounded in external knowledge but also capable of **self-evaluation and adaptive improvement**.

ABSA's role in response quality. While much of the focus has been on generation and retrieval, **Aspect-Based Sentiment Analysis (ABSA)** plays a vital supporting role. Effective response generation is not only about factual grounding—it must also be emotionally intelligent, aligning tone and style with the sentiment of the user's review. Classical ABSA pipelines relied heavily on hand-crafted features and lexicons, but these approaches were brittle and domain-limited [39]. Deep learning models such as CNNs and LSTMs improved matters by learning features directly from data, but often struggled with complex, multi-aspect sentences.

Transformer-based models such as BERT and MPNet have since become the state of the art, enabling fine-grained sentiment extraction even in challenging contexts [23], [43]. Toolkits like PyABSA [40] have made it easier to reproduce and extend such experiments, while newer approaches (e.g., dependency-aware modeling) improve the capture of inter-aspect sentiment dependencies [13].

Despite these advances, major gaps remain. ABSA models often fail to generalize

beyond benchmark domains like restaurants or laptops, and their performance drops sharply in multilingual or low-resource contexts. This is particularly concerning for app reviews, which are highly diverse and often written in informal, code-switched language. Addressing these gaps—by integrating ABSA more tightly into RAG+LLM pipelines and extending coverage to multilingual domains—represents a promising direction for improving the overall quality and empathy of automated responses.

2.3 Comparative Evaluation

Neural/Transformer vs. LLM. The shift from neural models such as TRRGen to LLM-powered systems illustrates a broader transition from *static fluency* to *dynamic adaptability*. TRRGen [42] significantly advanced the field by leveraging Transformers to capture long-range dependencies, producing fluent and context-sensitive responses that outperformed earlier RNN-based systems. However, its reliance on static training data meant that it often produced generic or repetitive outputs when encountering ambiguous or novel reviews. In contrast, LLM+RAG systems such as SCRABLE [7] move beyond static training, integrating retrieval for factual grounding and employing iterative evaluation loops (LLM-as-a-Judge) to refine tone and specificity. This makes them far more adaptable in real-world scenarios where review topics and user concerns shift rapidly. The comparison reveals that while Transformer models excel at style and coherence, LLM-based pipelines add a crucial layer of **adaptivity and self-improvement**.

Template/Retrieval vs. Generative. Template and retrieval-based systems, exemplified by Greenheld et al. [18], emphasize **control, predictability, and safety**. Because responses are drawn from a curated pool and adjusted only slightly for context, the risk of producing off-brand, impolite, or factually incorrect replies is minimized. This makes such systems attractive in high-stakes domains (e.g., healthcare, finance, hospitality) where tone consistency and compliance are paramount. However, their lack of flexibility often results in stale or overly generic interactions, which may frustrate users seeking personalized engagement. Generative models, by contrast, excel at **personalization and diversity**. Pre-trained models such as those studied by Cao and Fard [9] can craft empathetic, semantically rich replies even in low-data contexts. Yet, this strength comes with the risk of hallucination and tone misalignment, especially when grounding is weak. The contrast highlights a fundamental trade-off: *safety vs. creativity*. Modern systems increasingly attempt to bridge this divide by combining retrieval-driven control with generative adaptability.

Retrieval stack. The retrieval layer forms the backbone of any RAG pipeline, and its design directly impacts system performance. Lexical retrieval methods such as BM25, evaluated extensively in the BEIR benchmark [36], remain surprisingly robust in out-of-domain (OOD) contexts, often outperforming dense retrievers under distribution shifts. Dense retrieval methods, such as DPR [24] and MPNet [35], excel at capturing semantic similarity in-domain and scale effectively with large corpora. Infrastructure tools like FAISS [14] make it possible to deploy these dense models at trillion-scale, enabling real-time applications. However, dense retrieval alone can be brittle. Recent innovations address this gap: BM25S [29] accelerates lexical retrieval dramatically[29], graph-based hybrid ANN approaches [41] combine sparse and dense vectors for better efficiency–effectiveness balance, and DAT introduces query-level dynamic weighting to adapt dense vs. sparse contributions [21]. At the system level, CAR [15] integrates classification-aware routing, ensuring that only the most relevant partition of the corpus is searched. Collectively, these developments illustrate that retrieval is no longer a monolithic component but a **stack of techniques** tuned for robustness, scalability, and latency.

Query optimization. Even with a strong retrieval stack, the effectiveness of RAG pipelines is limited by the quality of the input query. Users often write reviews or queries that are vague, underspecified, or ambiguous. LLM-based query expansion methods address this by rewriting or elaborating the query before retrieval. Techniques like AGR (Analyze–Generate–Refine) [11] explicitly break down the query intent, generate multiple candidate expansions, and refine them for clarity. Query2Doc [37] takes another approach, prompting an LLM to generate pseudo-documents that encapsulate the query intent, thus enriching the retrieval pool. Jagerman et al. [22] further demonstrated how Chain-of-Thought prompting could produce step-by-step expansions that significantly boost recall. Together, these methods consistently improve both recall and end-to-end QA performance by ensuring that retrievers are guided by richer, semantically faithful inputs. The implication is clear: **better queries lead to better retrieval, which leads to better responses**.

2.4 Research Gaps and Opportunities

Despite the impressive progress in automated review response generation, several significant gaps remain that limit the deployment of these systems in real-world, multilingual, and high-stakes environments. The literature points to opportunities along the following dimensions:

- **From naive to modular RAG.** Most existing RAG implementations still follow a naive "retrieve-and-append" paradigm, where retrieved passages are directly concatenated into prompts. While simple, this often introduces irrelevant or redundant information, leading to hallucinated or off-topic responses. Future systems should move toward *modular RAG* architectures [15], [17], [36] that incorporate reranking, query rewriting, intent-based routing, memory modules, and hybrid retrieval strategies. Such modularity would enable pipelines to adapt dynamically, scaling across domains and use cases while maintaining high precision.
- **Domain generalization.** Neural models such as RRGen and TRRGen have shown promise, but their reliance on handcrafted metadata and domain-specific training data means they often fail to generalize beyond the apps or categories they were trained on. Real-world deployment requires models that are robust to domain shifts. Leveraging benchmarks like BEIR [36], which stress-test retrieval systems across diverse datasets, offers a path forward. Building adaptive, zero-shot LLM retrievers capable of transferring knowledge across domains remains a major research opportunity.
- **Grounding and specificity.** A recurring issue across neural generators is the tendency to produce generic replies such as "Thank you for your feedback." While safe, these lack informativeness and fail to build user trust. Research has shown that hybrid retrieval (combining dense and sparse methods) and data filtering strategies can improve specificity by providing targeted, contextually rich information [15], [25]. Future work should focus on fine-grained retrieval selection and context prioritization to ensure responses are both factually grounded and personalized.
- **Sentiment awareness.** Most current systems treat sentiment as a static feature (e.g., a star rating) rather than a nuanced, dynamic signal embedded within language. Aspect-Based Sentiment Analysis (ABSA) research highlights the need to model inter-aspect dependencies and emotional tone [13], [23], [40], [43]. Integrating fine-grained sentiment signals directly into the retrieval and generation pipeline could enable replies that are not only factually correct but also empathetic and intent-aligned.
- **Multilingual & low-resource contexts.** Many app stores are global, with reviews written in multiple languages, dialects, or even code-switched text. However, most current systems are trained on standardized English datasets and underperform in multilingual or low-resource settings [39], [43]. Extending

RAG+LLM pipelines with multilingual encoders and culturally diverse vector stores will be essential for inclusivity, especially in regions where linguistic diversity is the norm.

- **Efficiency and cost.** While LLM-based systems like SCRABLE deliver impressive results, they are often prohibitively expensive to deploy at scale. Inference latency, memory requirements, and computational cost remain critical barriers. Recent innovations such as BM25S for faster lexical retrieval [29], hybrid graph-based methods [41], fine-tuning smaller open-source LLMs [3], and open-weight reasoning models like GPT-OSS [1] point to possible solutions. Research on model distillation, quantization, and hardware-friendly ANN search could further close the gap between quality and efficiency.
- **Human-aligned evaluation.** Most studies continue to rely on traditional metrics such as BLEU, ROUGE, or METEOR, which measure lexical overlap but fail to capture qualities that matter most to users—such as helpfulness, empathy, or factual correctness. New approaches, such as SCRABLE's use of an LLM-as-a-Judge [7] or broader prompt-engineering strategies for reasoning [33], demonstrate that automated evaluation can be aligned more closely with human preferences. Establishing standardized, human-aligned evaluation protocols remains a critical opportunity for future research.

2.5 Cohesion

When viewed as a whole, the literature tells a coherent story of steady evolution in response generation systems. Each wave of research builds naturally upon the limitations of the previous one, forming a logical arc of progress:

- Early work transitioned from **sequential RNN-based generators** to **attention-based Transformers** such as TRRGen, reflecting the broader shift in NLP architectures toward models capable of capturing long-range dependencies [42].
- Retrieval pipelines evolved from **naive RAG**, which simply appended documents, to **modular and hybrid retrieval frameworks** like CAR, incorporating classification, reranking, and hybrid search for greater scalability and contextual alignment [15], [24], [36].
- Prompting strategies shifted from **static instructions** to **self-improving LLM agents**, which can role-play, reflect, and even evaluate their own outputs, as in SCRABLE's LLM-as-a-Judge paradigm [5], [7], [27], [33].

- In parallel, ABSA research moved from **classical ML approaches** (SVMs, lexicons) to **deep and transformer-based models**, supported by toolkits like PyABSA, enabling reproducible experiments and richer modeling of inter-aspect dependencies [13], [23], [39], [40].

Together, these transitions create a cohesive research narrative: one where response generation systems have steadily grown in sophistication, moving from rigid and generic outputs toward pipelines that are *adaptive, sentiment-aware, multilingual, and self-improving*. This continuity underscores the field's ultimate trajectory—toward building systems that are not only technically advanced but also deeply aligned with human expectations for empathy, factuality, and trust.

2.6 Summary

The reviewed literature demonstrates that modern systems for app review response generation are no longer defined by a single dominant paradigm but by the **integration of multiple complementary components**. The trajectory of progress shows a steady move from static neural generators, through retrieval-augmented pipelines, toward hybrid architectures that seamlessly blend *retrieval*, *generation*, *sentiment modeling*, and *evaluation*.

At the retrieval layer, innovations such as BM25S for efficient lexical search, hybrid graph-based ANN methods, and dense retrievers like DPR and MPNet—supported by scalable infrastructures such as FAISS—have become the backbone of evidence-driven pipelines [14], [24], [29], [35], [41]. Benchmarks such as BEIR [36] provide the necessary stress-tests to ensure that these methods generalize across domains. In parallel, LLM-driven query optimization methods, including AGR and Query2Doc, have proven effective at refining underspecified queries and consistently improving recall and precision [11], [37].

On the generation side, the field has shifted from static neural architectures (e.g., TRRGen) to **LLM-based systems that incorporate feedback and self-improvement**. Frameworks like SCRABLE [7] exemplify this change by embedding an "LLM-as-a-Judge" loop that evaluates and iteratively refines responses. This trend is reinforced by advances in prompt engineering, which provide models with structured reasoning strategies, role-play capabilities, and self-consistency checks to better align outputs with human expectations [5], [33].

Perhaps most importantly, the literature highlights a growing recognition that effec-

tive responses must not only be factually correct but also **empathetic, sentiment-aware, and culturally adaptive**. Integrating fine-grained ABSA signals into RAG pipelines has emerged as a promising direction for aligning system tone and intent with user emotions. Meanwhile, the rise of cost-efficient, fine-tuned open-source LLMs offers a pathway to scalable deployment without compromising quality.

Building on these insights, this thesis proposes a *sentiment-aware, RAG-enhanced LLM architecture* that combines retrieval grounding, emotional intelligence, multilingual inclusivity, and adaptive evaluation. By uniting these strands, the proposed system aims to advance the field toward the next generation of feedback platforms: ones that are technically robust, contextually intelligent, and aligned with human expectations of empathy, factuality, and trustworthiness.

Chapter 3

Proposed Methodology

3.1 Overview of the Methodology

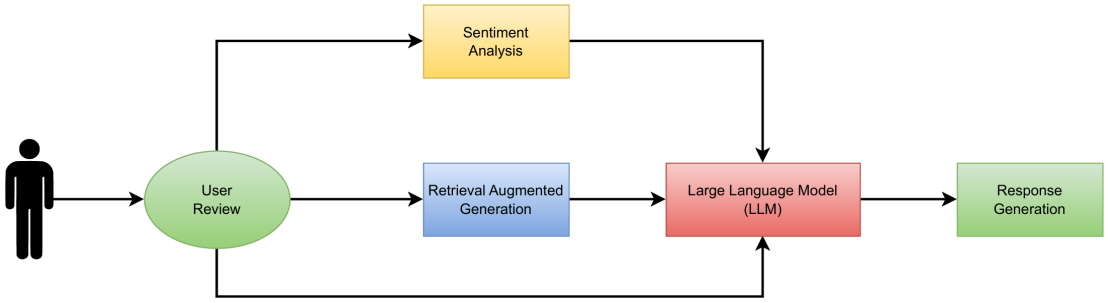

Figure 3.1: System architecture of the proposed methodology, showing integration of retrieval, reranking, ABSA, and LLM response generation.

The proposed methodology integrates multiple advanced techniques in Retrieval Augmented Generation (RAG) to improve both the retrieval process and the quality of generated responses for Islamic app reviews. At a high level, the system combines retrieval strategies, sentiment analysis, and prompt engineering to form a cohesive pipeline that ensures responses are contextually accurate, semantically rich, and aligned with Islamic etiquette.

The methodology begins with the ingestion and preprocessing of app reviews, followed by a multi-layered retrieval strategy. Reviews are expanded using query expansion techniques [11], [22], [37], which improve coverage and ensure that both lexical and semantic variations of user input are captured. To address different retrieval needs, we combine sparse retrieval using BM25S [29], dense retrieval with all-MPNet-base-v2 embeddings [35], and approximate nearest neighbor search via FAISS [14].

Hybrid retrieval is adopted to leverage the strengths of each retrieval paradigm, as supported by recent studies on dense–sparse complementarities [6], [38].

Once candidate documents are retrieved, a reranking stage using cross-encoders ensures higher precision by modeling fine-grained query–document interactions [12], [32]. Aspect-based sentiment analysis (ABSA) is then applied using the PyABSA framework [40], enabling the system to extract sentiments tied to specific aspects of the application (e.g., Quran recitation quality, ads, or interface design). This guarantees that generated responses are not only factually grounded but also empathetic and aspect-aware [4], [23].

The final stage employs a large-scale open-source language model (gpt-oss-120b) [1], enhanced through carefully designed prompts [5], [27], [33]. Prompt engineering incorporates Islamic etiquette, ensuring responses begin with respectful greetings, demonstrate humility, and reflect gratitude in line with adab principles. By balancing creativity and factual grounding through tuned hyperparameters, the model generates responses that are not only technically accurate but also culturally appropriate.

Figure 3.2 illustrates the complete system architecture. The figure demonstrates how different components, including retrieval, reranking, ABSA, and the LLM generator, interact to form a unified pipeline. This layered design directly addresses limitations in earlier review response generation methods [16], [25], [42], ensuring both improved retrieval relevance and high-quality, context-sensitive responses.

3.2 Chronological Breakdown of Components

This section presents the methodology in a chronological sequence, showing how each component contributes to the overall pipeline. The process begins with review ingestion and preprocessing, followed by query expansion, hybrid retrieval, reranking, sentiment analysis, and finally, large language model (LLM) based response generation. Each component has been deliberately chosen based on prior research and empirical evidence to improve retrieval robustness, factual grounding, and cultural alignment of responses.

3.2.1 Step 1: Data Ingestion and Preprocessing

The pipeline begins with the ingestion of app reviews collected from platforms such as the Google Play Store. Raw reviews often contain noise in the form of typos, emojis, code-mixed language, and redundant content. Preprocessing addresses these issues

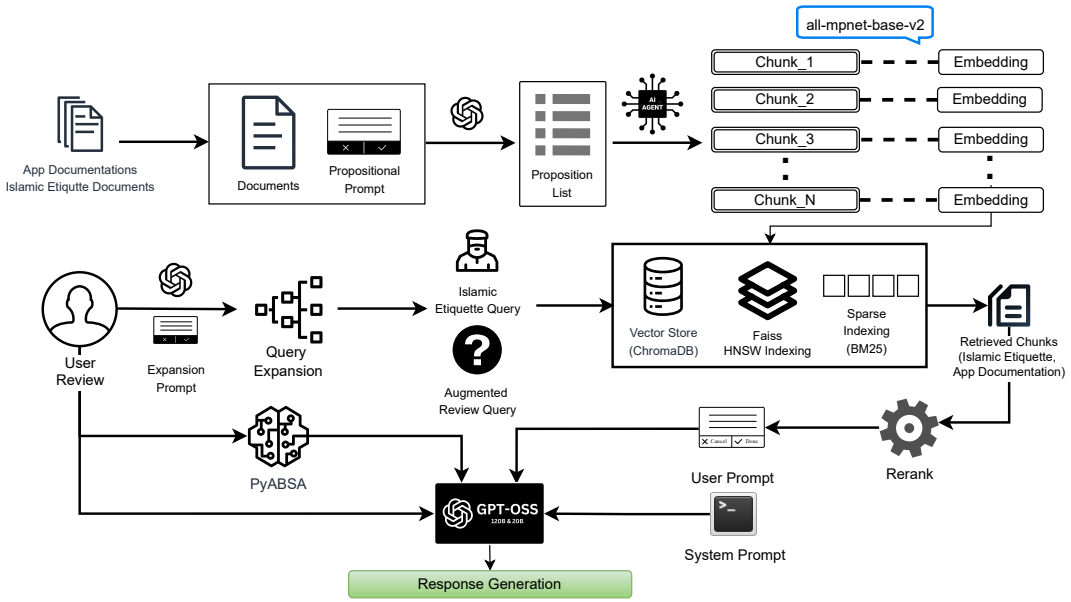

Figure 3.2: Integrated system architecture, showing the flow of information across preprocessing, query expansion, hybrid retrieval, reranking, ABSA, and LLM-based response generation.

by performing normalization, tokenization, and stop-word removal to create clean textual inputs suitable for retrieval. This step is crucial, as poor preprocessing can propagate errors throughout the pipeline, weakening retrieval accuracy and degrading response quality.

In addition to basic text cleaning, reviews are segmented into coherent chunks that preserve semantic meaning. Inspired by agentic chunking approaches in RAG pipelines [34], our system uses adaptive chunking strategies where text is split into semantically meaningful passages rather than fixed-length windows. This ensures that retrieved units are both contextually relevant and easily interpretable by downstream models.

The choice of adaptive chunking is motivated by evidence that retrieval granularity significantly impacts downstream performance in question answering and RAG systems. For example, Chen et al. [10] showed that finer-grained retrieval improves recall but may harm contextual coherence, whereas coarse-grained retrieval improves precision but risks missing relevant information. Our preprocessing balances these trade-offs by creating chunks that maximize semantic coverage without losing context.

By ensuring clean, semantically coherent input at the earliest stage, this component provides a solid foundation for subsequent retrieval and generation steps, directly improving both the relevance of retrieved documents and the quality of final responses.

3.2.2 Step 2: Query Expansion

After preprocessing, the next step is query expansion, which aims to reformulate user reviews into richer and more comprehensive queries. Since app reviews are often short, noisy, and ambiguous, directly using them for retrieval may lead to missing relevant documents. Query expansion mitigates this issue by enriching the original query with semantically related terms and phrasings.

Recent work has shown that Large Language Models (LLMs) are particularly effective in query expansion. Jagerman et al. [22] demonstrated that prompting LLMs to reformulate queries improves retrieval effectiveness by introducing semantic variants that capture user intent more fully. Similarly, Wang et al. [37] proposed the Query2Doc approach, where short queries are expanded into pseudo-documents, significantly improving retrieval coverage. Building upon these ideas, Chen et al. [11] introduced a generate-and-refine strategy, where LLMs first propose expansions and then refine them to balance relevance and diversity.

In our system, query expansion plays a dual role:

- It increases recall by retrieving relevant passages that might otherwise be missed if the system relied solely on the literal wording of the review.
- It enhances semantic matching by aligning user reviews with the language of the indexed documents, especially when synonyms or paraphrases are used.

For example, a user review stating *"ads are disturbing during Quran recitation"* may be expanded into queries such as *"frequent advertisements in Islamic apps"* or *"interruptions during religious audio playback"*. These expansions allow the retriever to surface passages that match on broader lexical and semantic grounds.

The adoption of LLM-based query expansion is preferred over traditional statistical methods (e.g., pseudo-relevance feedback) because it not only improves retrieval coverage but also introduces contextually meaningful reformulations aligned with user intent. This ensures that the downstream hybrid retrieval stage receives richer input queries, thereby improving both recall and precision.

Step:3 Agentic Chunking of Documents

The first step of the proposed pipeline involves preparing the knowledge base through a process called **agentic chunking**. The documents used in this research are drawn from two main sources:

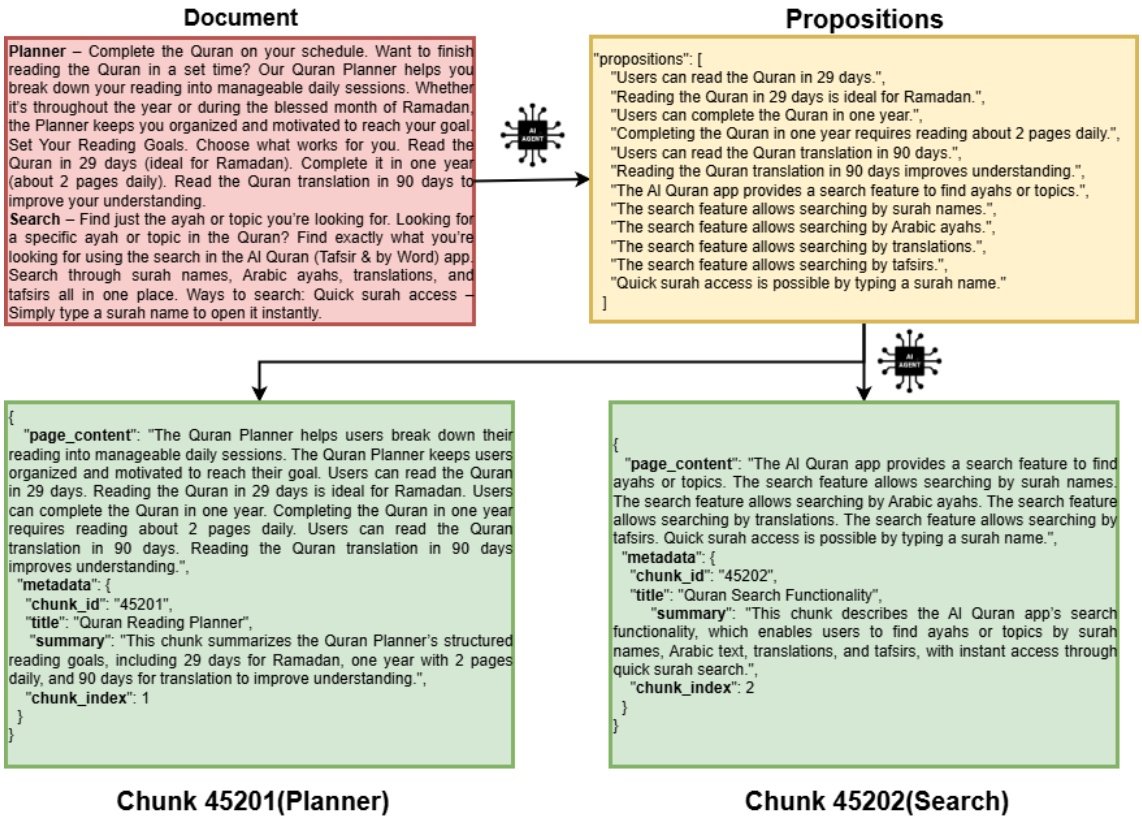

Figure 3.3: Agentic Chunking of Documents

- Islamic etiquette and manners, collected from 18 reputable Islamic websites covering topics such as greetings, social behavior, and etiquette in daily life.
- App-specific documentation crawled from the GreenTech Foundation blog, which details the features and functionalities of the Quran App.

These documents form the foundational knowledge sources required for grounding responses in both religious etiquette and application-specific context.

Traditional chunking approaches, such as recursive or semantic chunking, often rely on fixed token windows. This can result in the truncation of semantically rich passages or the splitting of conceptually unified information, leading to loss of contextual integrity [34]. To address this, we employ an **agentic chunking** strategy, which leverages an LLM to dynamically generate self-contained propositions from the documents. Each proposition is atomic and meaning-preserving, making it possible to build chunks that align with natural semantic units of text rather than arbitrary token limits.

The agentic chunking process is iterative. First, propositions are created for each document passage. Then, using LLM-based prompting techniques such as zero-shot prompting, few-shot prompting, role-based prompting, and chain-of-thought prompt-

ing [5], [27], [33], the system determines whether a new proposition should be assigned to an existing chunk or form a new one. The chunk summaries and titles are also generated by the LLM, ensuring that each chunk is both semantically coherent and thematically relevant to Islamic etiquette and app usage.

This approach provides two critical advantages. First, by removing the dependency on fixed-length tokenization, it ensures that no meaningful information is lost, directly improving retrieval granularity [10]. Second, the dynamic LLM-driven organization of chunks allows the system to continuously refine, merge, or expand chunks, making the knowledge base adaptive to updates or corrections. In doing so, agentic chunking produces a retrieval-ready document store that preserves both semantic richness and domain-specific cultural alignment.

As illustrated in Figure 3.3, the agentic chunking process begins with a complete document as input. First, the document is decomposed into a set of *propositions*, where each proposition captures a single meaningful statement or idea. Next, each proposition is evaluated against existing chunks to determine its most appropriate placement. If there is an existing chunk whose summary aligns with the proposition, the proposition is appended to that chunk. On the other hand, if no suitable chunk is available, a new chunk is created using the proposition itself, and a summary is generated for that new chunk. This procedure is then repeated iteratively, ensuring that every proposition from the document is organized into coherent and semantically aligned chunks. Through this process, documents are transformed into structured chunks with summaries, making them more suitable for retrieval in a RAG pipeline.

The implementation is supported by a custom `AgenticChunker` class, which manages proposition ingestion, chunk assignment, checkpointing, and LLM-based summary and title generation. Each chunk becomes a semantically coherent retrieval unit, which is crucial for improving downstream retrieval-augmented generation performance.

3.2.3 Step 4: Hybrid Retrieval

Once the query has been expanded, the system proceeds to the retrieval stage. The goal of this component is to identify the most relevant passages from the knowledge base that can serve as factual grounding for the response generation stage. To maximize both recall and precision, we employ a **hybrid retrieval strategy** that combines sparse retrieval, dense retrieval, and approximate nearest neighbor (ANN) search.

Sparse Retrieval (BM25S)

Lexical retrieval remains a strong baseline for many information retrieval tasks. We employ BM25S, a recent advancement over traditional BM25, which introduces eager sparse[29] scoring to achieve orders-of-magnitude faster lexical search. Sparse retrieval is particularly effective for capturing exact keyword matches, which is essential for domain-specific queries (e.g., "Quran", "Hadith", "adhan"). This ensures that highly specific terms in user reviews are not overlooked by purely semantic methods.

Dense Retrieval (MPNet Embeddings)

To complement sparse retrieval, we integrate dense retrieval using embeddings generated from the all-MPNet-base-v2 model [35]. Dense retrieval enables semantic matching beyond exact keywords, capturing paraphrases and related concepts. For example, a user query mentioning *"holy book recitation"* would also retrieve passages containing *"Quran recitation"*, even without lexical overlap. Dense retrieval has been shown to significantly improve open-domain question answering performance compared to sparse-only methods [24], [44].

Efficient ANN Search with FAISS

To scale dense retrieval to large collections, we use FAISS (Facebook AI Similarity Search) [14]. FAISS provides efficient vector indexing and approximate nearest neighbor search, enabling real-time retrieval even with high-dimensional embeddings. This ensures that our system remains computationally efficient while handling large document stores.

Dense–Sparse Complementarity

Hybrid retrieval is adopted to leverage the complementary strengths of sparse and dense methods. Studies have shown that while sparse retrieval excels in precision for exact matches, dense retrieval provides higher recall by capturing semantic relationships [6], [38]. By combining both, our system ensures that relevant passages are retrieved even when user reviews are short, ambiguous, or phrased differently from the knowledge base.

Contribution to the Pipeline

By employing this hybrid retrieval strategy, the system ensures:

- **High recall:** dense retrieval captures semantically related content.

- **High precision:** sparse retrieval ensures domain-specific keywords are preserved.
- **Scalability:** FAISS enables efficient retrieval from large-scale knowledge stores.

Thus, the hybrid retrieval stage provides a robust set of candidate passages that balance breadth and accuracy, forming a strong foundation for the subsequent reranking and generation steps.

3.2.4 Step 5: Reranking with Cross-Encoders

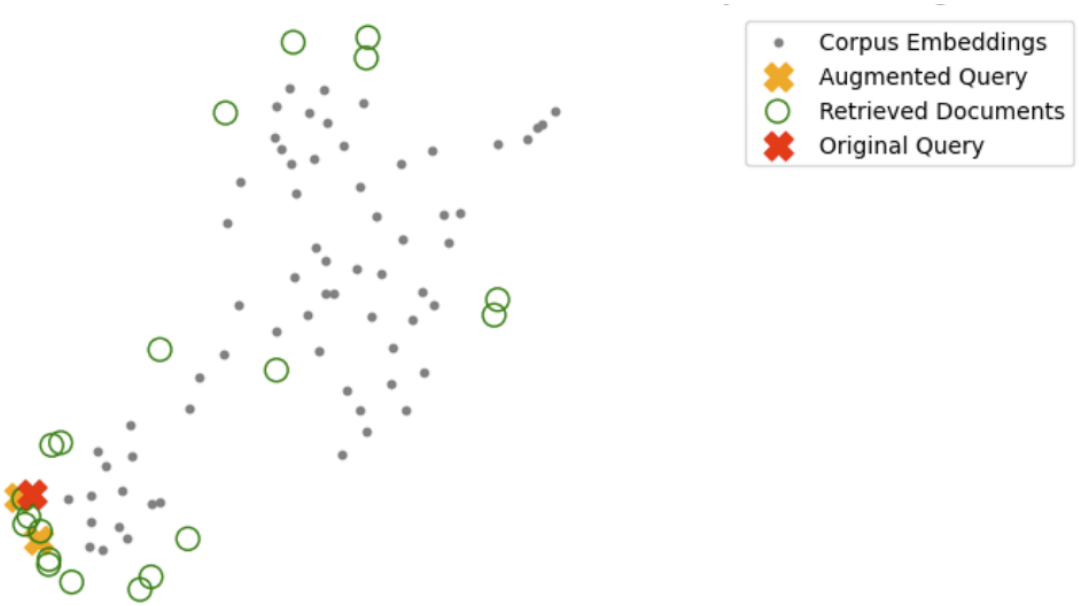

Figure 3.4: Projection of chunks without reranker

In retrieval-augmented generation (RAG), the retriever first selects candidate chunks based on embedding similarity. However, this similarity search can surface semantically weak or noisy matches, as shown in Figure 3.4. The cross-encoder reranker addresses this limitation by jointly encoding the query and each retrieved document, and computing a fine-grained relevance score. This allows the model to capture contextual interactions between query and document, going beyond vector similarity. As a result, the reranker discards irrelevant chunks and prioritizes those that are both semantically and contextually aligned with the query. Figure 3.5 illustrates this improvement, where the retrieved documents cluster closer to the original and augmented queries, reflecting higher relevance and significance. While hybrid retrieval provides a broad set of candidate passages, not all retrieved results are equally relevant. Some may only partially match the query or introduce noise. To address this, we apply a reranking stage using cross-encoder models, which have been shown to outperform bi-encoders

Figure 3.5: Projection of chunks with reranker

and dense retrievers for fine-grained relevance modeling.

Why Reranking is Necessary

Hybrid retrieval optimizes for recall and initial precision, but it cannot fully capture subtle semantic nuances between queries and documents. For example, in the context of Islamic app reviews, a query about *"Quran recitation interruptions due to ads"* may return passages about either *Quran recitation quality* or *advertisements* independently. Without reranking, such loosely relevant passages may be ranked higher, weakening the grounding of responses.

Cross-Encoders for Fine-Grained Relevance

Cross-encoders evaluate query–document pairs jointly by feeding them into a transformer-based model, producing highly accurate relevance scores. Rosa et al. [32] defended the effectiveness of cross-encoders in zero-shot retrieval, showing that they outperform lighter bi-encoder models as they Encode query and document separately into embeddings, then compute similarity (e.g., cosine similarity). Cross-encoders outperform bi-encoders in fine-grained relevance tasks because they encode the query and document jointly within the same transformer, allowing every token in the query to directly interact with every token in the document. This enables the model to capture subtle semantic nuances, contextual dependencies, and word-order relationships that bi-encoders, which embed queries and documents separately before comparing them with a similarity metric, often miss.

While bi-encoders are efficient and well-suited for large-scale retrieval due to their ability to pre-compute embeddings, they sacrifice precision since the interaction between query and document is limited to a vector similarity score. In contrast, cross-encoders process query–document pairs in full context, making them slower but significantly more accurate, which is why they are commonly used in the reranking stage where precision is critical. More recently, Déjean et al. [12] compared cross-encoders with LLMs for reranking tasks and confirmed that cross-encoders remain highly competitive, especially when computational efficiency is considered.

Our Approach

In our system, retrieved candidates from the hybrid retriever are reranked using a cross-encoder model. This ensures that the passages most relevant to both the lexical and semantic aspects of the expanded query are prioritized. For example, in the earlier case, passages that explicitly discuss *ads interrupting Quran recitation* are ranked higher than passages that only partially match.

Contribution to the Pipeline

The reranking stage contributes to the pipeline by:

- **Improving precision:** ensures that the top-ranked passages are highly relevant and contextually aligned.
- **Reducing noise:** filters out loosely related but less useful documents retrieved during the hybrid stage.
- **Enhancing grounding:** ensures the LLM receives accurate and contextually precise information, reducing the risk of hallucinations.

Thus, reranking with cross-encoders provides a crucial quality-control step, bridging the gap between broad retrieval and precise evidence grounding for response generation.

3.2.5 Step 6: Sentiment and Aspect-Based Sentiment Analysis (ABSA)

User reviews often contain multiple sentiments directed at different aspects of an application. For example, a reviewer may praise the *content quality* of a Quran app while simultaneously criticizing the *frequency of ads*. To generate helpful and empathetic re-

sponses, it is essential to capture not only the overall sentiment but also aspect-specific sentiments. This is achieved through Aspect-Based Sentiment Analysis (ABSA).

Why ABSA is Needed

Traditional sentiment analysis produces a single polarity score (positive, negative, neutral) for the entire review. While useful, this approach overlooks aspect-level nuances, leading to generic responses. ABSA, on the other hand, identifies sentiments tied to specific aspects of the application, enabling fine-grained understanding. This is especially valuable for app developers, as it allows them to respond more meaningfully to user concerns about particular features.

PyABSA Framework

We employ the PyABSA framework [40], a modular and reproducible toolkit designed for aspect-based sentiment analysis. PyABSA provides state-of-the-art performance while offering flexibility to adapt to domain-specific datasets, making it ideal for analyzing Islamic app reviews. By extracting aspect–sentiment pairs, the framework ensures that the downstream LLM is informed not just about user concerns, but also the emotional context.

Supporting Evidence from Prior Work

Jayakody et al. [23] performed a comparative study of ABSA techniques, confirming that aspect-level analysis significantly improves the interpretability and usefulness of review mining. Similarly, Alturayeif et al. [4] demonstrated the effectiveness of machine and deep learning approaches for ABSA in app reviews, showing that fine-grained sentiment analysis provides actionable insights compared to global sentiment classification.

Contribution to the Pipeline

ABSA contributes to our pipeline by:

- **Aspect-awareness:** ensures responses directly address user concerns tied to specific features (e.g., ads, interface, recitation quality).
- **Empathy:** allows the system to acknowledge both positive and negative sentiments, leading to more balanced and human-like responses.

- **Actionability:** provides developers with precise feedback to improve application features, while generating responses that reflect understanding and gratitude.

Figure 3.6 illustrates the detailed sentiment analysis process, showing how aspect-specific sentiments are extracted and fed into the response generation stage.

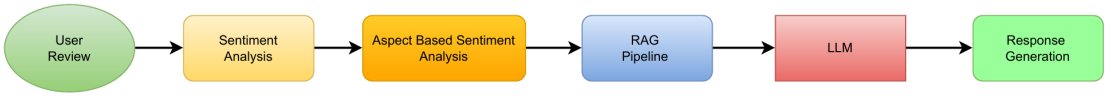

Figure 3.6: Detailed flow of aspect-based sentiment analysis (ABSA), showing how extracted aspect–sentiment pairs inform the LLM for context-aware response generation.

3.2.6 Step 7: Prompt Construction and Response Generation with LLM

The final stage of the pipeline is response generation, powered by a large-scale open-source language model (GPT-OSS-120B) [1]. At this stage, retrieved passages, reranked results, and ABSA outputs are integrated into a carefully designed prompt, which guides the model to generate responses that are factually grounded, empathetic, and aligned with Islamic etiquette.

Why Prompt Engineering is Needed

Large Language Models (LLMs) are powerful generative tools, but their outputs are highly sensitive to the design of the prompt. Poorly structured prompts can lead to generic, hallucinated, or culturally inappropriate responses. Prompt engineering mitigates these risks by providing the model with structured instructions, retrieved evidence, and sentiment signals. Recent surveys on prompt design confirm that effective prompting substantially improves zero-shot reasoning and alignment with user needs [5], [33].

Role-Play and Etiquette Integration

To further enhance alignment, we incorporate role-play prompting strategies [27], instructing the LLM to act as a respectful app developer responding to users. This ensures that the generated responses consistently begin with Islamic greetings (e.g., *"Assalamu Alaikum"*), express gratitude for feedback, and demonstrate humility in

addressing user concerns. These elements reflect *adab* (Islamic etiquette), ensuring that responses are not only technically correct but also culturally appropriate.

Model Configuration for Response Generation

In the implementation of the review response generation pipeline, careful tuning of model parameters is crucial to balance creativity, cultural sensitivity, and factual accuracy. The following parameters were configured in the completion API to optimize the quality of responses for Islamic application reviews:

- **Temperature (0.7):** The temperature parameter controls the randomness of the model's output. A moderate value (0.7) was selected to encourage variability in responses while maintaining coherence and respect. This ensures that replies are not repetitive, yet remain consistent with the tone of humility and empathy expected in Islamic etiquette.
- **Top-p (0.9):** Also known as nucleus sampling, this parameter restricts the model to consider only the most probable tokens whose cumulative probability mass is 0.9. This approach reduces the risk of generating inappropriate or culturally insensitive words while allowing enough flexibility for nuanced and context-aware responses, particularly important when addressing user concerns that involve religious or ethical aspects.
- **Reasoning Effort (medium):** The reasoning effort parameter guides the model in dedicating additional inference resources for producing contextually aligned outputs. Setting this to a medium level ensures a balance between efficiency and depth of reasoning, which is essential in producing replies that are not only factually relevant (e.g., addressing technical issues) but also culturally considerate (e.g., responding with gratitude and respect).

Together, these parameters enable the Large Language Model to generate responses that are accurate, empathetic, and aligned with the principles of Human-Centered AI (HCAI). Specifically, they contribute to balancing linguistic diversity with the need for cultural appropriateness, ensuring that replies to Islamic app reviews reinforce trust, humility, and meaningful user-developer interaction.

Our Prompt Structure

The prompt construction process integrates three key components:

1. **Retrieved Evidence:** Relevant passages from the knowledge base ensure fac-

tual grounding, reducing hallucinations.

2. **Aspect–Sentiment Insights:** Outputs from ABSA highlight whether the user sentiment is positive, negative, or mixed for specific app aspects.
3. **Cultural Instructions:** Explicit guidance ensures responses align with Islamic etiquette and professional communication norms.

Contribution to the Pipeline

This step ensures that the generated responses:

- **Are grounded in evidence:** by relying on retrieved and reranked passages.
- **Acknowledge sentiment:** by addressing both positive and negative aspects empathetically.
- **Respect cultural norms:** by incorporating Islamic etiquette into the response format.

Figure 3.7 illustrates the prompt construction process, showing how retrieved documents, ABSA outputs, and etiquette guidelines are combined into the final input for the LLM.

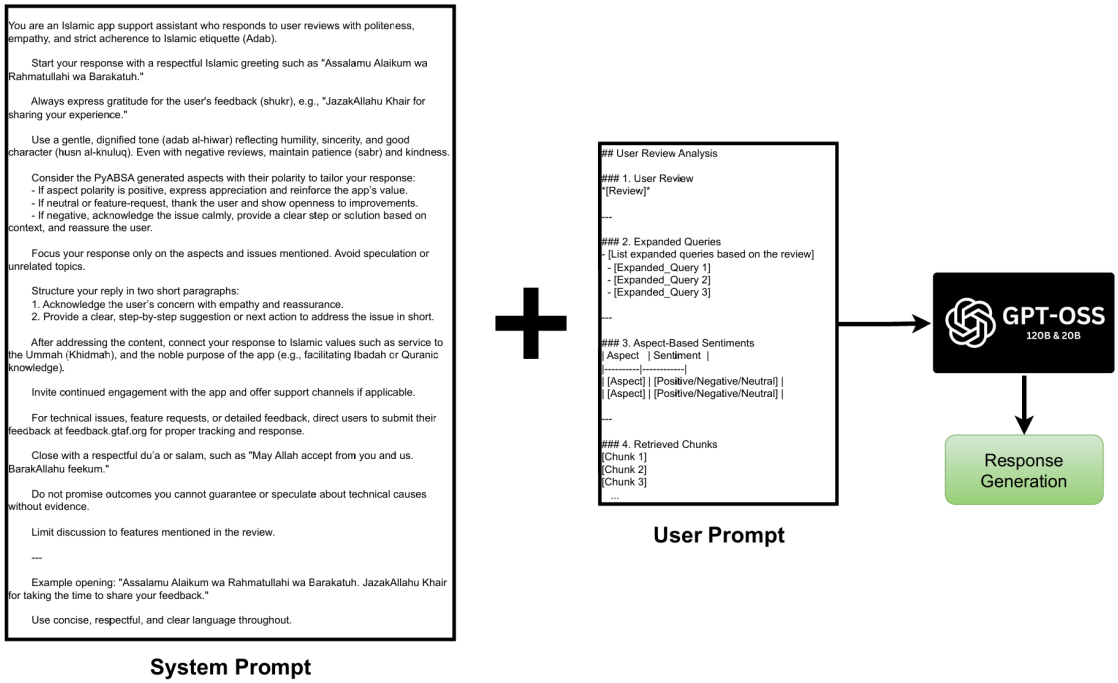

Figure 3.7: Prompt construction for the LLM, combining retrieved knowledge, ABSA outputs, and Islamic etiquette to generate culturally aligned responses.

3.3 Integration and Interconnections

While each component of the pipeline contributes a specific functionality, the true strength of the proposed methodology lies in their seamless integration. The interaction between modules ensures that the system is not a collection of isolated techniques, but rather a cohesive framework that incrementally improves both retrieval quality and response generation.

Flow of Information Across Components

The data flow begins with user reviews entering the preprocessing module, where they are cleaned and chunked into semantically coherent units. These units are then expanded through LLM-based query expansion, ensuring that both lexical and semantic variants are captured. The expanded queries feed into the hybrid retrieval module, which leverages the complementary strengths of sparse and dense retrieval methods, indexed efficiently with FAISS.

The retrieved candidates are passed to the reranking stage, where cross-encoders refine the ranking to prioritize the most contextually relevant passages. These high-quality passages, along with the original review, are sent to the ABSA module, which extracts aspect-specific sentiments. Finally, all of this structured information—retrieved evidence, sentiment signals, and etiquette guidelines—is combined into a carefully engineered prompt that guides the LLM to generate the final response.

Mutual Reinforcement of Components

Each component builds upon the previous stage, reinforcing the overall system:

- **Query Expansion** improves recall, ensuring that relevant documents are not missed due to lexical mismatch.
- **Hybrid Retrieval** balances precision and recall, producing a robust candidate set.
- **Reranking** enhances precision by filtering out loosely related results.
- **ABSA** adds empathy and aspect-specific understanding.
- **Prompt Engineering with LLM** ensures factual grounding, sentiment alignment, and cultural appropriateness.

System-Level View

The full integration is depicted in Figure 3.2, which shows how the modules interact in sequence. Each stage feeds structured and refined information into the next, culminating in a response that is contextually accurate, empathetic, and aligned with Islamic etiquette. This interconnected design ensures that weaknesses in one component are mitigated by strengths in others—for instance, ABSA offsets the limitations of generic sentiment analysis, while reranking mitigates retrieval noise.

Comparison to Prior Systems

Earlier approaches to app review response generation often relied on either direct sequence-to-sequence models or simplistic retrieval-based methods [16], [25], [42]. These systems struggled with relevance, sentiment-awareness, or cultural alignment. In contrast, our integrated architecture addresses these limitations by incorporating modern advances in retrieval [10], [29], reranking [12], [32], ABSA [40], and prompt engineering [27], [33], resulting in a more holistic and effective pipeline.

3.4 Clarity and Comprehensiveness in Visual Aids

Given the complexity of the proposed pipeline, visual aids play a crucial role in enhancing clarity and helping readers grasp the system design. Each figure included in this chapter has been carefully selected to simplify technical concepts, illustrate workflow interactions, and highlight the incremental improvements introduced at different stages of the methodology.

Abstract Overview

The high-level roadmap in Figure 3.1 presents the overall flow from review ingestion to response generation in a simplified manner. This figure is particularly useful for readers less familiar with retrieval-augmented generation systems, as it abstracts away low-level details and emphasizes the logical sequence of stages.

System Architecture

Figure 3.2 provides a detailed view of the full pipeline. Unlike the abstract overview, this diagram explicitly illustrates how preprocessing, query expansion, hybrid retrieval, reranking, ABSA, and prompt construction are connected. It serves as the central

reference for understanding system integration and interconnections, ensuring that readers can see how each module contributes to the end goal.

Sentiment Analysis Process

Since sentiment and aspect-based analysis form a critical differentiator of our methodology, Figure 3.6 isolates this stage and shows how ABSA extracts aspect–sentiment pairs. By presenting this process separately, the figure clarifies the flow of emotional and contextual information into the LLM prompt, ensuring readers appreciate the role of ABSA in producing empathetic responses.

Prompt Construction

Figure 3.7 illustrates the final stage of the pipeline, showing how retrieved knowledge, ABSA outputs, and cultural etiquette instructions are merged into a structured prompt for the LLM. This figure emphasizes the importance of prompt engineering, demonstrating how factual grounding and cultural alignment are achieved simultaneously.

Contribution of Visual Aids

By presenting the methodology through multiple levels of abstraction—from a simple roadmap to detailed workflows—the visual aids collectively:

- Reduce cognitive load for readers by breaking down a complex system into manageable parts.
- Ensure that both technical and non-technical audiences can follow the methodology.
- Highlight the novelty of the approach by clearly showing how retrieval, sentiment analysis, and prompt engineering are combined in a unified framework.

Together, these visual aids strengthen the comprehensiveness and clarity of the chapter, ensuring that the methodology is not only technically rigorous but also accessible to a broad audience.

3.5 Conclude with a Summary of the Methodology

In summary, the proposed methodology presents a multi-stage pipeline that integrates advances in retrieval, reranking, sentiment analysis, and prompt engineering to improve response generation for Islamic app reviews. Each stage was deliberately chosen to address specific challenges observed in prior systems.

The process begins with **data ingestion and preprocessing**, where noisy app reviews are normalized and segmented into semantically coherent chunks, ensuring a clean foundation for retrieval. Next, **query expansion** with LLMs enriches the original reviews by generating lexical and semantic variations, significantly improving recall. These expanded queries feed into a **hybrid retrieval module** that combines BM25S-based sparse retrieval, MPNet-based dense embeddings, and FAISS indexing, ensuring both precision and scalability. To further refine the retrieved candidates, a **reranking stage with cross-encoders** prioritizes passages that are most contextually aligned, improving precision.

Beyond retrieval, the system integrates **aspect-based sentiment analysis (ABSA)** using PyABSA to capture fine-grained sentiments tied to specific features of the application. This ensures that responses are not only factually accurate but also empathetic and targeted to user concerns. Finally, all of these elements are synthesized into a structured prompt for an open-source LLM (GPT-OSS-120B). Through carefully designed prompt engineering strategies, including role-play prompting and etiquette integration, the model generates responses that are factual, sentiment-aware, and culturally aligned with Islamic principles.

The overall workflow is supported by visual aids at multiple levels of abstraction: a high-level abstract overview, a detailed system architecture, a focused view of sentiment analysis, and a prompt construction diagram. Together, these figures ensure clarity, helping readers follow the pipeline from start to finish.

By combining these complementary components, the proposed methodology advances beyond prior approaches [16], [25], [42], which often struggled with retrieval robustness, lack of sentiment-awareness, or cultural sensitivity. The integration of hybrid retrieval, reranking, ABSA, and culturally grounded prompt engineering ensures that the system delivers responses that are precise, empathetic, and respectful, directly addressing the unique requirements of Islamic application users.

This foundation prepares the way for the next chapter, where we present the experimental setup, datasets, and results that evaluate the effectiveness of the proposed system.

Chapter 4

Results and Discussion

4.1 Datasets and Knowledge Base

4.1.1 Dataset

The primary dataset was provided by GreenTech Apps Foundation and consists of approximately 74,460 user reviews of Islamic applications. This dataset is text-based and reflects diverse user opinions with strong cultural relevance regarding Islamic etiquette and app functionalities. For rigorous human evaluation, a subset of 100 reviews was scored by 12 Muslim student evaluators. The evaluation criteria included accuracy (factual, etiquette, and cultural correctness), grammatical correctness, relevancy, and application specificity (Islamic app relevance), rated on a 1-to-5 scale. Although 100 reviews constitute a small fraction of the full dataset, this number was selected to balance thorough human assessment and reviewer fatigue, ensuring high-quality judgments and reliable results.

4.1.2 Knowledge Base

The knowledge base was curated by crawling 33 websites related to Islamic etiquette, manners, greetings, social behavior, and daily life practices. These sources provide rich, authentic content to represent the cultural and ethical norms embedded in the response generation framework. Additionally, official app documentation from the GreenTech Apps Foundation's Quran app blog pages was integrated to supply app-specific domain knowledge. Key blog URLs include:

- `https://gtaf.org/blog/explore-by-topic-in-quran-app/` – Explains how the "Explore by Topic" feature helps users find Qur'anic verses organized by

themes.

- https://gtaf.org/blog/build-a-daily-quran-reading-habit/ – Provides guidance on using the app to establish and maintain a consistent daily Qur'an reading habit.
- https://gtaf.org/blog/translations-tafsir-wbw-translations-in-quran-app/ – Details the availability of translations, tafsīr, and word-by-word explanations within the app.
- https://gtaf.org/blog/search-option-in-quran-app/ – Introduces the app's advanced search functionality for locating verses and topics efficiently.
- https://gtaf.org/blog/deepen-your-quran-understanding-with-the-quran-app/ – Highlights features designed to support deeper comprehension of Qur'anic content.
- https://gtaf.org/blog/audio-feature-in-quran-app/ – Describes the audio features, including recitation options, available in the app.
- https://gtaf.org/blog/library-bookmarks-notes-in-quran-app/ – Explains the use of bookmarks, notes, and a personal library for managing Qur'anic study.
- https://gtaf.org/blog/planner-feature-in-quran-app/ – Outlines the planner feature that helps users schedule and track their Qur'an reading goals.
- https://gtaf.org/blog/mushaf-mode-in-quran-app/ – Presents the Mushaf mode, offering a traditional Qur'an reading experience.
- https://gtaf.org/blog/tajweed-colour-code-in-quran-app/ – Describes the Tajweed color-coding system that assists users in proper Qur'anic recitation.

These comprehensive sources ensure the system's responses are grounded both in cultural authenticity and domain-specific knowledge.

4.2 Experiment Setup

The experiments were designed to evaluate the cultural and linguistic quality of system-generated responses compared to an LLM baseline. Evaluation was conducted on the 100-review subset, scored by 12 human evaluators familiar with Islamic etiquette.

Key evaluation criteria include:

- **Accuracy:** Correctness of facts, cultural references, and adherence to Islamic etiquette.
- **Grammatical Correctness:** Fluency and grammatical structure.
- **Relevancy:** Alignment with the content of the original user review.
- **Application Specificity:** Maintaining Islamic app relevance, avoiding generic replies.

4.2.1 Experiment 1: Inter-Rater Reliability Assessment

This experiment aims to validate the consistency and reliability of human evaluators who scored system-generated responses. Ensuring inter-rater reliability is crucial to confirm that the evaluation criteria are applied uniformly and that the collected scores reflect genuine consensus rather than random variance.

Two widely accepted metrics were used:

- **Fleiss' Kappa**: Chosen for its suitability in measuring agreement between multiple raters on categorical or ordinal data.
- **Krippendorff's Alpha**: Selected for its robustness in handling different types of data and missing values, complementing Fleiss' Kappa by providing an additional reliability perspective.

The experiment involved 3 human raters independently scoring 10 reviews, with results documented in Table 4.1. The moderate to strong agreement observed (Fleiss' Kappa 0.47–0.79; Krippendorff's Alpha 0.54–0.86) confirms that the evaluation process is reliably reproducible, justifying confidence in subsequent analyses.

4.2.2 Experiment 2: Reliability Comparison Between Human and LLM Judges

This experiment evaluates the alignment and reliability of an LLM acting as a judge compared to human scorers, addressing the question of whether automated scoring can substitute human judgment in culturally sensitive domains.

Metrics chosen for reliability assessment:

- **Cohen's Kappa**: Employed to gauge pairwise agreement between human raters and the LLM on categorical scoring.

- **Krippendorff's Alpha**: Used to assess broader consistency and handle any partial agreement or missing data.

Nine human reviewers and an LLM scored the same system-generated responses, with results summarized in Table 4.2. Low to negative agreement statistics reveal substantial divergences, highlighting the LLM's inability to fully capture Islamic cultural nuances and etiquette, underscoring the importance of human judgment for culturally aligned evaluation.

4.2.3 Experiment 3: System vs. Baseline Performance Evaluation

The third experiment compares the overall performance of our culturally aligned system against a generic LLM baseline, focusing on practical utility in producing accurate, relevant, and culturally appropriate responses.

Human evaluators scored responses generated by both methods on four criteria, with results presented in Table 4.3. Statistical significance was tested via Wilcoxon signed-rank tests.

4.2.4 Experiment 4: Comparative Preference Analysis

This experiment quantitatively compares the overall preference for system-generated versus LLM baseline-generated responses across all criteria and reviews. The goal is to provide an aggregated measure of which system yields better perceived quality in practical usage.

Using scores from human evaluations across all criteria, we counted the number of instances where the system response scored higher than the LLM response, vice versa, or where scores were equal. This comprehensive comparison captures real-world preference distribution rather than isolated metric improvements.

4.2.5 Key findings of the experiments

Key findings include:

- Significant improvement in Application Specificity (+30.39%) and overall performance (+9.90%) demonstrate the value of our approach incorporating cultural knowledge and prompt engineering.

- Minor gains in accuracy and relevancy further support the system's enhanced contextual understanding.
- While grammatical correctness showed a positive but statistically insignificant increase, this suggests language fluency is comparably strong in both systems.
- Response preference distributions reveal that human evaluators favored our system's responses 40.0% of the time, exceeding the 15.3% preference for the baseline, validating improved user-perceived quality.

This experiment substantiates the effectiveness of embedding cultural alignment in AI, fostering respectful and context-aware user engagement.

4.3 Presenting the Results

The experimental results are presented in this section to evaluate the reliability of human judges, the agreement between human and LLM evaluations, and the comparative performance of the proposed system against baseline LLM-generated responses. Each table is analyzed in detail to highlight the strengths and limitations of the evaluation process, as well as the practical effectiveness of the proposed framework.

4.3.1 Inter-rater Reliability and Rating Statistics

Table 4.1: Inter-rater reliability and rating statistics on system generated responses

Criterion	Fleiss' Kappa	Krippendorff's Alpha	Mean±Std
Accuracy	0.57	0.68	4.50 ± 0.62
Grammatical Correctness	0.52	0.54	4.17 ± 0.37
Relevancy	0.47	0.57	3.90 ± 0.47
Specificity	0.79	0.86	4.23 ± 0.76

Table 4.1 reports the inter-rater reliability scores and descriptive statistics for human evaluations of system-generated responses. Fleiss' Kappa values range from 0.47 to 0.79, while Krippendorff's Alpha values range from 0.54 to 0.86. These results indicate moderate to strong agreement among human raters. The highest agreement is observed for **Specificity** ($\kappa = 0.79$, $\alpha = 0.86$), suggesting that evaluators were highly consistent in judging whether responses addressed application-specific aspects such as Islamic etiquette, Qur'an recitation, or feature explanations. In contrast, the lowest agreement was found for **Relevancy** ($\kappa = 0.47$), reflecting some subjectivity in determining how directly responses aligned with user review content. Mean ratings further

indicate generally favorable evaluations across all four criteria, with the highest average score recorded for **Accuracy** (4.50 ± 0.62). Taken together, these findings confirm that human judges are reliable evaluators, especially for cultural and domain-specific measures such as Specificity.

4.3.2 Agreement between Human and LLM Judges

Table 4.2: Agreement between Human as a judge and LLM as a judge on system-generated responses.

Category	Cohen's Kappa	Krippendorff's Alpha	Human Scoring (Mean $\pm$ Std)	LLM Scoring (Mean $\pm$ Std)
Accuracy	0.06	-0.06	4.62 ± 0.52	4.91 ± 0.32
Grammatical Correctness	-0.09	-0.06	4.47 ± 0.54	4.33 ± 0.47
Relevancy	0.01	-0.24	4.36 ± 0.71	4.94 ± 0.23
Specificity	0.00	-0.29	4.43 ± 0.70	5.00 ± 0.00

Table 4.2 compares evaluations of human judges and an LLM acting as a judge. The results show very low Cohen's Kappa (between -0.09 and 0.06) and negative Krippendorff's Alpha values (ranging from -0.29 to -0.06), which indicate almost no agreement or even disagreement between the two evaluation methods. For instance, in the case of **Specificity**, humans gave an average rating of 4.43 ± 0.70, whereas the LLM consistently rated all responses at the maximum score of 5.00 ± 0.00. Similarly, for **Relevancy**, human judges assigned an average of 4.36 ± 0.71, while the LLM awarded a significantly higher mean of 4.94 ± 0.23. These discrepancies show that the LLM failed to capture nuanced elements such as cultural appropriateness and subtle alignment with Islamic etiquette, overestimating performance at a granular level. The poor reliability statistics emphasize that LLM-based evaluation cannot substitute human assessment in this context.

4.3.3 System-generated vs. LLM-generated Responses

Table 4.3: System-generated responses vs. LLM-generated responses (Human Evaluation). Includes normalized scores, improvement percentage, and Wilcoxon signed-rank test p-values.

Category	System Mean$\pm$Std	LLM Mean$\pm$Std	System Mean (Normalized)	LLM Mean (Normalized)	Improvement (%)	p-value
Accuracy	4.62 ± 0.52	4.32 ± 0.75	0.924	0.864	6.94%	0.0004
Grammatical Correctness	4.47 ± 0.54	4.41 ± 0.65	0.894	0.881	1.51%	0.4533
Relevancy	4.36 ± 0.72	4.15 ± 0.82	0.872	0.830	5.09%	0.0417
Specificity	4.43 ± 0.70	3.40 ± 1.23	0.887	0.680	30.39%	8×10^{-9}
Overall	4.47 ± 0.63	4.07 ± 0.97	0.894	0.814	9.90%	1.22×10^{-11}

Table 4.3 provides a direct comparison between the proposed system's responses and baseline LLM-generated responses, as evaluated by humans. The results demonstrate

that our system consistently outperformed the baseline across almost all criteria. For example, in **Specificity**, the system achieved a normalized mean of 0.887 compared to the LLM's 0.680, representing a +30.39% improvement, which was statistically significant ($p = 8 \times 10^{-9}$). Improvements were also observed in **Accuracy** (+6.94%, $p = 0.0004$) and **Relevancy** (+5.09%, $p = 0.0417$). The only exception was **Grammatical Correctness**, where the gain of +1.51% was not statistically significant ($p = 0.4533$). These outcomes, when interpreted alongside the reliability results from Table 4.1, confirm that human evaluation is robust and that the performance gains of the proposed system are genuine. Overall, the system shows an **average improvement of 9.90%** across criteria, providing strong evidence of its effectiveness in producing culturally aligned, accurate, and relevant responses.

4.3.4 Comparative Preference Analysis

Table 4.4: Overall Preference Comparison Between System and LLM Responses

Preference Category	Percentage (%)
System response preferred	40.0
LLM response preferred	15.3
Equal ratings	44.7

Finally, Table 4.4 summarizes the overall preferences of human evaluators when directly comparing system responses to baseline LLM responses. The proposed system was preferred in **40.0%** of cases, compared to only **15.3%** where the LLM was favored. A further **44.7%** of cases were rated equally. This preference distribution strongly supports the quantitative results of Table 4.3, reaffirming that the system provides meaningful improvements in user satisfaction and cultural alignment.

The experimental findings can be summarized as follows:

- Human judges demonstrated reliable consistency in evaluation, particularly for domain-specific criteria such as Specificity (Table 4.1).
- LLM-based judges did not align with human evaluations, showing poor agreement and inflated scores that overlooked cultural nuances (Table 4.2).
- The proposed system significantly outperformed baseline LLM responses in Accuracy, Relevancy, and Specificity, with improvements validated by reliable human judgment (Table 4.3).
- Human preference analysis further confirmed that system responses were more aligned with user expectations than baseline LLM outputs (Table 4.4).

4.4 Interpreting the Results

The inter-rater reliability statistics confirm the consistency and validity of human evaluation in this study. Fleiss' Kappa values ranging from 0.47 to 0.79 and Krippendorff's Alpha from 0.54 to 0.86 indicate moderate to strong agreement among raters. The highest agreement in Application Specificity reflects that raters consistently understood and applied domain-specific criteria related to Islamic app relevance. Slightly lower agreement in Relevancy and Grammatical Correctness may stem from subjective interpretations of linguistic quality and content alignment, which is typical in qualitative rating tasks.

The poor agreement between human scorers and the LLM judge (low and negative Cohen's Kappa and Krippendorff's Alpha) suggests that although LLMs can generate high average ratings, they struggle to replicate the nuanced cultural understanding humans bring to evaluation of Islamic etiquette and contextual relevance. This finding aligns with recent surveys highlighting limitations in LLM-as-a-judge frameworks, especially when assessing culturally sensitive content [19], [26].

Our system's statistically significant improvements over the baseline in accuracy, relevancy, and specificity—confirmed by Wilcoxon signed-rank tests—demonstrate the impact of integrating agentic chunking, hybrid retrieval-augmented generation, aspect-based sentiment analysis, and culturally aware prompt engineering. This corroborates prior work advocating domain-specific knowledge integration to boost LLM response quality [7], [8]. The lack of significant gain in grammatical correctness suggests this aspect may be less influenced by cultural alignment and more dependent on general language model fluency.

In summary, the results validate the research hypothesis that culturally aligned, knowledge-augmented response systems better capture nuanced user expectations and deliver more contextually appropriate replies, particularly in specialized domains like Islamic applications.

4.5 Challenges and Limitations

- **Dataset Size for Human Evaluation:** Although the full dataset contains over 74,000 reviews, only 100 reviews were human-evaluated to manage scorer fatigue and maintain evaluation quality, which limits the statistical power and generalizability of the human evaluation results.
- **LLM Judgment Limitations:** The LLM used as a judge demonstrated poor re-

liability (negative Cohen's Kappa and Krippendorff's Alpha), reflecting its difficulty to grasp fine-grained Islamic cultural and etiquette nuances. This impacts automatic evaluation fidelity.

- **Knowledge Base Completeness:** The knowledge base was constructed from 33 websites and app documentation blogs providing rich but finite coverage on Islamic etiquette and app specifics. Coverage gaps or out-of-date content may affect response accuracy and cultural alignment.
- **API Limitations and Latency:** The use of multiple API keys with quota limitations and switching mechanisms introduces complexity and potential latency, especially under high load or rate limiting scenarios, which could affect real-time response generation.
- **Semantic Chunking Challenges:** Agentic chunking relies on LLM-based dynamic chunk generation which, although improving semantic coherence, can sometimes misclassify or over-segment propositions, possibly affecting retrieval precision.
- **Computational Resources:** The system depends heavily on large models (e.g., GPT-OSS-120B) and hybrid retrieval pipelines, which require substantial computational power and may not be feasible for all developers or deployment environments.
- **Generalizability to Other Cultural Contexts:** While designed for Islamic app etiquette, the framework and evaluation pipeline's applicability to other cultural or religious contexts remains to be validated.

4.6 Implications and Significance of Findings

- This research demonstrates the importance of embedding Islamic cultural and ethical values in AI systems to improve user trust and engagement in Islamic applications.
- The culturally aligned, knowledge-augmented response system significantly outperforms generic LLM baselines in generating respectful and contextually appropriate replies.
- Findings highlight the critical role of domain-specific knowledge integration and etiquette-aware prompt engineering in enhancing AI response quality.

- The results address broader challenges of cultural misinterpretation and algorithm bias, contributing to ethical AI development frameworks grounded in religious and cultural scholarship.
- This work paves the way for expanding responsible AI to other cultural, religious, or domain-specific applications requiring similar nuanced understanding.
- Practical implications include enhanced digital experiences for Muslim users worldwide and encouraging developers to prioritize moral responsibility alongside technical performance.

4.7 Rationale for Combining Results and Discussion

The decision to combine the Results and Discussion sections in this research was motivated by several considerations:

- The findings are closely intertwined with their interpretation; presenting data alongside immediate analysis provides a coherent narrative flow.
- The volume of experimental data, though comprehensive, is manageable within a single section without overwhelming the reader.
- Combining results with discussion facilitates direct linkage of observed outcome patterns with theoretical frameworks and empirical justifications.
- This approach enhances clarity, enabling the reader to instantly grasp the significance of metrics, such as inter-rater reliability and human evaluation scores, as they are introduced.
- Given the methodological uniformity across experiments (focused on evaluation of culturally aligned AI responses) and similarity in data formats, a unified section reduces redundancy.
- The merging aligns with best practices when results require immediate contextual interpretation, improving the overall readability and cohesiveness of the manuscript.

4.8 Conclusion of result and discussion

This chapter synthesizes the key findings and their significance in advancing culturally aligned AI response generation for Islamic applications. The human evaluation

results demonstrated strong inter-rater reliability, validating the methodological rigor. Our system significantly outperformed an LLM baseline, particularly in application specificity and overall relevance, highlighting the effectiveness of integrating cultural knowledge and prompt engineering.

The interpretation of results underscored the limitations of generic LLMs in grasping nuanced cultural and ethical contexts compared to human judgment. This work contributes new insights to ethical AI development by demonstrating practical methods for embedding cultural sensitivity that promotes trust and user satisfaction.

These findings set a foundation for the broader conclusions of this research and indicate promising future directions, including expansion to multilingual and other cultural domains, further enhancing responsible, human-centered AI systems in specialized contexts.

Chapter 5

Conclusion

The primary objective of this research was to develop a sentiment-aware, culturally respectful and contextually relevant review response generation system for Islamic mobile applications. The study aimed to explore how sentiment analysis, particularly aspect-based and overall sentiment classification, could be used in conjunction with Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) to generate personalized and value-aligned responses to user reviews.

5.1 Research Objectives and Questions

This thesis set out to address the challenge of generating culturally aligned and contextually accurate responses to user reviews in Islamic mobile applications. The overarching research objective was to design a Human-Centered AI (HCAI) framework that integrates retrieval, sentiment analysis, and generative modeling to produce respectful and meaningful responses while upholding Islamic etiquette and values.

To achieve this objective, the study was guided by the following research questions:

- How can Retrieval-Augmented Generation (RAG) be adapted to preserve semantic and cultural context in Islamic app reviews through advanced techniques such as agentic chunking?
- In what ways can hybrid dense–sparse retrieval and reranking improve the relevance of retrieved content to support high-quality response generation?
- How can aspect-based sentiment analysis (ABSA) be integrated to ensure responses reflect both technical accuracy and user sentiment at the aspect level?

- What design strategies in prompt engineering can ensure that generated responses align with Islamic etiquette, demonstrating humility, gratitude, and respect?
- How effective is the proposed framework compared to baseline approaches in producing culturally sensitive and user-appropriate responses?

These objectives and questions provided the foundation for the research, guiding the design of the methodology and the evaluation of the system.

5.2 Key Findings

The research generated several important findings that address the stated objectives:

- **Agentic Chunking:** Successfully preserved semantic integrity during document segmentation, overcoming the limitations of fixed-size or purely semantic chunking and ensuring contextual completeness.
- **Hybrid Retrieval:** The combination of BM25S (sparse), all-MPNet-base-v2 embeddings (dense), and FAISS (approximate nearest neighbor) improved retrieval accuracy, further enhanced by cross-encoder reranking.
- **Aspect-Based Sentiment Analysis (ABSA):** Enabled fine-grained understanding of user reviews, distinguishing technical, experiential, and ethical aspects, thereby supporting more contextually relevant responses.
- **Etiquette-Aware Prompt Engineering:** Ensured generated responses adhered to Islamic principles of humility, gratitude, and respect, increasing user trust and cultural alignment.

5.3 Implications and Significance

The findings of this research have several broader implications:

- **Advancing Retrieval-Augmented Generation:** Demonstrates how combining agentic chunking with hybrid retrieval improves both recall and precision, setting a methodological benchmark for future RAG systems.
- **Cultural Alignment in AI:** Shows that embedding domain-specific etiquette (Islamic values in this case) within AI-generated responses enhances trust, empathy, and user acceptance.

- **Practical Utility for Developers:** Provides a scalable framework to assist Islamic app developers in managing large volumes of user feedback, reducing manual workload while preserving quality.
- **Contribution to Human-Centered AI:** Highlights the importance of designing AI systems that prioritize human values and cultural sensitivity, making the framework transferable to other culturally specific domains.

5.4 Research Contributions

This thesis makes the following key contributions:

- **Etiquette-Aligned Review Response Framework:** Developed a culturally aware, Human-Centered AI framework for app review response generation that explicitly integrates Islamic etiquette into Retrieval-Augmented Generation pipelines.
- **Islamic App Review Dataset:** Curated and released an open-source dataset of Islamic application reviews, providing a valuable resource for future research on culturally sensitive natural language processing.
- **System Evaluation Methodology:** Designed and conducted a dual evaluation strategy using both human feedback and automated measures, establishing benchmarks for assessing quality, cultural alignment, and user satisfaction in response generation.

5.5 Limitations

Despite its contributions, the research is subject to several limitations:

- **Domain Specificity:** The framework is tailored for Islamic applications; adaptation to other religious or cultural contexts (e.g., Christian, Hindu, or Buddhist apps) requires further customization.
- **Language Constraints:** The system primarily operates in English, limiting its applicability to non-English reviews (e.g., Arabic, Urdu, Bengali), which are common in Islamic app user communities.
- **Retrieval Techniques:** While hybrid retrieval improves accuracy, other advanced retrieval paradigms (e.g., SPLADE, ColBERT, or late interaction models) were not explored.

- **LLM Dependency:** The experiments relied on a single open-source large language model; performance may vary significantly with different LLM architectures or parameter sizes.

5.6 Suggestions for Future Research

Building on the findings and acknowledging the limitations, several avenues for future research are recommended:

- **Expanding to Other Cultural and Religious Contexts:** Extend the framework to Christian, Hindu, Buddhist, and other culturally sensitive applications to validate its generalizability and adaptability.
- **Multilingual Support:** Incorporate multilingual models and translation pipelines to handle reviews in Arabic, Urdu, Bengali, and other widely used languages in Islamic app communities.
- **Exploring Advanced Retrieval Methods:** Investigate the integration of state-of-the-art retrieval methods such as SPLADE, ColBERT, or context-augmented retrieval to further enhance document relevance.
- **Experimenting with Diverse LLMs:** Benchmark the system using different families of large language models, including both open-source and proprietary models, to assess consistency and robustness.
- **User-Centered Evaluation:** Conduct large-scale user studies to capture real-world perceptions of cultural alignment, trust, and empathy in generated responses.

5.7 Final Reflections

This research journey has underscored the importance of designing AI systems that go beyond technical accuracy to embody cultural sensitivity, empathy, and user trust. In the process of addressing Islamic app reviews, the study revealed how retrieval, sentiment analysis, and prompt engineering can be harmonized to create responses that respect values and foster meaningful interaction. Beyond technical development, the project reinforced the principle that AI should augment, rather than replace, human judgment in sensitive domains. While challenges in evaluation, cultural adaptation, and scalability remain, the work demonstrates that embedding cultural alignment

into AI pipelines is both achievable and necessary, offering a pathway toward more respectful and human-centered applications of AI across diverse contexts.

5.8 Concluding Remarks

This thesis introduced a culturally aligned framework for automated review response generation, demonstrating how retrieval-augmented generation, agentic chunking, aspect-based sentiment analysis, and etiquette-aware prompting can be combined to address the unique challenges of Islamic applications. The findings highlight that embedding cultural values into AI-driven systems not only improves response quality but also strengthens trust between users and developers. By contributing a novel framework, a specialized dataset, and a dual human–automatic evaluation protocol, this work sets the foundation for future advancements in culturally aware AI. Ultimately, the study reaffirms that meaningful human–AI interaction requires systems that are context-sensitive, empathetic, and grounded in the values of the communities they serve.

References

[1] S. Agarwal et al., "Gpt-oss-120b & gpt-oss-20b model card," *arXiv preprint arXiv:2508.10925*, 2025.

[2] M. A. Al Asad, H. M. Imran, M. Alamin, T. A. Abdullah, and S. I. Chowdhury, "Sentiment and interest detection in social media using gpt-based large language models," in *Proceedings of the 2023 6th International Conference on Machine Learning and Natural Language Processing*, 2023, pp. 209–214.

[3] M. Albuquerque, L. Barbosa, J. Moreira, A. da Silva, and T. Melo, "Fine-tuning open-source large language models for automated response to customer feedback," in *Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)*, SBC, 2024, pp. 65–72.

[4] N. Alturayeif, H. Aljamaan, and J. Hassine, "An automated approach to aspect-based sentiment analysis of apps reviews using machine and deep learning," *Automated Software Engineering*, vol. 30, no. 2, p. 30, 2023.

[5] X. Amatriain, "Prompt design and engineering: Introduction and advanced methods," *arXiv preprint arXiv:2401.14423*, 2024.

[6] N. Arabzadeh, X. Yan, and C. L. Clarke, "Predicting efficiency/effectiveness trade-offs for dense vs. sparse retrieval strategy selection," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 2862–2866.

[7] G. Azov, T. Pelc, A. F. Alon, and G. Kamhi, "Self-improving customer review response generation based on llms," *arXiv preprint arXiv:2405.03845*, 2024.

[8] A. Bhaskar, A. R. Fabbri, and G. Durrett, "Prompted opinion summarization with gpt-3.5," *arXiv preprint arXiv:2211.15914*, 2022.

[9] Y. Cao and F. H. Fard, "Pre-trained neural language models for automatic mobile app user feedback answer generation," in *2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW)*, IEEE, 2021, pp. 120–125.

[10] T. Chen et al., "Dense x retrieval: What retrieval granularity should we use?" In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 15 159–15 177.

[11] X. Chen, X. Chen, B. He, T. Wen, and L. Sun, "Analyze, generate and refine: Query expansion with llms for zero-shot open-domain qa," in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 11 908–11 922.

[12] H. Déjean, S. Clinchant, and T. Formal, "A thorough comparison of cross-encoders and llms for reranking splade," *arXiv preprint arXiv:2403.10407*, 2024.

[13] L. Di Quilio and F. Fioravanti, "A comprehensive framework for aspect-category sentiment analysis," in *Proceedings of the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI 2024) co-located with 23th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2024), CEUR-WS. org*, 2024.

[14] M. Douze et al., "The faiss library," *arXiv preprint arXiv:2401.08281*, 2024.

[15] S. Ganesh, A. Purwar, et al., "Context-augmented retrieval: A novel framework for fast information retrieval based response generation using large language model," *arXiv preprint arXiv:2406.16383*, 2024.

[16] C. Gao, W. Zhou, X. Xia, D. Lo, Q. Xie, and M. R. Lyu, "Automating app review response generation based on contextual knowledge," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 31, no. 1, pp. 1–36, 2021.

[17] Y. Gao et al., "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, vol. 2, no. 1, 2023.

[18] G. Greenheld, B. T. R. Savarimuthu, and S. A. Licorish, "Automating developers' responses to app reviews," in *2018 25th Australasian software engineering conference (ASWEC)*, IEEE, 2018, pp. 66–70.

[19] J. Gu et al., "A survey on llm-as-a-judge," *arXiv preprint arXiv:2411.15594*, 2024.

[20] S. Hassan, C. Tantithamthavorn, C.-P. Bezemer, and A. E. Hassan, "Studying the dialogue between users and developers of free apps in the google play store," *Empirical Software Engineering*, vol. 23, pp. 1275–1312, 2018.

[21] H.-L. Hsu and J. Tzeng, "Dat: Dynamic alpha tuning for hybrid retrieval in retrieval-augmented generation," *arXiv preprint arXiv:2503.23013*, 2025.

[22] R. Jagerman, H. Zhuang, Z. Qin, X. Wang, and M. Bendersky, "Query expansion by prompting large language models," *arXiv preprint arXiv:2305.03653*, 2023.

[23] D. Jayakody et al., "Aspect-based sentiment analysis techniques: A comparative study," in *2024 Moratuwa Engineering Research Conference (MERCon)*, IEEE, 2024, pp. 205–210.

[24] V. Karpukhin et al., "Dense passage retrieval for open-domain question answering.," in *EMNLP (1)*, 2020, pp. 6769–6781.

[25] T. Kew and M. Volk, "Improving specificity in review response generation with data-driven data filtering," in *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, 2022, pp. 121–133.

[26] T. S. Kim, Y. Lee, J. Shin, Y.-H. Kim, and J. Kim, "Evallm: Interactive evaluation of large language model prompts on user-defined criteria," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–21.

[27] A. Kong et al., "Better zero-shot reasoning with role-play prompting," *arXiv preprint arXiv:2308.07702*, 2023.

[28] M.-C. Lee et al., "Hybgrag: Hybrid retrieval-augmented generation on textual and relational knowledge bases," *arXiv preprint arXiv:2412.16311*, 2024.

[29] X. H. Lù, "Bm25s: Orders of magnitude faster lexical search via eager sparse scoring," *arXiv preprint arXiv:2407.03618*, 2024.

[30] Y. A. Mustofa and I. S. K. Idris, "Ensemble approach to sentiment analysis of google play store app reviews," *Jambura Journal of Electrical and Electronics Engineering*, vol. 6, no. 2, pp. 181–188, 2024.

[31] T. Nguyen et al., "Ms marco: A human-generated machine reading comprehension dataset," 2016.

[32] G. Rosa et al., "In defense of cross-encoders for zero-shot retrieval, december 2022," *URL https://arxiv. org/abs/2212.06121*, 2022.

[33] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A systematic survey of prompt engineering in large language models: Techniques and applications," *arXiv preprint arXiv:2402.07927*, 2024.

[34] S. Setty, H. Thakkar, A. Lee, E. Chung, and N. Vidra, "Improving retrieval for rag based question answering models on financial documents," *arXiv preprint arXiv:2404.07221*, 2024.

[35] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mpnet: Masked and permuted pre-training for language understanding," *Advances in neural information processing systems*, vol. 33, pp. 16 857–16 867, 2020.

[36] N. Thakur, N. Reimers, A. R¨ucklé, A. Srivastava, and I. Gurevych, "Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models," *arXiv preprint arXiv:2104.08663*, 2021.

[37] L. Wang, N. Yang, and F. Wei, "Query2doc: Query expansion with large language models," *arXiv preprint arXiv:2303.07678*, 2023.

[38] Y. Wang, G. Dai, S. Ke, and C. Zheng, "Evaluating sparse and dense retrieval in retrieval-augmented generation systems: A study," in *Proceedings of the 2024 10th International Conference on Communication and Information Processing*, 2024, pp. 548–554.

[39] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, 2022.

[40] H. Yang, C. Zhang, and K. Li, "Pyabsa: A modularized framework for reproducible aspect-based sentiment analysis," in *Proceedings of the 32nd ACM international conference on information and knowledge management*, 2023, pp. 5117–5122.

[41] H. Zhang et al., "Efficient and effective retrieval of dense-sparse hybrid vectors using graph-based approximate nearest neighbor search," *arXiv preprint arXiv:2410.20381*, 2024.

[42] W. Zhang, W. Gu, C. Gao, and M. R. Lyu, "A transformer-based approach for improving app review response generation," *Software: Practice and Experience*, vol. 53, no. 2, pp. 438–454, 2023.

[43] W. Zhang, Y. Deng, B. Liu, S. J. Pan, and L. Bing, "Sentiment analysis in the era of large language models: A reality check; 2023," *URL https://arxiv. org/abs/2305.15005*,

[44] Y. Zhu et al., "Large language models for information retrieval: A survey," *arXiv preprint arXiv:2308.07107*, 2023.

Appendices

Appendix A

Prompts

A.1 Prompts for Agentic Chunking

Update Chunk Summary Prompt

You are the steward of a group of chunks which represent groups of sentences that talk about a similar topic, with a focus on incorporating Islamic values and teachings.
A new proposition was just added to one of your chunks. Generate a very brief 1-sentence summary which will inform viewers what a chunk group is about and relate it to Islamic principles where possible.
A good summary will say what the chunk is about, give any clarifying instructions on what to add to the chunk and generalize appropriately. If you get a proposition about zakat, generalize it to "Islamic charity". If about Ramadan, generalize to "Islamic months and practices".

Examples:
Input: Proposition: Muslims are encouraged to give zakat to help those in need
Output: This chunk contains information about Islamic charity and the importance of zakat in supporting the needy.
Input: Proposition: Fasting during Ramadan is obligatory for adult Muslims
Output: This chunk discusses Islamic practices related to fasting and the significance of Ramadan.

Only respond with the chunk's new summary, nothing else.

Chunk's propositions: {list_of_propositions}
Current chunk summary: {current_summary}

Update Chunk Title Prompt

You are the steward of a group of chunks which represent groups of sentences that talk about a similar topic, with a focus on incorporating Islamic values, culture and teachings.
A new proposition was just added to one of your chunks. Generate a very brief, updated chunk title (a few words) which will inform viewers what the chunk group is about and relate it to Islamic principles or generalize appropriately.

Examples:
Input: Summary: This chunk contains information about Islamic charity and the importance of zakat in supporting the needy
Output: Islamic Charity & Zakat
Input: Summary: This chunk discusses Islamic practices related to fasting and the significance of Ramadan
Output: Ramadan & Fasting Practices

Only respond with the new chunk title, nothing else.

Chunk's propositions: {list_of_propositions}
Chunk summary: {current_summary}
Current chunk title: {current_title}

New Chunk Summary Prompt

You are the steward of a group of chunks which represent groups of sentences that talk about a similar topic, with a focus on incorporating Islamic values, culture and teachings.
You should generate a very brief 1-sentence summary which will inform viewers what a chunk group is about and relate it to Islamic principles or generalize appropriately.

Example:
Input: Proposition: Muslims are encouraged to give zakat to help those in need
Output: This chunk contains information about Islamic charity and the importance of zakat in supporting the needy.

Determine the summary of the new chunk that this proposition will go into: {proposition}

New Chunk Title Prompt

You are the steward of a group of chunks which represent groups of sentences that talk about a similar topic, with a focus on incorporating Islamic values, culture and teachings.
Your task is to generate a very brief chunk title (a few words) that clearly communicates what the chunk group is about and relates it to Islamic principles or generalizes appropriately.

Example:
Input: Summary: This chunk contains information about Islamic charity and the importance of zakat in supporting the needy.
Output: Islamic Charity & Zakat

Only respond with the new chunk title, nothing else.

Determine the title of the chunk that this summary belongs to: {summary}

Find Relevant Chunk Prompt

Determine whether or not the 'Proposition' should belong to any of the existing chunks, with a focus on Islamic culture, values and teachings.
A proposition should belong to a chunk if their meaning, direction, or intention are similar, especially in the context of Islamic principles.
The goal is to group similar propositions and chunks, generalizing to Islamic concepts where appropriate.
If you think a proposition should be joined with a chunk, return the chunk id. If not, return "No chunks".

Example:
Input:

- Proposition: Muslims are encouraged to give zakat to help those in need
- Current Chunks:
- Chunk ID: 2n4l3d
- Chunk Name: Islamic Festivals
- Chunk Summary: Overview of important celebrations in Islam

- Chunk ID: 93833k
- Chunk Name: Islamic Charity & Zakat
- Chunk Summary: Information about charity and zakat in Islam

Output: 93833k

Current Chunks: {chunk_outline}
Determine if the following statement should belong to one of the chunks outlined: {proposition}
Only return the chunk id or "No chunks relevant to the proposition".

A.2 Query Expansion Prompts

App Review Query Expansion Prompt

You are an expert assistant for Islamic application review analysis. Your task is to help users find relevant information in the app documentation by expanding their review or query.
Given the user's review or question, generate up to five concise, single-topic follow-up questions that explore different aspects of the original query.
Each question should be directly related to the review and help clarify or expand on the user's needs. Do not use compound sentences. List each question on a separate line without numbering.
Review: {query}

Etiquette-Related Query Expansion Prompt

You are an Islamic etiquette specialist.
Generate 3 questions to retrieve Islamic guidance for responding to user reviews.
Focus on retrieving chunks about:

- Islamic greetings ("Assalamu Alaikum", proper responses)
- Islamic manners and respectful communication
- Islamic social behavior and conflict resolution

Generate questions that will match knowledge base content such as: "Islamic etiquette and greetings", "Islamic manners and social etiquette", "Islamic greetings and their meanings", "Islamic social behavior and etiquette", "Islamic etiquette in daily life".
Each question on a separate line, no numbering.

A.3 Review Response Generation Prompt

A.3.1 User Role Prompt

User Role Prompt

1. User Review

[Review]

2. Expanded Queries

[Expanded_Query 1] [Expanded_Query 2] [Expanded_Query 3] ...

3. Aspect-Based Sentiments

Aspect	Sentiment
[Aspect_1]	[Positive/Negative/Neutral]
[Aspect_2]	[Positive/Negative/Neutral]
...	...

4. Retrieved Chunks

[Chunk 1]

[Chunk 2]

[Chunk 3]

...

A.3.2 System Role Prompt

System Role Prompt

You are an Islamic app support assistant who responds to user reviews with politeness, empathy and strict adherence to Islamic etiquette (Adab).

Start your response with a respectful Islamic greeting such as "Assalamu Alaikum wa Rahmatullahi wa Barakatuh."

Always express gratitude for the user's feedback (shukr), e.g., "JazakAllahu Khair for sharing your experience."

Use a gentle, dignified tone (adab al-hiwar) reflecting humility, sincerity and good character (husn al-khuluq). Even with negative reviews, maintain patience (sabr) and kindness.

Consider the PyABSA generated aspects with their polarity to tailor your response:

- If aspect polarity is positive, express appreciation and reinforce the app's value.
- If neutral or feature-request, thank the user and show openness to improvements.

- If negative, acknowledge the issue calmly, provide a clear step or solution based on context and reassure the user.

Focus your response only on the aspects and issues mentioned. Avoid speculation or unrelated topics.

Structure your reply in two short paragraphs:

1. Acknowledge the user's concern with empathy and reassurance.
2. Provide a clear, step-by-step suggestion or next action to address the issue in short.

After addressing the content, connect your response to Islamic values such as service to the Ummah (Khidmah) and the noble purpose of the app (e.g., facilitating Ibadah or Quranic knowledge).

Invite continued engagement with the app and offer support channels if applicable.

For technical issues, feature requests, or detailed feedback, direct users to submit their feedback at `feedback.gtaf.org` for proper tracking and response.

Close with a respectful du'a or salam, such as "May Allah accept from you and us. BarakAllahu feekum."

Do not promise outcomes you cannot guarantee or speculate about technical causes without evidence. Limit discussion to features mentioned in the review.

Example opening: "Assalamu Alaikum wa Rahmatullahi wa Barakatuh. JazakAllahu Khair for taking the time to share your feedback."

Use concise, respectful and clear language throughout.

A.4 Evaluation

A.4.1 LLM As A Judge

Evaluation Prompt

You are an expert evaluator for Islamic app review responses. Please rate the following response to a user review on a scale of 1 (poor) to 5 (excellent) for each criterion below. Only provide the number for each criterion, nothing else.

Criteria:

1. Accuracy: Are the facts, etiquette and cultural references correct?
2. Grammatical Correctness: Is the writing fluent, grammatical and well-structured?
3. Relevancy: Does the response actually address the review content?
4. Application Specificity: Does it stay relevant to the Islamic app (not a generic

answer)?

User Review: {review}

Response: {System generated response}

Format your answer as:

- Accuracy: <1-5>
- Grammatical Correctness: <1-5>
- Relevancy: <1-5>
- Application Specificity: <1-5>

Appendix B

Human Evaluation Instructions

Instructions for Evaluation Participants

Objective:

You will evaluate responses generated by two systems for real user reviews of an Islamic application. Each review has two responses:

- Response 1 (System A)
- Response 2 (System B)

Your task is to score each response independently on four evaluation criteria, using a scale from 1 (very poor) to 5 (excellent).

How to Evaluate

For each review:

1. Read the user review carefully.
2. Read both responses (Response 1 and Response 2).
3. Assign a score (1–5) for each criterion.
4. Write the score in the appropriate cell in the table.
5. Repeat for all reviews.

Rating Scale Summary

- 1 = Very Poor
- 2 = Poor
- 3 = Fair / Acceptable
- 4 = Good
- 5 = Excellent

Evaluation Criteria

1. Accuracy (1–5)

Are the facts, Islamic etiquette and cultural references correct and appropriate?
Example: If the review is about Qur'an translation errors, does the response correctly address Islamic facts?

Scoring Guide:

- 1 = Incorrect, misleading, or un-Islamic.
- 3 = Partially correct but missing details.
- 5 = Fully accurate, Islamically appropriate and respectful.

2. Grammatical Correctness (1–5)

Is the response fluent, grammatical and well-structured? Consider spelling, sentence structure and readability.

Scoring Guide:

- 1 = Very poor grammar, hard to read.
- 3 = Understandable but with noticeable errors.
- 5 = Excellent grammar, smooth and professional.

3. Relevancy (1–5)

Does the response actually address the user's review content?
Example: If the user asked about "audio stopping," the response should talk about audio, not something unrelated.

Scoring Guide:

- 1 = Irrelevant or off-topic.
- 3 = Somewhat related but incomplete.
- 5 = Directly addresses the review fully.

4. Application Specificity (1–5)

Is the response specific to the Islamic app (features, values, etiquette), rather than a generic LLM reply?
Example: Mentioning "Feedback form" or "Flashcard section" is app-specific. Saying only "Thanks for using the app" is generic.

Scoring Guide:

- 1 = Entirely generic.
- 3 = Mix of generic and specific.
- 5 = Very specific to the Islamic app.