Evaluating and Stabilizing Retrieval Augmented LLMs

TA: Jesse Zhang

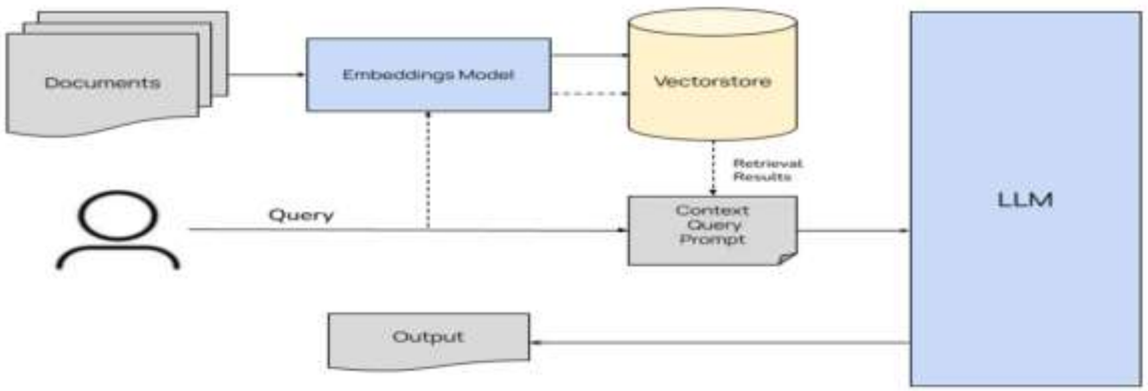
Query Craft

Sriram Gurazada   Kavya Sree Polavarapu   Mahema Reddy Nelaturi   Saisrinath Narra

Abstract:

- Focus: Enhancing consistency, relevance, debiasing in Retrieval-Augmented LLMs.
- Techniques: debiasing for unbiased outputs, dynamic chunking, and advanced hybrid similarity searches
- Validation: Metrics like BLEU, ROUGE-L, and WEAT ensure robust performance.

Challenges and Objectives



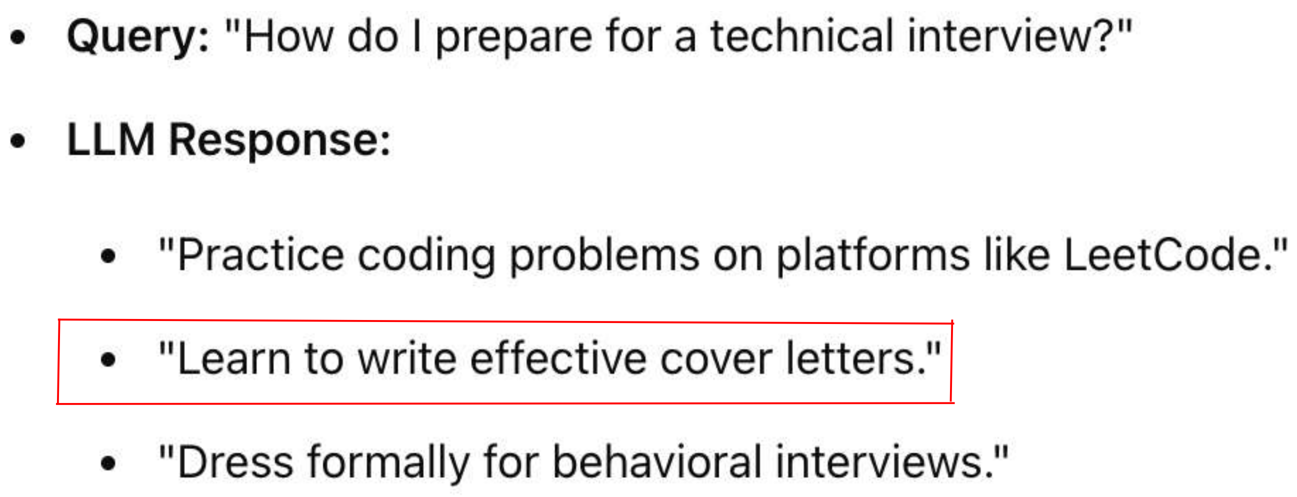
RAG

Bias

HE is a Nurse
SHE is a Nurse

HE is muscular
SHE is muscular

Inconsistent and irrelevant



- **Query:** "How do I prepare for a technical interview?"
- **LLM Response:**
  - "Practice coding problems on platforms like LeetCode."
  - "Learn to write effective cover letters."
  - "Dress formally for behavioral interviews."

Inefficiency and accuracy

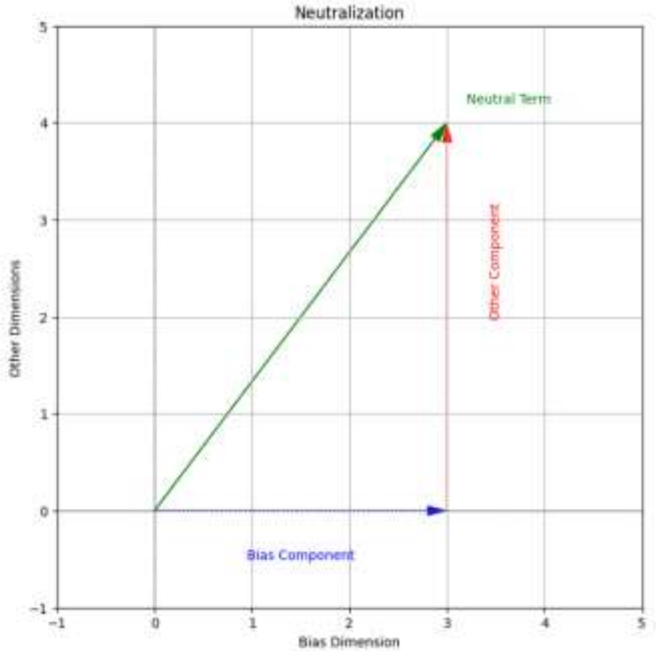
“mitochondria is the power”

“House of the cell”

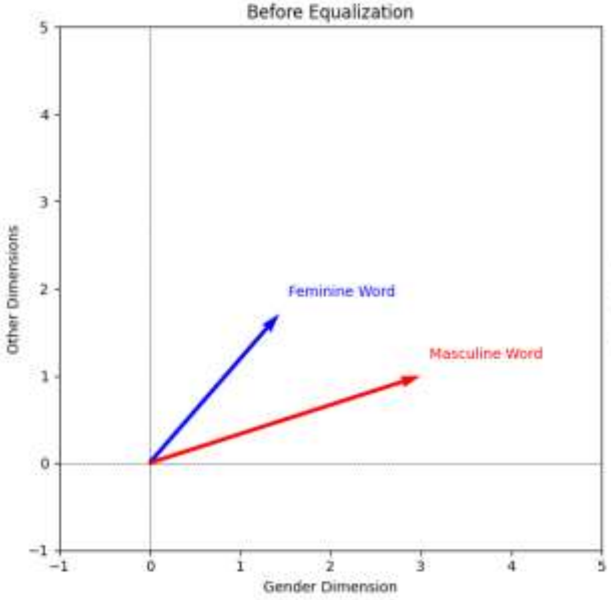
Methodology

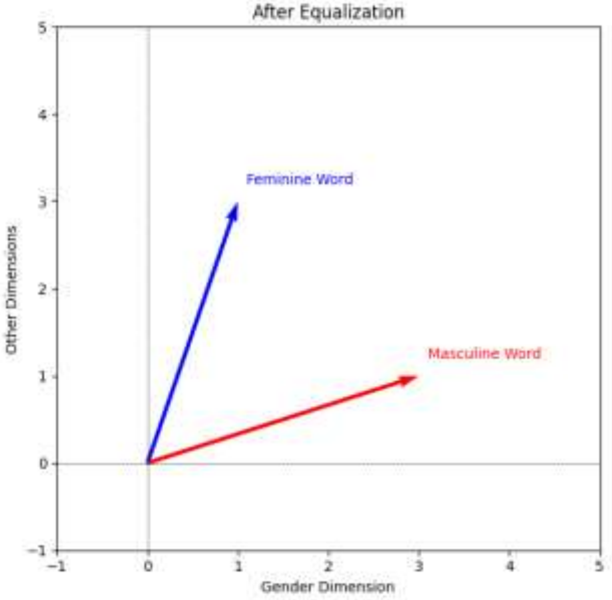
Approach:

- **Context Retrieval:**
  - Hybrid Similarity Searches using BM25, Cosine, FAISS, ANN
  - Dynamic Chunking, Semantic Chunking



- **Bias Mitigation:**
  - Neutralization of neutral embeddings
  - Equalization of vector embeddings.



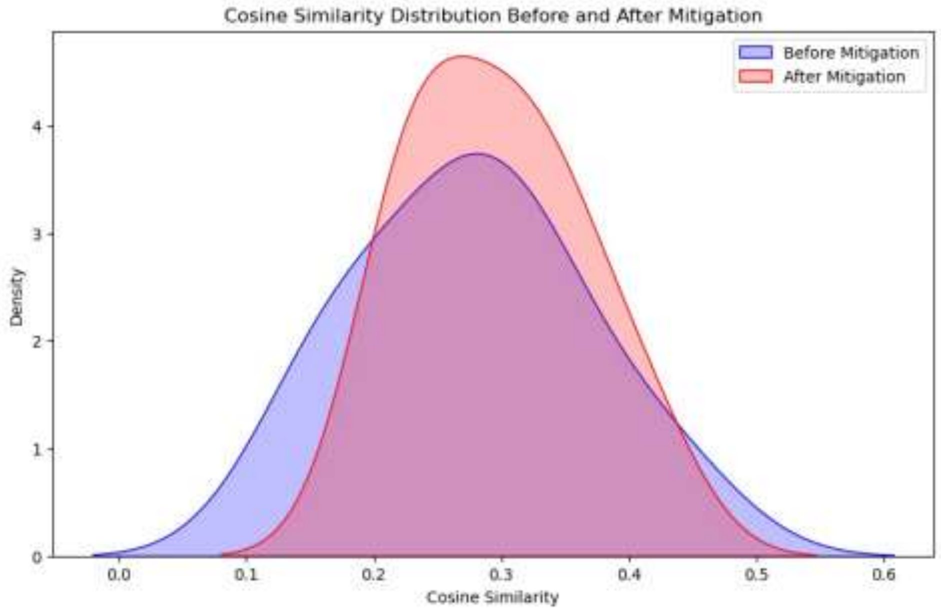


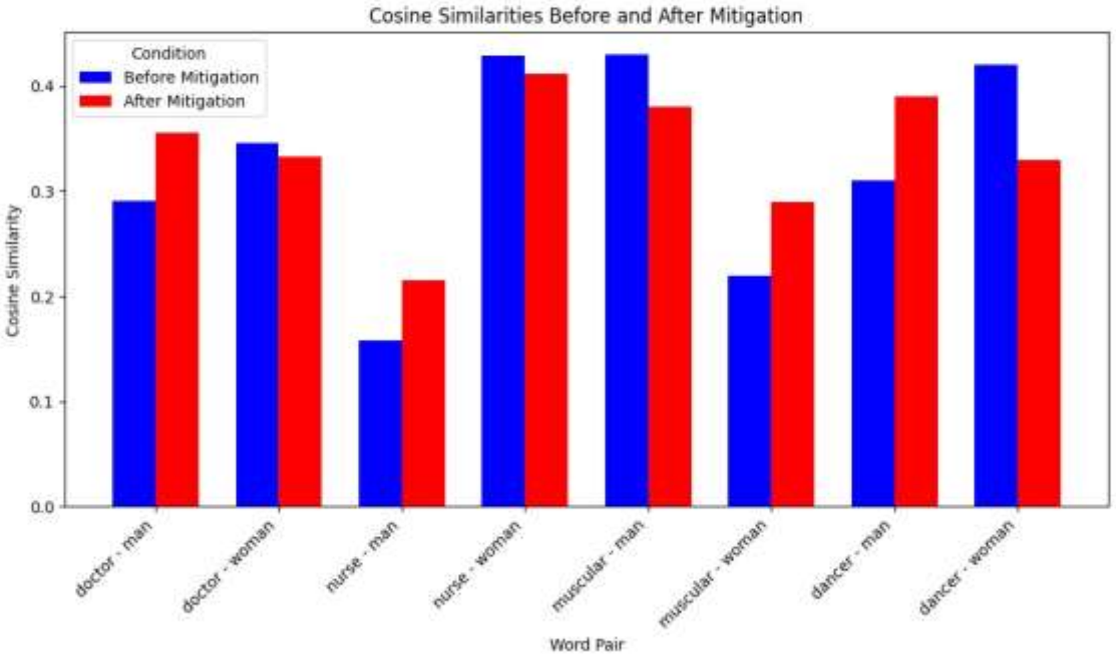
Experiments and Results

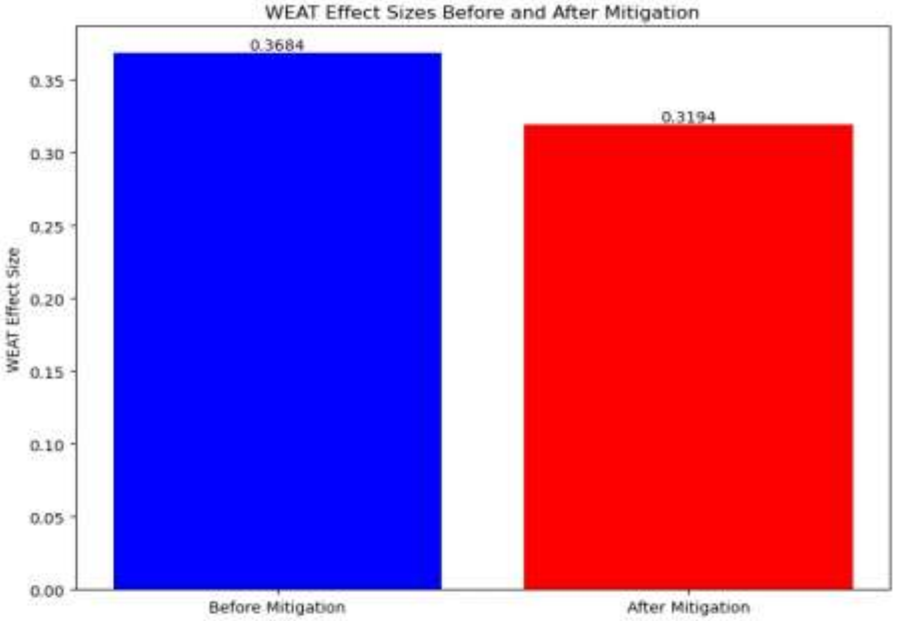
Key Findings:

Gender Bias Mitigation:

- Improved association neutrality for neutral terms and decreased WEAT effect size.


Cosine Similarity Distribution Before and After Mitigation


Cosine Similarities Before and After Mitigation


WEAT Effect Sizes Before and After Mitigation

Experiments and Results

What We Have Done:

- **FAISS + Flat Index:**
  - Performs **exact nearest neighbor search** in dense embedding space.

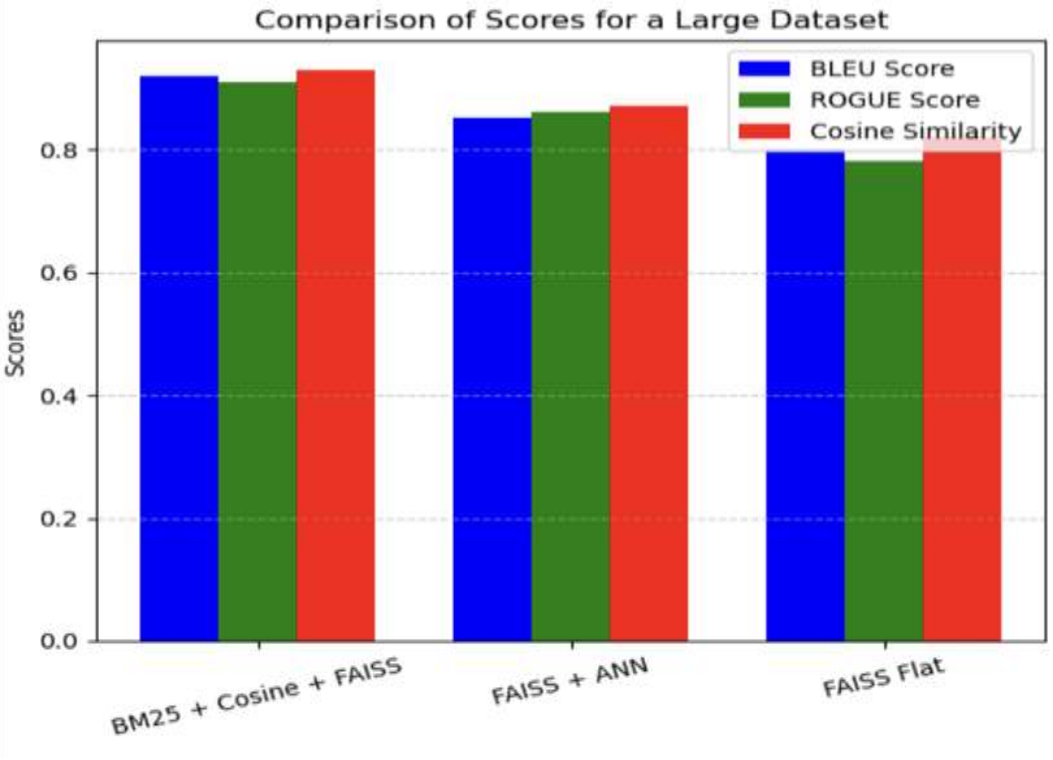
Combined Methods:

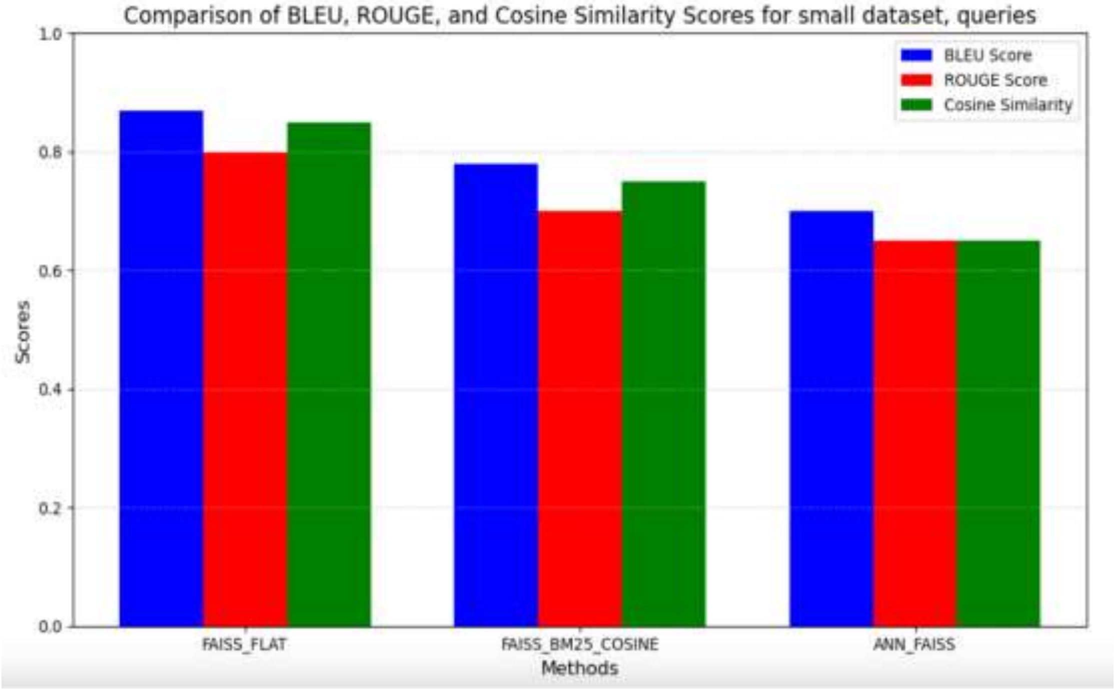
- **FAISS + BM25 + Cosine Similarity:**
  - Sparse keyword retrieval (BM25) + semantic similarity (FAISS).
  - Cosine similarity re-ranks combined results.
- **FAISS + Inverted Index + ANN:**
  - Inverted index for efficient candidate filtering.
  - FAISS ANN for dense vector semantic search.

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.

Metrics Overview:

1. **Cosine Similarity:**
   - ○ Higher cosine similarity means embeddings align closely with the context of the query.
2. **BLEU Score:**
   - ○ Compares text similarity at word/phrase level.
3. **ROUGE Score:**
   - ○ Overlap of unigrams, bigrams, Longest Common Subsequence (LCS).





Future Scope

Enhancements and Next Steps:

1. Knowledge Graphs Integration:
   - Knowledge graphs help connect related concepts in a structured way, this reduces confusion and ensures more accurate, reliable answers.
2. Implement Similarity Ranking:
   - Develop advanced ranking algorithms for retrieved contexts based on relevance and quality.
3. Generalization to Other Biases:
   - Expand debiasing techniques to address racial and occupational biases, validated through additional WEAT tests.