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

K.M. Tahlil Mahfuz Faruk[1], **Mushfiqur Rahman Talha**[1], **H.M. Kawsar Ahamad**[1], **Mohammad Galib Shams**[2], **Nabil Mosharraf Hossain**[3], **Syed Rifat Raiyan**[1], **Md Kamrul Hasan**[1], **Hasan Mahmud**[1], **Riasat Islam**[3]
[1]Systems and Software Lab (SSL), Department of Computer Science and Engineering, Islamic University of Technology, Gazipur, Bangladesh
[2]Delineate Inc., Cambridge, MA, USA
[3]Greentech Apps Foundation, Huntingdon, United Kingdom
{tahlilmahfuz, mushfiqurtalha, kawsarahamad, rifatraiyan, hasank, hasan}@iut-dhaka.edu, galib@delineate.pro, {nabil, riasat.islam}@gtaf.org

Abstract

Automated review response systems have advanced considerably, yet most fail to incorporate Islamic etiquette, values, and cultural norms, which are essential for meaningful engagement with users who are adherents of the Islamic faith. 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 confirms significant improvements in accuracy ($p = 0.0004$), relevancy ($p = 0.0417$) and specificity ($p = 8 \times 10^{-9}$), while grammatical correctness shows 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.

1 Introduction

Artificial intelligence has transformed user engagement with applications, especially through LLM-powered review response systems Ramesh and Kumar [2025], Gao and Smith [2023]. Due to the sheer volume of feedback, where nearly 40% of apps garner over 10,000 reviews while developers respond to fewer than 3%, manual responses are pretty much impractical Hassan et al. [2018]. Despite broader advances, mainstream technologies provide inadequate support for Islamic applications. In an effort to ameliorate this, we present ADAB, a multifaceted pipeline tailored to Islamic applications, which

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Muslims in ML.

synergistically integrates sentiment analysis, aspect extraction, information retrieval, and generative modeling. Using Retrieval-Augmented Generation (RAG), reviews are analyzed to produce responses reflecting Islamic etiquette. The key challenges of our work span semantic context preservation, cultural alignment, and precise retrieval. The system supports developers via etiquette-aware prompt engineering advancing culturally sensitive, human-centered AI in this domain Gao et al. [2021], Hassan et al. [2018]. It addresses limitations in existing systems that lack cultural sensitivity. Failure to align automated responses with Islamic etiquette can inadvertently offend users, erode trust, and diminish engagement—underscoring the necessity of culturally aware AI communication systems. Combining dense and sparse retrieval, reranking, and aspect-based sentiment analysis (ABSA), the pipeline generates contextually relevant, empathetic responses. Prompt engineering techniques further embed humility, gratitude, and respect aligned with Islamic values Yang et al. [2023], Azov et al. [2024]. This research contributes ADAB, an etiquette-aligned framework tailored for Islamic applications that generates accurate, respectful, and empathetic responses consistent with Islamic etiquette. The contribution lies in creatively integrating and applying existing techniques (RAG, ABSA, prompt engineering) for Islamic app review responses, not in proposing new algorithms. It has been evaluated using a hybrid methodology that combines human assessments with LLM-based judgments, providing a culturally appropriate approach to evaluation. In this paper, our key contributions are as follows:

- We introduce **ADAB**, a culturally-aligned automated review response framework integrating **hybrid RAG**, **aspect-based sentiment analysis (ABSA)**, and **etiquette-aware prompt engineering** to produce Islamically appropriate responses.
- We incorporate a **agentic chunking mechanism** that dynamically aligns retrieval units with semantic and cultural coherence rather than fixed token limits.
- We conduct a **dual human–LLM evaluation**, revealing significant improvements in accuracy, relevancy, and application specificity, and highlighting the current inability of LLMs to capture nuanced cultural etiquette.

2 Related Work

Automated review response generation has evolved significantly. Early approaches by Gao et al. [2019] and Greenheld et al. [2018] established foundational frameworks using machine learning techniques. These were later enhanced by Zhang et al. [2023], who introduced transformer-based architectures, demonstrating improved response quality through contextual understanding. Jagerman et al. [2023] explored query expansion techniques using LLMs, while Amatriain [2024] provided comprehensive guidelines for prompt engineering, establishing best practices for LLM-based text generation. Kong et al. [2023] advanced role-play prompting for zero-shot reasoning and Azov et al. [2024] developed self-improving mechanisms for customer review responses. Recent works have focused on domain-specific improvements. Gao et al. [2021] incorporated contextual knowledge for app review responses, Cao and Fard [2021] leveraged pre-trained models for mobile app feedback and Albuquerque et al. [2024] fine-tuned open-source LLMs for customer feedback automation. Data quality enhancements were addressed by Kew and Volk [2022] through data-driven filtering approaches, while Katsiuba et al. [2022] developed comprehensive customer feedback management systems. For evaluation, the emerging strategy is to utilize LLM as a Judge Gu et al. [2024], Zhou et al. [2022], Zheng et al. [2023], Kim et al. [2024]. Despite these advances, existing literature lacks Islamic cultural values, etiquette, and communication norms.

3 Proposed Methodology

3.1 Agentic Chunking of Documents

The proposed pipeline incorporates a knowledge base using **agentic chunking**, which addresses the limitations of fixed-token chunking that often truncate semantically rich passages and disrupt contextual integrity Setty et al. [2024]. Our documents are sourced from 18 Islamic websites detailing etiquette and manners, alongside app-specific documentation from the Greentech Apps Foundation blog on Quran App features. The agentic chunking method leverages an LLM to dynamically generate atomic, meaning-preserving propositions from the text, aligning chunks with natural semantic units

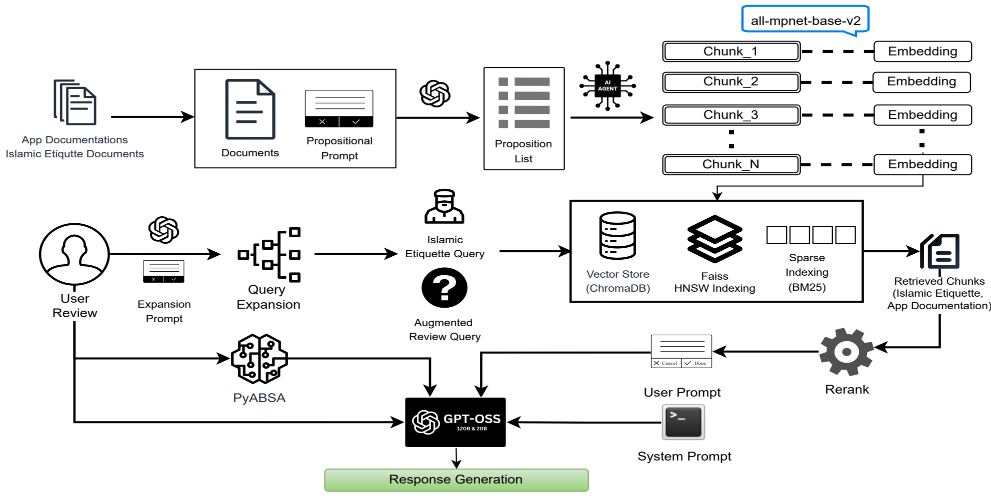

Figure 1: System architecture of ADAB showing the flow of information across Agentic Chunker, Retriever, Reranking, ABSA, and Prompt Engineering

rather than arbitrary token limits. This process is iterative: propositions are created for each passage, then LLM-based prompting techniques, including zero-shot, few-shot, role-based, and chain-of-thought prompting Sahoo et al. [2024], Amatriain [2024], Kong et al. [2023] determine whether to assign new propositions to existing chunks or form new ones. The LLM also generates chunk summaries and titles, ensuring semantic coherence and thematic relevance to Islamic etiquette and app usage. This approach enhances retrieval granularity by preserving meaningful context Chen et al. [2024a].

3.2 Hybrid Retriever with Query Expansion and Reranking

User reviews are often short and ambiguous. To overcome this, our approach employs **query expansion** to reformulate reviews into richer, more comprehensive queries. Recent studies demonstrate that LLMs are effective for this task. For instance, Jagerman et al. [2023] showed that LLM-based reformulation introduces semantic variants that better capture user intent. Wang et al. [2023] proposed the Query2Doc method, transforming short queries into pseudo-documents, thereby enhancing retrieval coverage. Chen et al. [2024b] developed a generate-and-refine strategy where expansions are iteratively proposed and refined to balance relevance and diversity. In our system, query expansion serves two major purposes: it increases recall by retrieving passages that might be missed using literal queries, and it enhances semantic matching, aligning user input with indexed document language.

Following query expansion, the system employs a **hybrid retrieval strategy** which combines sparse lexical retrieval, dense semantic retrieval and approximate nearest neighbor (ANN) search. Sparse retrieval using BM25S Lù [2024] efficiently captures exact keyword matches. Complementing this, dense retrieval leverages embeddings from the all-MPNet-base-v2 model Song et al. [2020] to identify semantically related passages. For example, matching "holy book recitation" with "Quran recitation." To enable scalability over large document sets, FAISS HNSW Douze et al. [2024] provides high-performance ANN search for dense vectors. The hybrid approach exploits the strengths of each method, with sparse retrieval ensuring high precision and dense retrieval increasing recall by capturing semantic similarities Wang et al. [2024], Arabzadeh et al. [2021]. Additionally, the system incorporates ChromaDB's cosine similarity search over embedded documents.

Hybrid retrieval generates a broad set of candidate passages but can include noise or partial matches. To refine results, we use **cross-encoder** models that jointly encode query-passage pairs, enabling detailed semantic interactions and highly accurate relevance scoring Rosa et al. [2022], Déjean et al. [2024]. In contrast, **bi-encoders** encode queries and documents separately for fast similarity computation but miss subtle token-level nuances. While bi-encoders scale well, their reranking precision is limited. Prioritizing accuracy over speed, our system first retrieves candidates via hybrid methods, then reranks them with cross-encoders to capture lexical and semantic relevance.

By extracting detailed aspect-sentiment pairs Yang et al. [2023], our system informs the downstream language model of both specific user concerns and their emotional context. Prior research Jayakody et al. [2024], Alturayeif et al. [2023] confirms that ABSA significantly enhances interpretability and actionable insights compared to global sentiment classification. Incorporating ABSA allows our pipeline to generate empathetic responses that address particular features and reflect user sentiment.

3.3 Prompt Construction and Response Generation with LLM

The pipeline generates responses using the GPT-OSS-120B Agarwal et al. [2025]. This model, with 117 billion parameters, offers powerful reasoning capabilities and supports flexible prompting with configurable temperature, top-p, and reasoning effort parameters. In this system, prompt engineering is critical to guide the model toward generating factually grounded, empathetic, and culturally aligned replies. Poor prompt design can cause generic or hallucinated outputs, while well-structured prompts mitigate these risks Sahoo et al. [2024], Amatriain [2024]. We employ role-play prompting, instructing the LLM to respond as a respectful app developer, starting and ending with Islamic greetings, expressing gratitude, and demonstrating humility Kong et al. [2023]. The prompt integrates three key elements: relevant retrieved passages ensure factual grounding; aspect-level sentiment context; and explicit cultural instructions align responses with Islamic etiquette. GPT-OSS-120B, provided by Cerebras AI, is configured with temperature 0.7, top-p 0.9, and medium reasoning effort, balancing creativity and precision. It generates responses within 2.5 seconds for each review on average, allowing nuanced, context-aware response generation that respects cultural norms and user emotions.

4 Experiments and Results

Table 1: Agreement between Human-as-a-judge and LLM-as-a-judge on system-generated responses.

Category	Cohen's Kappa	Krippendorff's Alpha	Human Scoring ($\mu \pm \sigma$)	LLM Scoring ($\mu \pm \sigma$)
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 2: System-generated responses *vs.* LLM-generated responses (Human Evaluation). Includes normalized scores, improvement percentage, and Wilcoxon signed-rank test p-values.

Category	System ($\mu \pm \sigma$)	LLM ($\mu \pm \sigma$)	Norm. System Mean (μ_{System})	Norm. LLM Mean (μ_{LLM})	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}

Initially, we conduct a partial evaluation on the review dataset provided by Greentech Apps Foundation, with the long-term target of extending the evaluation to 20%-30% of the dataset. From this dataset of reviews, 100 reviews are scored by 12 scorers (Muslim Students) across four criteria: Accuracy (factual, etiquette, and cultural correctness), Grammatical Correctness, Relevancy (addressing review content), and Application Specificity (Islamic app relevance), following analogous human evaluation Azov et al. [2024], Bhaskar et al. [2022]. Inter-rater reliability ranged from moderate to good, with Krippendorff's Alpha of Accuracy (0.68), Grammatical Correctness (0.54), Relevancy (0.57), and Specificity (0.86), confirming reliable human evaluation. Human *vs.* LLM judge comparison (Table 1) reveals poor agreement between 12 human evaluators and LLM scoring, with Cohen's Kappa (-0.09 to 0.06) and negative Krippendorff's Alpha values (-0.29 to -0.06). It is evident that current Large Language Models (LLMs) lack the capacity to adequately comprehend and represent

Islamic etiquette and cultural norms, and our results present the same. This finding aligns with recent research showing limitations of automated evaluation for culturally sensitive content Kim et al. [2024]. Performance comparisons (Table 2) through human evaluation show our culturally-aligned system significantly outperforms the LLM baseline across all criteria, achieving notable improvements in Specificity (Islamic app relevance) ($+30.39\%$) and Overall performance ($+9.90\%$). Direct comparison reveals that our system generates superior responses in 40.0% of cases. This demonstrates the effectiveness of agentic chunking, hybrid RAG, and prompt engineering for culturally sensitive response generation. Statistical significance was assessed using the Wilcoxon signed-rank test: improvements are significant for *accuracy* ($p = 0.0004$), *relevancy* ($p = 0.0417$) and especially *application specificity* ($p = 8 \times 10^{-9}$), while *grammatical correctness* shows only a negligible change between the two systems ($p = 0.4533$). This indicates that both systems are grammatically correct, but ADAB produces more specific, relevant, and accurate responses aligned with the application context. The overall improvement across all criteria is also significant ($p = 1.22 \times 10^{-11}$), confirming that ADAB consistently outperforms the baseline.

5 Discussion

This work's primary contribution lies in the creative integration and practical application of established advanced techniques such as RAG, ABSA, and etiquette-aware prompt engineering specifically tailored for culturally aligned response generation in Islamic app review contexts. We acknowledge that while no fundamentally new algorithm is proposed, the novelty resides in the thoughtful combination of these methods to address the complex challenges, such as context loss in chunking, retrieval of irrelevant documents, hallucinations, and cultural context sensitivity in AI-generated responses. Regarding dataset considerations, although our dataset consists of nearly $74{,}000$ reviews, the human evaluation was initially limited to 100 reviews to ensure the quality and feasibility of in-depth assessment. The annotators were Muslim university students familiar with Islamic etiquette and cultural sensitivities; they were not religious scholars, as the evaluation focused on general manners and respectful communication rather than jurisprudential or theological issues. Our evaluation confirms the hypothesis that an LLM as a judge cannot reliably capture nuanced cultural and etiquette requirements integral to Islamic app review responses. The significant inconsistencies observed between human and LLM scores reinforce the critical role of culturally informed human evaluators in validating AI output quality. To enhance transparency and accessibility for readers, some direct qualitative comparisons between ADAB-generated and Baseline LLM responses are provided in Appendix C.

6 Limitations & Future Works

Future efforts will focus on expanding the human evaluation to cover a larger portion (20%–30%) of the nearly $74{,}000$ review dataset. Adoption and generalization of this framework for multilingual support and broader cultural contexts, including other religions and diverse user groups, can enhance its global applicability. Improving automated evaluation methods to better capture cultural nuances remains a priority. Lastly, presenting more qualitative examples and increasing the diversity of raters will further strengthen the framework's robustness and transparency.

7 Conclusion

This research introduced ADAB, a culturally aligned response generation system tailored for Islamic app reviews. It successfully outperforms baseline LLMs in producing respectful, context-aware, and culturally aligned replies as verified by reliable human evaluation. The findings highlight that embedding cultural values not only improves response quality but also strengthens engagement with the app and its users. Ultimately, the study emphasizes that successful human–AI interaction depends on systems that are context-aware, empathetic, and aligned with community values. Future directions include extending ADAB to multilingual contexts and other faith-based or culturally nuanced domains, advancing toward a broader vision of value-aligned AI that engages users not only intelligently but also respectfully.

References

S. Agarwal, L. Ahmad, J. Ai, S. Altman, A. Applebaum, E. Arbus, R. K. Arora, Y. Bai, B. Baker, H. Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

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)*, pages 65–72. SBC, 2024.

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*, 30 (2):30, 2023.

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

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*, pages 2862–2866, 2021.

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.

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

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)*, pages 120–125. IEEE, 2021.

T. Chen, H. Wang, S. Chen, W. Yu, K. Ma, X. Zhao, H. Zhang, and D. Yu. Dense x retrieval: What retrieval granularity should we use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15159–15177, 2024a.

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*, pages 11908–11922, 2024b.

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

M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024.

C. Gao, J. Zeng, X. Xia, D. Lo, M. R. Lyu, and I. King. Automating app review response generation. In *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 163–175. IEEE, 2019.

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)*, 31(1):1–36, 2021.

Y. Gao and J. Smith. Review generation with large language models: Techniques and challenges. In *Proceedings of the 2023 Conference on Natural Language Processing*, pages 156–165, 2023.

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

J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

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*, 23: 1275–1312, 2018.

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

D. Jayakody, K. Isuranda, A. Malkith, N. De Silva, S. R. Ponnamperuma, G. Sandamali, and K. Sudheera. Aspect-based sentiment analysis techniques: A comparative study. In *2024 Moratuwa Engineering Research Conference (MERCon)*, pages 205–210. IEEE, 2024.

D. Katsiuba, T. Kew, M. Dolata, and G. Schwabe. Supporting online customer feedback management with automatic review response generation. 2022.

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)*, pages 121–133, 2022.

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*, pages 1–21, 2024.

A. Kong, S. Zhao, H. Chen, Q. Li, Y. Qin, R. Sun, X. Zhou, E. Wang, and X. Dong. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*, 2023.

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

V. Ramesh and S. Kumar. Advancements in large language models for automated review responses. *Journal of Artificial Intelligence Research*, 73:112–130, 2025.

G. Rosa, L. Bonifacio, V. Jeronymo, H. Abonizio, M. Fadaee, R. Lotufo, and R. Nogueira. In defense of cross-encoders for zero-shot retrieval, december 2022. *URL https://arxiv. org/abs/2212.06121*, 2022.

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.

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.

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*, 33:16857–16867, 2020.

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

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*, pages 548–554, 2024.

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*, pages 5117–5122, 2023.

W. Zhang, W. Gu, C. Gao, and M. R. Lyu. A transformer-based approach for improving app review response generation. *Software: Practice and Experience*, 53(2):438–454, 2023.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba. Large language models are human-level prompt engineers. In *The eleventh international conference on learning representations*, 2022.

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>

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.

C ADAB *vs.* Baseline LLM Review Response Examples

Review Example 1

App Review:
Very useful to me. Very greatfull to all admins and other stuff.

ADAB Generated Response:
Assalamu Alaikum wa Rahmatullahi wa Barakatuh. JazakAllahu Khair an kum for sharing your kind words and for appreciating the efforts of our admins. It warms our hearts to know that the app serves you well and that the support you receive brings benefit to your Qur'anic journey.

To keep enjoying the features you love, you may continue exploring the Quran Planner, audio recitations and topic searches and feel free to reach out to the admins. If you have any further suggestions or encounter any issue, kindly submit them at feedback.gtaf.org so we can serve you better. May Allah accept your gratitude and bless your continued use of the app. BarakAllahu feekum.

Baseline LLM Response:
Thank you so much for your kind words! We're delighted to hear that you find the app useful and that our admin team has been helpful. Your gratitude encourages us to keep improving the app and its content for everyone. If you have any suggestions, questions, or need assistance, please don't hesitate to reach out to us through the "Help & Feedback" section. May Allah bless you and grant you success in all your endeavors. JazakAllahu Khair! — The Support Team

Review Example 2

App Review:
I never uninstalled this app. This app is advertisement free. I can easily see the prayer times, sehri and iftar time from any location. This app has many other useful features like qibla compass, many important Dua, Al Quran and many more.

ADAB Generated Response:
Assalamu Alaikum wa Rahmatullahi wa Barakatuh. JazakAllahu Khair for sharing your wonderful experience with the app; it fills our hearts with gratitude to hear that you find the ad-free design, accurate prayer times, Sehri-Iftar schedule, Qibla compass, Dua collection and Quran features beneficial for your daily Ibadah.
To make the most of the app, you may try the Quran Planner: tap the "Planner" icon, set a realistic daily reading goal (*e.g.*, 2 pages or 5 minutes), enable a reminder at a convenient time and watch your progress grow. If you have ideas for new tools or wish to report any detail, please submit them at `feedback.gtaf.org` so we can continue serving the Ummah. May Allah accept your good deeds and keep you steadfast on the path of knowledge. BarakAllahu feekum.

Baseline LLM Response:
Thank you so much for your kind words and for supporting our ad-free app! We're delighted to hear that the prayer times, Sehri & Iftar schedules, Qibla compass, Duas, Qur'an and other features are helping you in your daily worship. If you ever have any questions, suggestions, or need assistance with anything, please don't hesitate to reach out. May Allah bless you and keep you steadfast on His path.

Review Example 3

App Review:
This is the best Quranic program I have ever worked with.

ADAB Generated Response:
Assalamu Alaikum wa Rahmatullahi wa Barakatuh. JazakAllahu Khair for sharing your wonderful experience; it brings great joy to hear that you consider this program the best Quranic companion you have worked with. Your kind words encourage us to continue serving the Ummah with humility and dedication.

We invite you to explore additional tools such as the Quran Planner, audio recitations and the word-by-word analysis to deepen your connection with the Qur'an and to let us know any ideas you may have by submitting feedback at feedback.gtaf.org. May Allah accept your good deeds and increase you in knowledge. BarakAllahu feekum.

Baseline LLM Response:
Assalamu Alaikum wa Rahmatullāh, Thank you so much for your wonderful feedback! We're thrilled to hear that you consider our app the best Quranic program you've worked with. Your kind words motivate us to keep improving and adding features that support your study and love of the Quran.
If you have any suggestions, questions, or need assistance, please don't hesitate to let us know—we're here to help. May Allah bless you and make your recitation and reflection even more rewarding. JazakAllāhu khair!

D Agentic Chunking of Documents

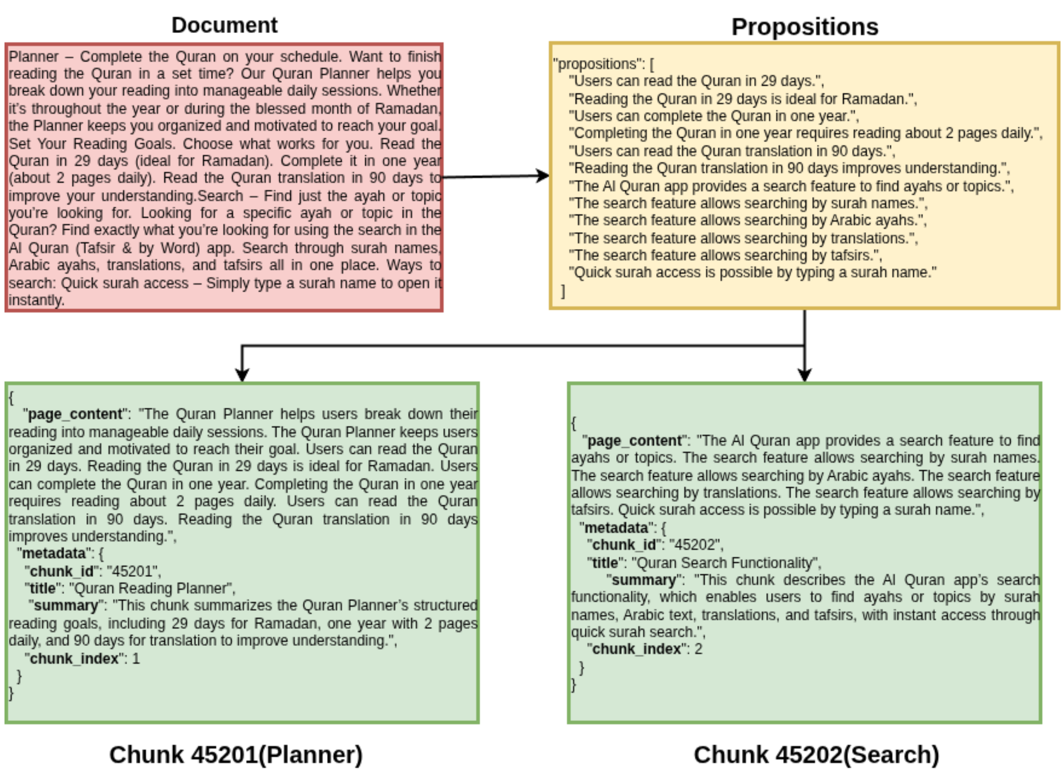

Figure 2: Agentic Chunking of Documents

E Rerenking with cross-encoder

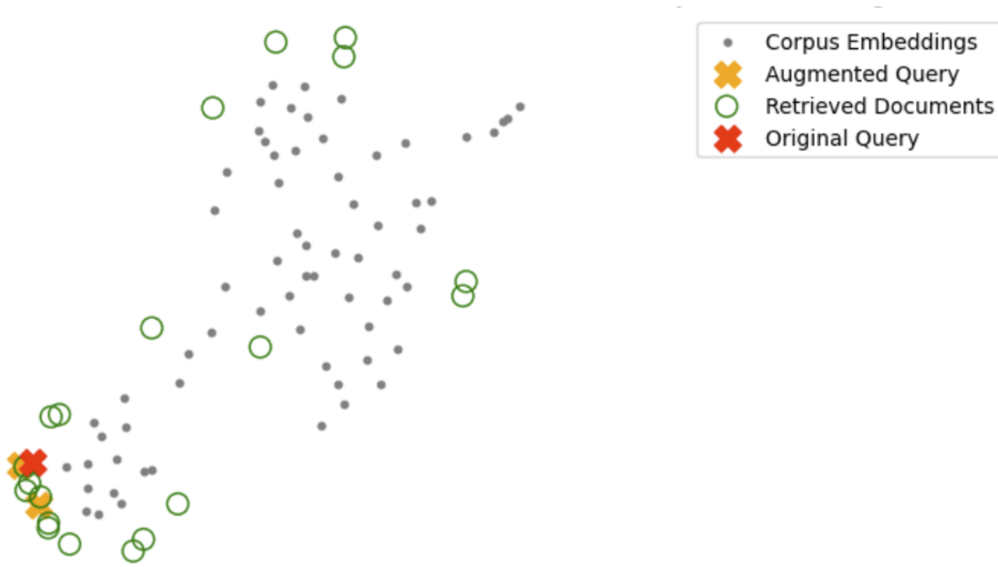

Figure 3: Projection of chunks without reranker. Irrelevant or loosely related documents are retrieved alongside relevant ones.

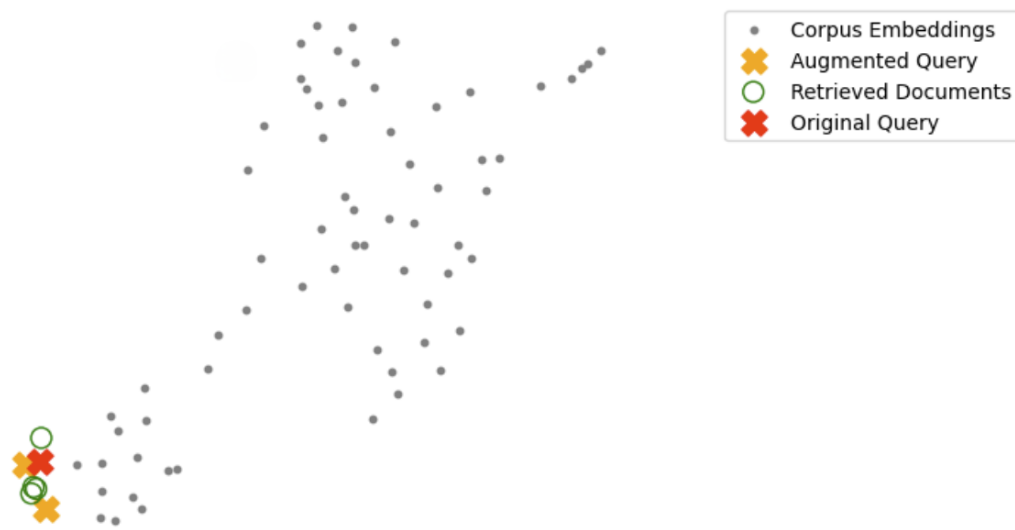

Figure 4: Projection of chunks with reranker. Relevant passages cluster closer to the original and expanded queries.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately describe the RAG-based pipeline for Islamic app review responses with 9.90% overall improvement and 40% direct comparison advantage, which matches the experimental results in Table 2

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Future work section 6 consists of the limitations and the future works.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (*e.g.*, independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This work presents an empirical system combining existing techniques (Hybrid RAG, ABSA, prompt engineering) without novel theoretical contributions requiring formal proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides detailed methodology section 3, model specifications (GPT-OSS-120B with temperature 0.7, top-p 0.9), evaluation setup (100 reviews, 12 scorers, 4 criteria), and dataset/knowledge base sources (18 Islamic websites, Greentech Apps Foundation's 74,000 review dataset).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper mentions providing an open-source Islamic app review dataset from Greentech Apps Foundation and describes implementation details for the hybrid RAG pipeline in proposed methodology section 3.1 and 3.2.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Experiements and Results section 4 specifies evaluation with 100 reviews, 12 Muslim student scorers, 4 evaluation criteria (1-5 scale), inter-rater reliability testing with 3 raters on 10 reviews and model configurations.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 1 and Table 2 report standard deviations, Fleiss' Kappa (0.47-0.79), Krippendorff's Alpha (0.54-0.86) and Cohen's Kappa values providing statistical reliability measures.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The Proposed Methodlogy section 3.3 part includes computer resources and time of execution(2.5s) for each review. Memory and compute workers are not relevant here as it is using a Cerebras AI provider resource to access GPT-OSS-120B through API to generate the responses.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research promotes cultural sensitivity, human-centered AI and respectful engagement with Muslim communities, aligning with ethical AI development principles.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses positive impacts of cultural alignment in human centered AI. However, It's biased for Muslims but the same framework could help other religious user groups in the same way and this is included in the future works & conclusion section 6 of this research.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The system generates culturally appropriate review responses and doesn't present high-risk applications requiring special safeguards beyond standard ethical AI practices.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper properly cites all models and datasets used: GPT-OSS-120B, MPNet, FAISS, ChromaDB, BM25S, PyABSA and Greentech Apps Foundation dataset with appropriate references.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces the first open-source Islamic app review dataset, documenting its sources (18 Islamic websites, Greentech Apps Foundation blog) and evaluation methodology.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The paper mentions involvement 12 Muslim student evaluators. Also the appendix B part describes all the instructions provided to the human evaluators.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The human evaluation involved rating system-generated responses without collecting personal data or involving significant risks that would typically require IRB approval.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are central to the methodology: GPT-OSS-120B for response generation, LLM-based agentic chunking, query expansion and prompt engineering techniques are thoroughly described in section 3

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.