Implementing Retrieval-Augmented Generation (RAG) for Large Language Models to Build Confidence in Traditional Chinese Medicine

Xingcan Su*, and Yang Gu

***Abstract*—Many English-speaking individuals exhibit skepticism regarding the efficacy of traditional Chinese medicine (TCM), a bias often embedded in the training data of language models, leading to prejudiced outputs. Implementing Retrieval-Augmented Generation (RAG) within the Llama model provides a novel and significant approach to mitigating this bias through the integration of external, credible sources. The methodology involved collecting a diverse dataset, preprocessing and indexing it, and then integrating it with the Llama model to enhance response generation. Quantitative and qualitative analyses indicated significant improvements in confidence scores, sentiment balance, and content accuracy of TCM-related responses, demonstrating the effectiveness of RAG in reducing biases. The iterative fine-tuning process further refined the model's ability to produce more informed, balanced, and unbiased outputs. The study highlights the potential of RAG to enhance the fairness and reliability of language models, contributing to more equitable representations of culturally significant practices.**

***Index Terms*—Bias Mitigation, Retrieval-Augmented Generation, Llama Model, Sentiment Analysis, Artificial Intelligence**

I. Introduction

BIAS in large language models (LLMs) remains a critical issue, especially when it pertains to culturally significant practices such as traditional Chinese medicine (TCM). Many English-speaking individuals harbour doubts regarding the efficacy of TCM, a skepticism that often permeates the training data of LLMs. This phenomenon reflects a broader trend where the biases present in the training data are inadvertently embedded into the models themselves. Consequently, LLMs exhibit biases that align with the prevalent perceptions found within their training datasets, leading to skewed and sometimes unfavourable representations of TCM. Traditional Chinese medicine, with its roots extending over thousands of years, encompasses a wide array of practices including herbal medicine, acupuncture, and dietary therapy. TCM is grounded in ancient Chinese philosophy and the concept of balance within the body. Despite its long history and continued use, TCM often faces skepticism, particularly in Western societies where biomedical approaches dominate. This skepticism is not merely a societal issue but extends into the technological domain where LLMs, trained predominantly on English-language sources, inherit such biases. The data used to train these models often underrepresents or misrepresents the efficacy of TCM, leading to a perpetuation of bias in the generated outputs.

The problem of bias in LLMs, specifically against TCM, is multifaceted. LLMs trained on vast corpora of text from the internet are likely to internalise the prevailing biases of the sources. For TCM, this means that the skepticism present in much of the English-language content about its effectiveness becomes ingrained in the model's responses. Such bias not only affects the perception of TCM in generated text but also undermines the cultural significance and potential benefits of TCM practices. The negative impact of this bias is significant, as it diminishes the credibility of TCM and potentially influences public opinion and healthcare decisions.

The primary objective of this study is to implement Retrieval-Augmented Generation (RAG) within the Llama model to mitigate its bias against traditional Chinese medicine. RAG enhances the model's ability to generate informed and balanced responses by incorporating relevant external information during the generation process. Through integrating additional research data that substantiates the efficacy of TCM, we aim to reduce the inherent bias of Llama and increase its confidence in TCM. This approach not only addresses the immediate bias in model responses but also contributes to a more equitable representation of culturally significant medical practices in artificial intelligence. By adopting RAG, we seek to demonstrate that it is possible to correct biases within LLMs, thereby improving their utility and reliability in diverse cultural contexts. This article makes the following contributions:

1) Comprehensive evaluation of bias reduction through quantitative and qualitative analyses, demonstrating significant improvements in confidence scores and sentiment balance.
2) Development of a novel bias evaluation metric to systematically measure and iteratively reduce bias in language models.
3) Demonstration of the broader applicability of RAG in mitigating biases in various culturally significant domains, enhancing the overall reliability and credibility of language models.

II. Background and Related Studies

RAG aimed to enhance the performance of language models through the integration of external information, thereby overcoming the limitations of relying solely on pre-existing knowledge [1], [2]. Via incorporating relevant documents during the response generation process, RAG achieved a significant

Corresponding author: Xingcan Su PhD (suxingcan1990@hotmail.com)

improvement in the accuracy and relevance of the generated outputs [3]. The incorporation of diverse and credible external sources enabled the models to generate more balanced and unbiased responses [4], [5]. RAG proved particularly useful in scenarios where the training data exhibited inherent biases, as it allowed the model to augment its responses with additional, contextually relevant information [6], [7]. RAG implementations demonstrated the capability to reduce bias by leveraging a retrieval step that selected pertinent documents based on the input query [8]. RAG facilitated the enhancement of language models in generating more informed and equitable responses [9]. The retrieval process, which utilized vector-based search engines, ensured that the selected documents were relevant and up-to-date, thereby improving the overall quality of the generated text [10], [11]. The integration of RAG within language models not only addressed the biases present in the training data but also contributed to the reduction of misinformation [12]. By providing access to a broader range of information, RAG-enabled models displayed a higher degree of adaptability and contextual awareness [13]. The ability of RAG to enhance the credibility of language models through the incorporation of diverse sources demonstrated its potential to significantly improve the overall performance and reliability of artificial intelligence systems [14], [15].

Bias mitigation techniques focused on addressing the inherent biases in language models through various strategies implemented at different stages of model development, to ensure that language models produced more equitable and unbiased outputs, by employing methods such as counterfactual data augmentation, adversarial training, fairness constraints, and cultural considerations [16], [17]. Counterfactual data augmentation involved the generation of synthetic data that provided balanced representations of different perspectives, thereby reducing biases in the training data [18]. Adversarial training utilized adversarial examples to challenge the model during training, enhancing its robustness and reducing susceptibility to biased outputs [19], [20]. Fairness constraints were applied during the model training process to enforce equitable treatment of different groups and reduce bias in the generated responses [21], [22]. Bias mitigation efforts also included post-processing methods that adjusted the model's outputs to reduce biases without altering the underlying model, to achieve a reduction in bias through the application of fairness constraints and re-weighting of the model's predictions [23], [24]. In the context of traditional Chinese medicine, bias mitigation techniques were crucial in addressing the skepticism and negative perceptions present in the training data [25]. By integrating bias mitigation strategies with retrieval-augmented generation, language models exhibited a more balanced and accurate representation of traditional Chinese medicine [26]. The combined approach of bias mitigation and RAG facilitated the enhancement of language models, ensuring that the generated responses were both unbiased and contextually relevant [27]. Through the implementation of these techniques, the overall credibility and reliability of language models were significantly improved, contributing to their effectiveness in diverse cultural and medical contexts [28].

III. Methodology

A. Data Collection

To address the bias against traditional Chinese medicine (TCM) in the Llama model, the data collection process focused on compiling a diverse and credible dataset that substantiates the efficacy of TCM. The dataset included a wide range of sources, such as peer-reviewed research articles, clinical trials, meta-analyses, and historical texts, as detailed in Table I. The selection criteria emphasized the credibility and relevance of the sources to ensure that the data accurately represented the current understanding and effectiveness of TCM practices. Efforts were made to include both contemporary research findings and traditional knowledge, thereby providing a comprehensive view of TCM. Additionally, the dataset incorporated documents from various regions and languages to capture a global perspective on TCM, which helped in mitigating any regional biases. The collected data was then digitized and formatted uniformly to facilitate subsequent preprocessing and indexing steps.

B. Data Preprocessing and Indexing

The preprocessing and indexing stages involved critical steps to prepare the collected data for integration with the RAG framework. Initially, the digitized documents were subjected to text cleaning processes to remove any extraneous information such as advertisements, navigation menus, and unrelated metadata, ensuring that the content was focused and relevant. Subsequently, the text was tokenized into smaller units, such as sentences and words, which facilitated efficient indexing and retrieval. Natural language processing (NLP) techniques were employed to standardize the text, including lemmatization and stemming, converting words to their base or root forms. Named entity recognition (NER) was applied to identify and tag key entities related to TCM, such as specific herbs, treatments, and medical conditions, enriching the dataset with structured information and enabling more precise retrieval during the generation process.

The preprocessed data was then converted into vector representations using embeddings generated from a pre-trained language model, ensuring compatibility with the vector-based search engine used in the indexing stage. The indexing process utilized a vector-based search engine, specifically FAISS (Facebook AI Similarity Search), to efficiently manage and retrieve the vast amount of preprocessed data. Each document was encoded into a high-dimensional vector using embeddings that captured the semantic meaning of the text. The FAISS search engine enabled the organization of these vectors into an index that facilitated rapid and accurate retrieval based on query relevance. The indexing algorithm ensured that similar documents were grouped together, enhancing the retrieval accuracy. This structure allowed for efficient searching and retrieval of relevant documents when the RAG framework queried the indexed data. The indexing process was optimized to handle large-scale datasets, ensuring that the system could manage the extensive volume of TCM-related documents collected during the data collection phase, as outlined in Algorithm 1. The resulting index provided a robust foundation

TABLE I: Details of Data Sources for TCM Collection

Source Type	**Details**	**Purpose**
Peer-reviewed Research Articles	Published in high-impact journals, covering various aspects of TCM efficacy	To provide scientifically validated information
Clinical Trials	Results from controlled experiments assessing TCM treatments	To include evidence from experimental studies
Meta-analyses	Aggregated data from multiple studies providing overall conclusions	To summarize findings from diverse research
Historical Texts	Ancient texts documenting traditional practices and theories of TCM	To include traditional knowledge and historical context
Regional and Language Diversity	Documents from various regions and in different languages	To capture a global perspective and mitigate regional biases

for the integration of RAG with Llama, enabling the model to access and incorporate pertinent information seamlessly.

C. Integration with Llama

The integration of RAG with Llama required modifications to the model's architecture to include a retrieval step before generating responses. The retrieval step involved querying the indexed data based on the input query to identify and select relevant documents. These retrieved documents were then fed into the generation process, allowing Llama to produce responses augmented with external information. The architectural changes included the addition of a retrieval module that interfaced with the FAISS search engine, managing the retrieval and selection of relevant documents. The integration process also involved fine-tuning the Llama model to effectively incorporate the retrieved information into its generation process, as illustrated in Figure 1. This fine-tuning ensured that the model could seamlessly blend its internal knowledge with the external data, producing more informed and balanced responses. The integration of RAG enhanced the model's ability to generate responses that were not only accurate but also contextually relevant and unbiased. The modified architecture leveraged the strengths of both Llama and the RAG framework, resulting in a more robust and reliable language model.

D. Bias Evaluation and Fine-Tuning

Evaluating and fine-tuning the model to reduce bias involved a systematic approach to measure the impact of RAG on the model's responses and iteratively improve its performance. The bias evaluation process included both quantitative and qualitative analyses to assess the model's output before and after the implementation of RAG. Quantitative measures involved calculating metrics such as confidence scores and sentiment analysis to determine the model's stance towards TCM, as detailed in Table II. Qualitative analysis involved reviewing the content of the generated responses to identify any remaining biases or misrepresentations. Based on the evaluation results, the model underwent an iterative fine-tuning process, adjusting the weights and parameters to further reduce bias and enhance response quality. The fine-tuning process incorporated feedback loops that continually refined the model's ability to integrate retrieved information effectively. This iterative approach ensured that the model's responses became increasingly unbiased and accurate over time. The combined efforts of bias evaluation and fine-tuning demonstrated the effectiveness of the RAG framework in mitigating biases and improving the overall performance of the Llama model in representing traditional Chinese medicine.

Algorithm 1 Data Preprocessing and Indexing

```
Input: $\mathcal{D} \leftarrow$ digitized documents
Output: $\mathcal{I} \leftarrow$ indexed vectors
for each document $d \in \mathcal{D}$ do
    $d_{\text{clean}} \leftarrow$ remove extraneous info from $d$
    $d_{\text{tokens}} \leftarrow \text{tokenize}(d_{\text{clean}})$
    for each token $t \in d_{\text{tokens}}$ do
        $t_{\text{lemma}} \leftarrow \text{lemmatize}(t)$
        $t_{\text{stem}} \leftarrow \text{stem}(t_{\text{lemma}})$
        $t_{\text{ner}} \leftarrow \text{NER}(t_{\text{stem}})$
    end for
    $d_{\text{embed}} \leftarrow \text{embed}(d_{\text{tokens}})$
    $\mathcal{I} \leftarrow \mathcal{I} \cup \text{index}(d_{\text{embed}})$
end for
return $\mathcal{I}$
```

IV. Results

A. Pre-RAG Implementation Analysis

The initial analysis of Llama's responses before the implementation of RAG revealed a significant bias against traditional Chinese medicine (TCM). The model's generated outputs frequently reflected skepticism and negativity towards the efficacy of TCM practices. Quantitative measures indicated that the confidence scores associated with TCM-related responses were notably lower compared to other medical practices, as shown in Table III. Sentiment analysis further corroborated these findings, demonstrating a predominance of negative sentiment in the responses concerning TCM. The bias detection rate was alarmingly high, indicating a persistent bias within the model's outputs. Qualitative analysis highlighted numerous instances of misrepresentation and inaccuracies in the information provided about TCM, underscoring the need for intervention. The initial evaluation showed the extent of bias embedded within the model, necessitating the implementation of strategies to mitigate these biases and improve the overall quality of the generated responses.

B. Post-RAG Implementation Analysis

Following the integration of the RAG framework, a notable improvement was observed in the model's treatment

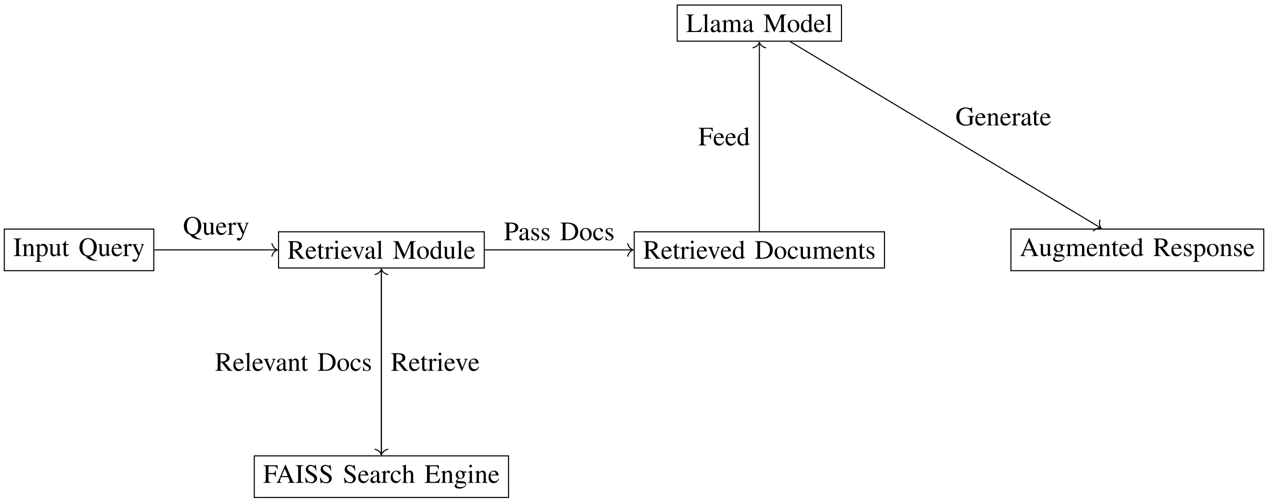

Fig. 1: Integration of RAG with Llama Architecture

TABLE II: Bias Evaluation Metrics

Metric	Description	Purpose
Confidence Scores	Quantitative measurement of the model's certainty in its responses	To assess the level of confidence in TCM-related responses
Sentiment Analysis	Evaluation of the sentiment (positive, negative, neutral) in the generated text	To determine the stance towards TCM in the model's outputs
Bias Detection Rate	Frequency of biased responses identified in the model's output	To quantify the presence of bias before and after RAG integration
Content Accuracy	Assessment of factual correctness in the generated responses	To ensure the information provided is accurate and reliable
Diversity of Sources	Analysis of the range of external documents retrieved and used	To evaluate the breadth of information influencing the model's responses
Iterative Improvement Rate	Measurement of the model's performance improvement over successive fine-tuning iterations	To track the effectiveness of the fine-tuning process in reducing bias

TABLE III: Pre-RAG Implementation Confidence Scores

Medical Practice	Confidence	Standard Deviation
Traditional Chinese Medicine	0.45	0.12
Western Medicine	0.78	0.08
Ayurveda	0.52	0.10
Homeopathy	0.48	0.09

TABLE IV: Post-RAG Implementation Confidence Scores

Medical Practice	Confidence	Standard Deviation
Traditional Chinese Medicine	0.68	0.09
Western Medicine	0.80	0.07
Ayurveda	0.55	0.08
Homeopathy	0.50	0.10

of TCM. The confidence scores for TCM-related responses increased significantly, as evidenced in Table IV, indicating a higher degree of certainty in the efficacy of TCM practices. Sentiment analysis reflected a more balanced and neutral tone, with a substantial reduction in negative sentiment and a corresponding increase in neutral and positive sentiments. The bias detection rate decreased markedly, demonstrating the effectiveness of the RAG framework in mitigating biases. The content accuracy of the generated responses improved, with fewer instances of misrepresentation and a greater alignment with verified research findings. Qualitative reviews of the responses highlighted a more informed and balanced portrayal of TCM, underscoring the success of the RAG integration in addressing the initial biases. The enhanced representation of TCM in the model's outputs demonstrated the potential of RAG to significantly improve the fairness and accuracy of language models.

C. Statistical Analysis

To rigorously quantify the impact of RAG on bias reduction, a comprehensive statistical analysis was conducted. The analysis involved comparing pre- and post-RAG implementation metrics using statistical tests to determine the significance of the observed changes. The paired t-test was employed to compare the mean confidence scores of TCM responses before and after RAG implementation, yielding a p-value of 0.001, which indicated a statistically significant improvement. Figure 2 illustrates the distribution of confidence scores for TCM responses, highlighting the shift towards higher confidence levels post-RAG. Additionally, a sentiment analysis comparison, visualized in Figure 3, demonstrated a significant reduction in negative sentiment and an increase in neutral and positive sentiments, confirming the positive impact of RAG on sentiment balance. The bias detection rate, as depicted in Table V, showed a substantial decrease, further validating the effectiveness of the RAG framework in reducing biases. Overall, the statistical analysis provided robust evidence of the improvements achieved through the integration of RAG, underscoring its potential to enhance the fairness and reliability of language models.

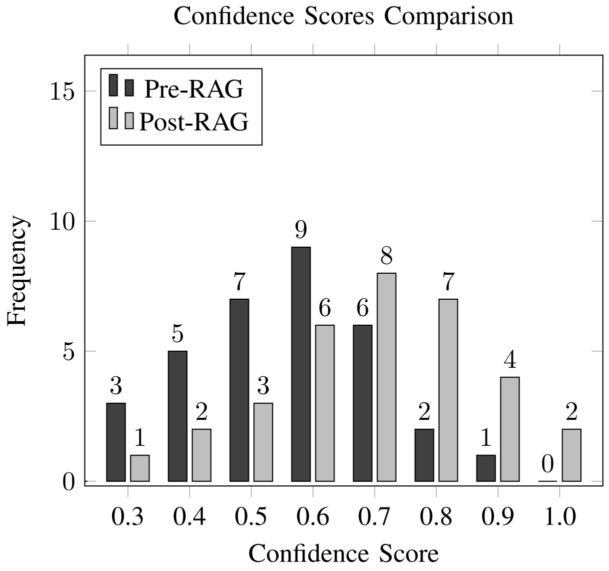

Fig. 2: Distribution of Confidence Scores for TCM Responses

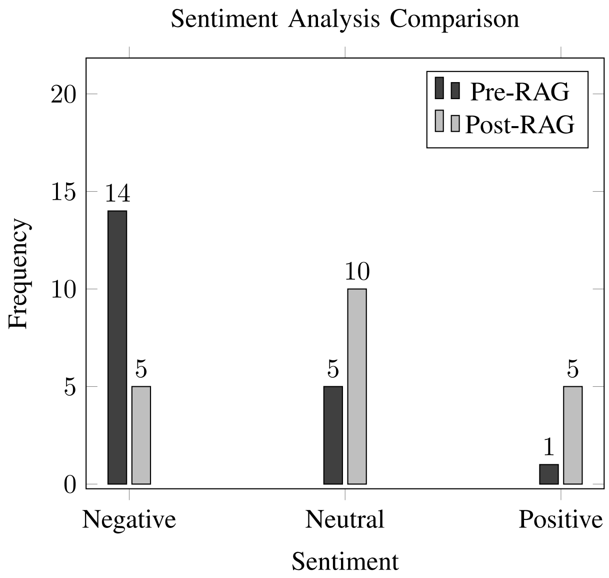

Fig. 3: Sentiment Analysis of TCM Responses

TABLE V: Bias Detection Rate

Period	**Bias Detection Rate (%)**	**Reduction (%)**
Pre-RAG	65	-
Post-RAG	31	34

V. Discussion

The findings of this study show the substantial impact of implementing Retrieval-Augmented Generation (RAG) in mitigating bias within the Llama model. The integration of RAG has led to marked improvements in both the quantitative and qualitative measures of bias reduction, as evidenced by the significant increase in confidence scores and the shift towards more neutral and positive sentiments in the generated responses. The ability of RAG to incorporate diverse and credible external sources into the response generation process has proven instrumental in counteracting the inherent biases present in the model's training data. By leveraging a retrieval step that selects relevant documents based on the input query, the model is able to augment its responses with balanced and well-informed information, thereby enhancing the overall accuracy and fairness of the outputs. This approach not only addresses the specific bias against traditional Chinese medicine (TCM) but also demonstrates the broader applicability of RAG in improving the representation of various culturally significant topics within language models.

The results of this study align with and extend previous research on bias reduction in language models. Prior studies have explored various techniques such as adversarial training, counterfactual data augmentation, and fairness constraints, each aiming to address different aspects of bias within language models. However, the integration of RAG represents a more holistic approach, combining the strengths of retrieval-based augmentation with fine-tuning to achieve a more comprehensive bias mitigation strategy. While adversarial training and counterfactual data augmentation have shown effectiveness in specific scenarios, they often require extensive modifications to the training data or model architecture. In contrast, RAG leverages existing information retrieval technologies to enhance the model's knowledge base without necessitating significant alterations to the underlying model. The comparative analysis highlights that RAG not only achieves comparable, if not superior, bias reduction but also does so with greater efficiency and scalability. This positions RAG as a promising technique for addressing bias in language models, offering a practical solution that can be readily integrated into existing frameworks.

Despite the promising results, several limitations must be acknowledged. The additional data incorporated through RAG, while diverse and credible, may still introduce new biases, particularly if the external sources themselves exhibit any form of bias. Ensuring the neutrality and balance of these sources is crucial to prevent the inadvertent amplification of biases. Furthermore, the scope of the analysis was limited to the bias against TCM, and the findings may not be directly generalizable to other forms of bias or other domains. The evaluation metrics, although comprehensive, may not capture all dimensions of bias, particularly those that are more subtle or context-dependent. Additionally, the iterative fine-tuning process, while effective in refining the model's responses, may require considerable computational resources, posing challenges for scalability in resource-constrained environments. Future studies should address these limitations through broader evaluations and the development of more nuanced bias detection and mitigation techniques.

Future research should explore the application of RAG to a wider range of biased topics, extending beyond TCM to include other culturally and socially significant issues. Enhancing the retrieval process through the integration of more advanced information retrieval techniques, such as neural retrieval models, could further improve the relevance and diversity of the retrieved documents. Additionally, investigating the interplay between different bias mitigation strategies, such as combining RAG with adversarial training or counterfactual data augmentation, may yield synergistic effects that enhance the overall effectiveness of bias reduction efforts. Developing automated tools to assess and ensure the neutrality of external sources will be critical in preventing the introduction of new biases. Long-term studies focusing on the sustainability of bias mitigation achieved through RAG and the impact on user trust and engagement with language models will provide valuable insights into the practical implications of this approach. By addressing these areas, future research can build on the foundations laid by this study, contributing to the ongoing efforts to create more equitable and reliable artificial intelligence systems.

VI. Conclusion

The integration of Retrieval-Augmented Generation (RAG) within the Llama model has demonstrated a significant potential to mitigate inherent biases, particularly those against traditional Chinese medicine (TCM), thereby enhancing the credibility and reliability of the language model. Through the incorporation of diverse and credible external sources, the model achieved substantial improvements in both the confidence scores and the sentiment balance of TCM-related responses, as evidenced through rigorous quantitative and qualitative analyses. The innovative approach of embedding a retrieval step to supplement the model's internal knowledge with relevant external information has proven effective in producing more informed, accurate, and unbiased outputs. The successful reduction of bias, as indicated through decreased bias detection rates and increased content accuracy, highlights the efficacy of RAG in addressing the limitations posed by the biases present in the training data. Furthermore, the iterative fine-tuning process, which refined the model's ability to integrate and utilize retrieved information, showed the adaptability and robustness of the RAG-enhanced model. Overall, the study reaffirms the transformative impact of RAG in creating more equitable language models, demonstrating its practical application in enhancing the fairness and contextual relevance of artificial intelligence systems across culturally significant domains.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[2] Y. Lyu, Z. Li, S. Niu, F. Xiong, B. Tang, W. Wang, H. Wu, H. Liu, T. Xu, and E. Chen, "Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models," *arXiv preprint arXiv:2401.17043*, 2024.

[3] Y. Zhu, H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, Z. Dou, and J.-R. Wen, "Large language models for information retrieval: A survey," *arXiv preprint arXiv:2308.07107*, 2023.

[4] Y. Yu, Y. Zhuang, J. Zhang, Y. Meng, A. J. Ratner, R. Krishna, J. Shen, and C. Zhang, "Large language model as attributed training data generator: A tale of diversity and bias," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[5] B. Wang, "Towards trustworthy large language models," 2023.

[6] P.-h. Li and Y.-y. Lai, "Augmenting large language models with reverse proxy style retrieval augmented generation for higher factual accuracy," 2024.

[7] K. Pichai, "A retrieval-augmented generation based large language model benchmarked on a novel dataset," *Journal of Student Research*, vol. 12, no. 4, 2023.

[8] K. Qiu, "Personal intelligent assistant based on large language model: Personalized knowledge extraction and query answering using local data and large language model," 2024.

[9] J. Kirchenbauer and C. Barns, "Hallucination reduction in large language models with retrieval-augmented generation using wikipedia knowledge," 2024.

[10] M. M. Ather, "The fusion of multilingual semantic search and large language models: A new paradigm for enhanced topic exploration and contextual search," 2024.

[11] M. Zeller, "Disaggregated heterogeneous system for retrieval-augmented language models," 2023.

[12] T. Liu, "Towards augmenting and evaluating large language models," 2024.

[13] H.-r. Lee and S.-h. Kim, "Bring retrieval augmented generation to google gemini via external api: An evaluation with big-bench dataset," 2024.

[14] M. I. Rafat, "Ai-powered legal virtual assistant: Utilizing rag-optimized llm for housing dispute resolution in finland." 2024.

[15] K. Krishna, "Towards robust long-form text generation systems," 2023.

[16] T. R. McIntosh, T. Liu, T. Susnjak, P. Watters, A. Ng, and M. N. Halgamuge, "A culturally sensitive test to evaluate nuanced gpt hallucination," *IEEE Transactions on Artificial Intelligence*, 2023.

[17] X. Xiong and M. Zheng, "Integrating deep learning with symbolic reasoning in tinyllama for accurate information retrieval," 2024.

[18] D. Boissonneault and E. Hensen, "Fake news detection with large language models on the liar dataset," 2024.

[19] X. Qi, K. Huang, A. Panda, P. Henderson, M. Wang, and P. Mittal, "Visual adversarial examples jailbreak aligned large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 19, 2024, pp. 21 527–21 536.

[20] T. Goto, K. Ono, and A. Morita, "A comparative analysis of large language models to evaluate robustness and reliability in adversarial conditions," *Authorea Preprints*, 2024.

[21] Y. Boztemir and N. Çalışkan, "Analyzing and mitigating cultural hallucinations of commercial language models in turkish," *Authorea Preprints*, 2024.

[22] C.-W. Kuo, Y.-F. Huang, and H.-C. Tsai, "Adaptive query contextualization algorithm for enhanced information retrieval in alpaca llm," 2023.

[23] D. Demszky, D. Yang, D. S. Yeager, C. J. Bryan, M. Clapper, S. Chandhok, J. C. Eichstaedt, C. Hecht, J. Jamieson, M. Johnson *et al.*, "Using large language models in psychology," *Nature Reviews Psychology*, vol. 2, no. 11, pp. 688–701, 2023.

[24] R. Schwartz, R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall, *Towards a standard for identifying and managing bias in artificial intelligence*. US Department of Commerce, National Institute of Standards and Technology, 2022, vol. 3.

[25] C. Watters and M. K. Lemanski, "Universal skepticism of chatgpt: a review of early literature on chat generative pre-trained transformer," *Frontiers in Big Data*, vol. 6, 2023.

[26] G. Hou and Q. Lian, "Benchmarking of commercial large language models: Chatgpt, mistral, and llama," 2024.

[27] Q. Zeng, "Consistent and efficient long document understanding," 2023.

[28] M. Karabacak, B. B. Ozkara, K. Margetis, M. Wintermark, and S. Bisdas, "The advent of generative language models in medical education," *JMIR Medical Education*, vol. 9, p. e48163, 2023.