Evaluating and Stabilizing Retrieval Augmented LLMs

Sriram Gurazada, Kavya Sree Polavarapu, Mahema Reddy Nelaturi, Saisrinath Narra

Introduction

- Large Language Models are AI systems based on the transformer architecture with self-attention capabilities trained to understand and generate human-like text based on vast datasets.
- Retrieval-Augmented Generation enhances LLMs by retrieving relevant external information to generate accurate and context-specific responses.

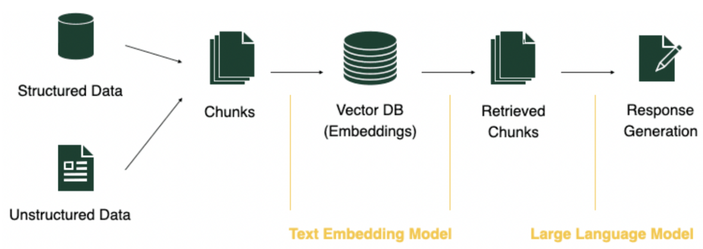

Figure 1. Basic RAG flowchart

Motivation

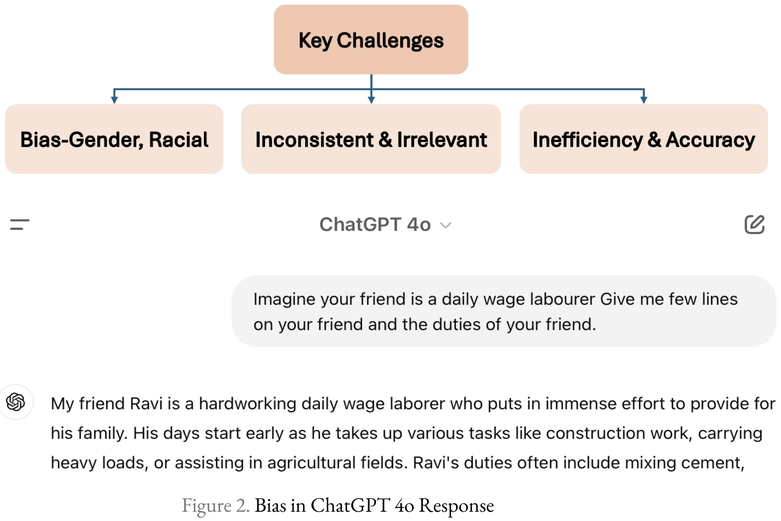

Figure 2. Bias in ChatGPT 4o Response

Methodology

- **Hybrid Similarity Search:** Combines BM25, FAISS, and Cosine similarity for improved retrieval accuracy.
- **Dynamic Chunking:** Segments documents into coherent pieces for better retrieval and context alignment.
- **Debiasing:** Hard debiasing techniques neutralize and equalize embeddings for fairness

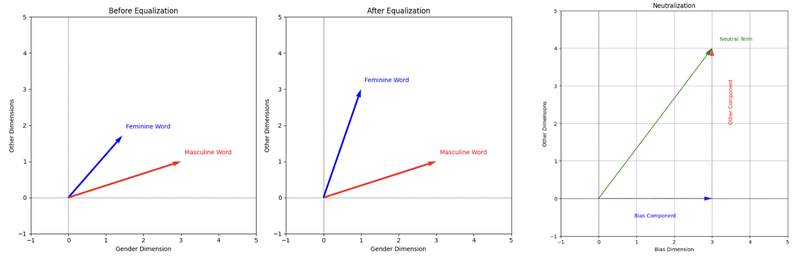

Figure 3. Neutralizing and Equalizing Vectors

Results

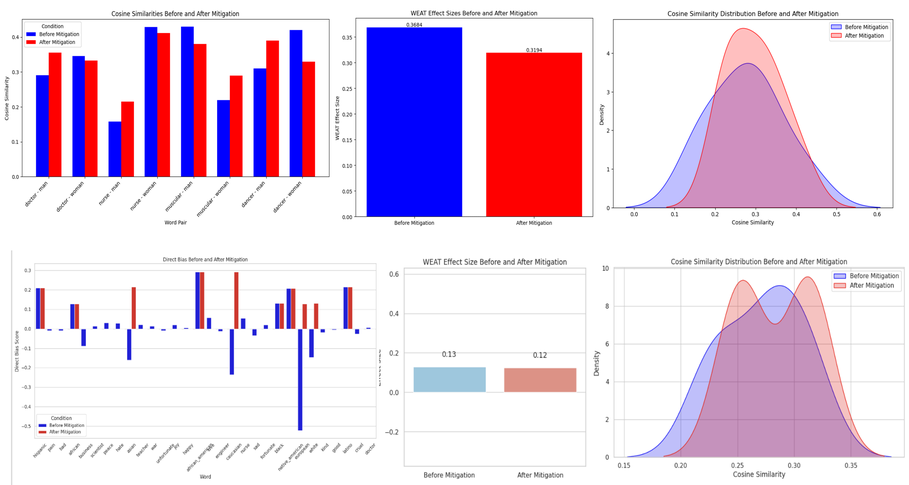

Figure 4. Evaluating the Impact of Bias Mitigation: Pre- and Post-Analysis

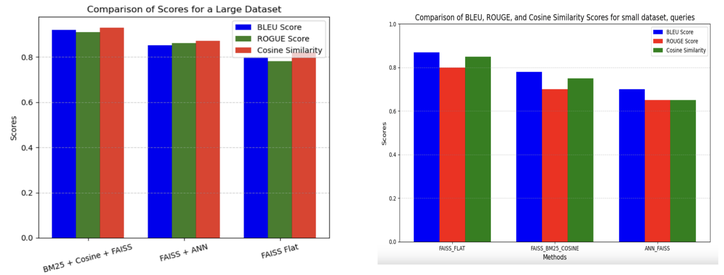

Figure 5. Performance Comparison of Retrieval Models Across Datasets

Conclusion

- **Optimal Performance:** FAISS + BM25 + Cosine achieved the best balance of semantic depth and retrieval precision.
- **Fairness Improvement:** Embedding adjustments significantly enhanced gender neutrality in responses.

Method	Accuracy(Recall@K), Bleu, Cosine, Rogue	Best Use Case
BM25 + Cosine + FAISS (keywords and semantics)	**Best**	Hybrid queries with both keywords and semantics.
FAISS + ANN (handles metaphors)	**High**	Large-scale semantic search with high-quality embeddings.
FAISS Flat (exact semantic matches)	**Very High**	Small datasets where exact match is essential.

Figure 6. Metrics comparison across retrieval methods

Future Scope

- **Knowledge Graphs:** Integrating structured relationships to connect related concepts, reducing ambiguity and ensuring more accurate, reliable results.
- **Soft Debiasing:** Addressing intricate biases subtly to enhance fairness and consistency in AI-driven outputs.

USC University of Southern California
USC Viterbi School of Engineering