Decoding Bias in Letters of Recommendation: A Word Embeddings Approach

Tanay Nagar

Advisors: Sarah Jung, Fred Sala, Shamya Karumbaiah
Thesis Department: Computer Science

Background

Context: Recent shift to pass/fail grading in medical schools and USMLE Step 1 has led to increased reliance on subjective evaluation methods like Letters of Recommendation (LoRs).

Problem: LoRs may contain implicit biases that affect how applicants are perceived.

Opportunity: Advances in Natural Language Processing (NLP) allow systematic investigation of bias in written evaluations.

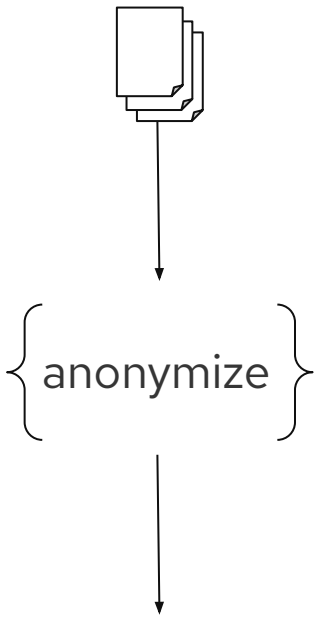
Research Overview and Objectives

Our Goal: To identify linguistic patterns that indicate implicit bias in LoRs for plastic surgery residency applicants based on gender and race/ethnicity.

Sub-Goals:

- ➔ Systematically measure bias using advanced NLP techniques.
- ➔ Utilized traditional **embedding approaches** (Word2Vec) along with **fine-tuned LLM embeddings** (BERT, RoBERTa, GPT-4, Llama) for deeper context capture.
- ➔ Evaluate the effectiveness of a novel method–**Embedding Bias Attention Mapping (EBAM)**–in directly visualizing intersectional biases via attention mechanisms.

Methods - Data



- **Dataset**: 5,679 plastic surgery residency applications (2017-2022) submitted via ERAS to a Midwestern academic medical center.
- **Data Processing**:
  - Anonymization using Part-of-Speech tagging and Named Entity Recognition.
- **Data Cleaning**
  - Eliminating common words and normalizing text (lemmatization) for consistency.
  - Generating descriptive word lists (e.g., top-N adjectives, nouns) to capture language patterns and downstream tasks.

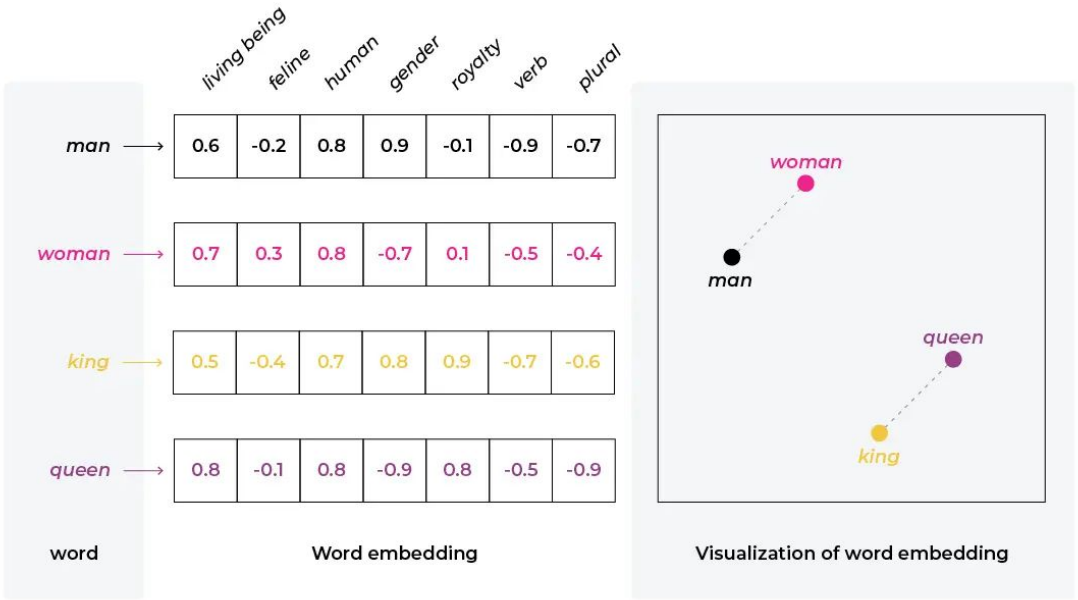
Embedding Training – Word2Vec & LLMs

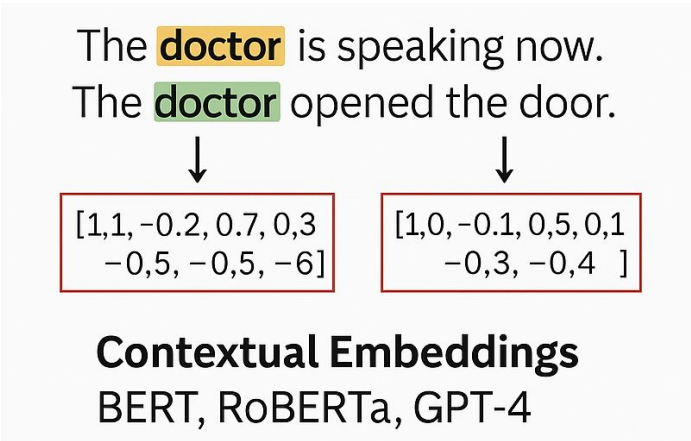
Word2Vec

- Trained on the cleaned LoR corpus
- Captures co-occurrence-based, static word semantics
- Grid search over hyperparameters (vector size, window, etc.)

LLM Fine-Tuning:

- BERT, RoBERTa, GPT-4, and LAMA fine-tuned on same corpus
- Captures contextual semantics & word meaning in specific usage
- Allows deeper understanding of how terms are used, not just how often





Bias Eval – Analogy, Nearest Neighbors & NLI

🧠 Analogy Tests

Purpose: Reveal conceptual gender associations

man : woman :: king : queen ✅

he : she :: assertive : ? gentle? ❌

Some outputs align with gendered stereotypes, others are incoherent.

🔍 Nearest Neighbors

Purpose: Reveal words that are semantically "close"

nurturing → soft, gentle, maternal

assertive → dominant, aggressive, bold

Biased words tend to cluster around stereotype-adjacent terms.

🧪 NLI Tests

Purpose: Detect subtle meaning shifts in gender-swapped sentences

"She is an ambitious leader." **0.2**

"He is an ambitious leader." **0.76**

- Entailment or contradiction?

Even neutral-looking sentences can be interpreted differently by models depending on gender.

Semantic Clustering & Visualization of Bias

K-Means Clustering

Purpose: Group words based on their bias-aligned embeddings to reveal hidden stereotype clusters.

Method:

- Applied K-Means clustering on adjectives/nouns
- Grouped based on proximity in embedding space
- Color-coded by bias direction (male/female/neutral)

Visualization Techniques

Violin & Box Plots: Bias distribution for top-N adjectives

3D PCA Scatter Plots: Dimensionality-reduced embeddings

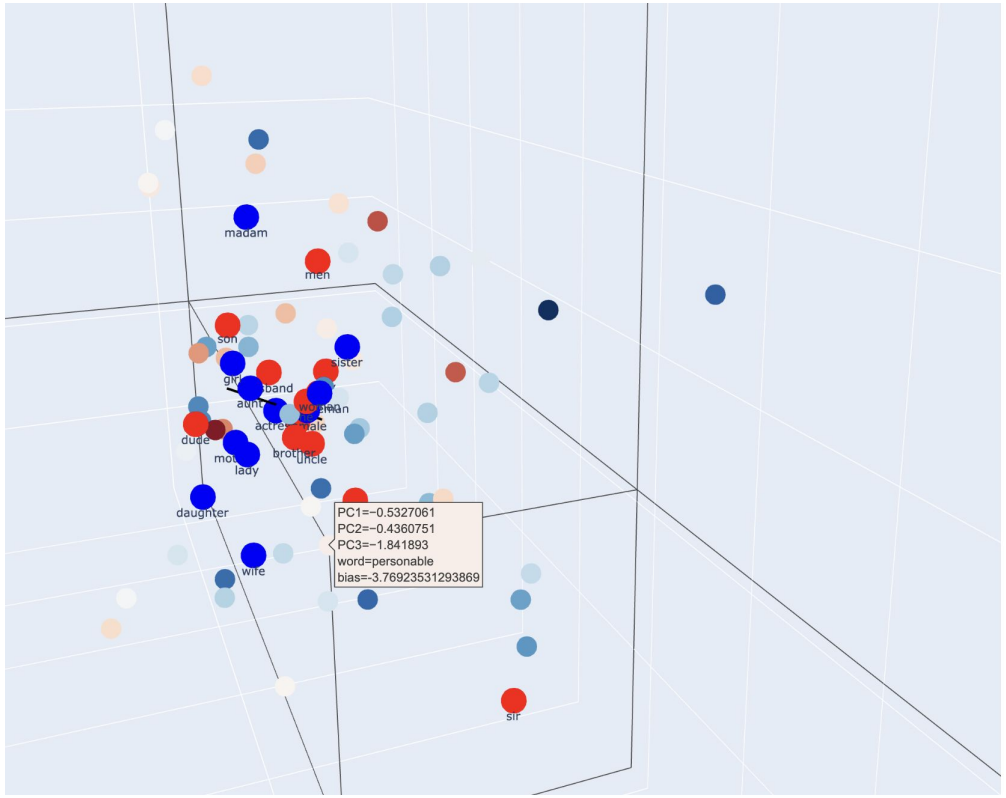
- Words color-coded by bias score
- Gender seed words used as anchors

Cosine Similarity Heatmaps:

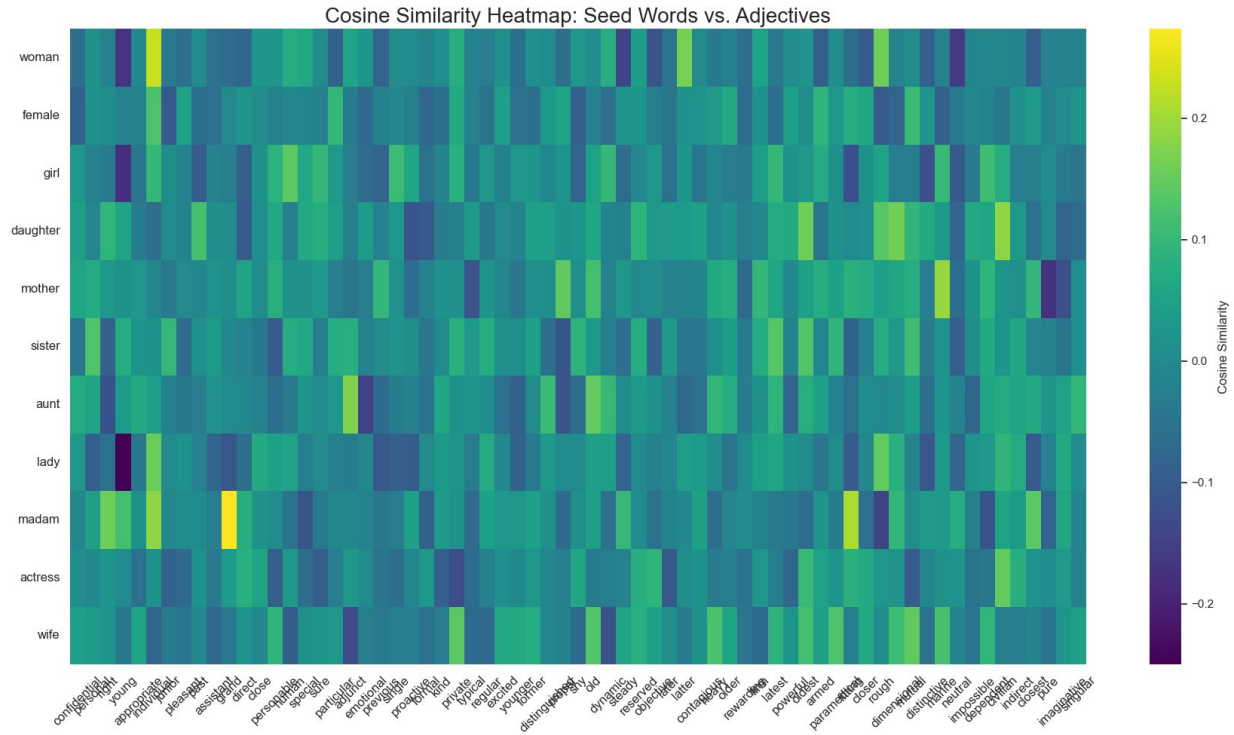
Semantic similarity across adjectives or nouns

Semantic Clustering & Visualization of Bias

K-Means Clustering



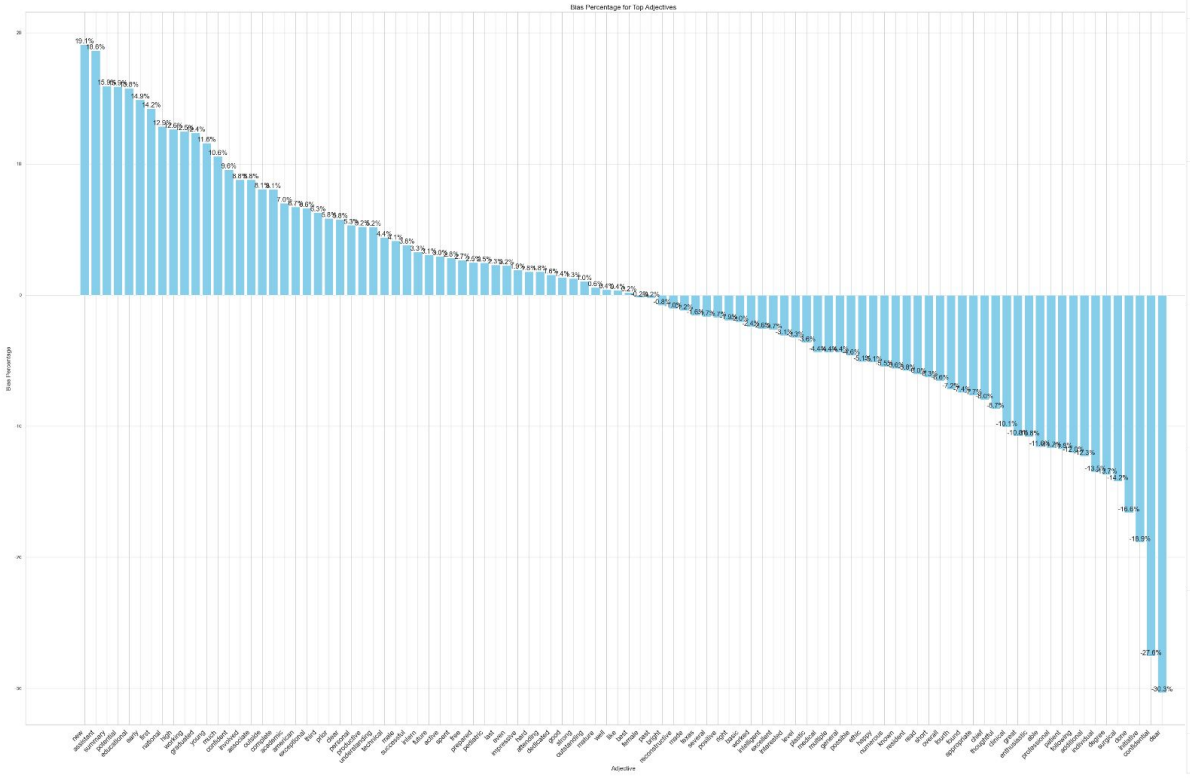
Visualization Techniques

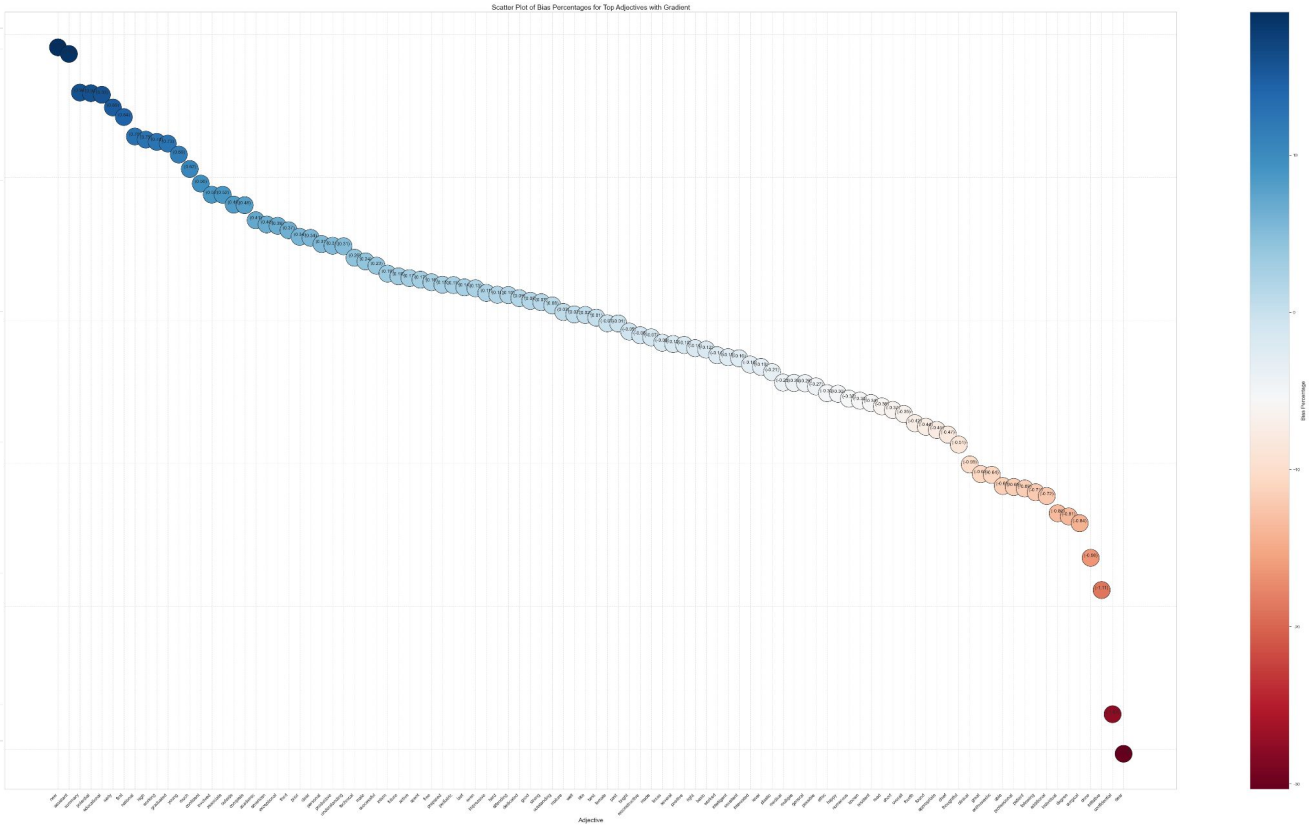


Semantic Clustering & Visualization of Bias

K-Means Clustering

Visualization Techniques



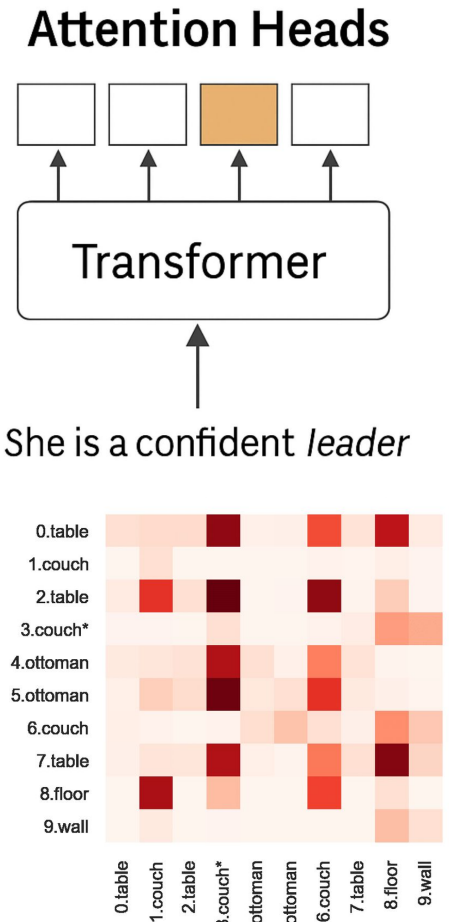


Embedding Bias Attention Mapping (EBAM)

- A novel method that analyzes the internal attention mechanisms of fine-tuned LLMs (like BERT, RoBERTa, GPT-4, LAMA).
- Goes beyond embeddings – looks inside the model's attention layers to find where bias resides.

How It Works

- Fine-tune LLMs on the LoR corpus.
- Select key word pairs (e.g., "he"/"she", "leader"/"helper").
- Extract and analyze attention weights from specific heads and layers.
- Visualize attention differences between gendered sentence variants.
- E.g., "She is a confident leader" vs. "He is a confident leader"
- Look for consistent patterns in which certain terms (like "confident") are attended to differently based on gender context.



Open-Source Tooling and Next Steps

🔧 Open-Source Tool

- Upload recommendation letter text
- View word-level bias scores and cluster visualizations
- Explore attention patterns using EBAM
- Run interactive tests (e.g., gender swaps, similarity heatmaps)

🔭 Future Work

- Expand EBAM to capture intersectional bias (e.g., gender × race/ethnicity)
- Integrate sentence-level comparisons for nuanced bias testing
- Extend the pipeline to other subjective texts (e.g., performance reviews, admissions essays)
    - Visualize trends across years to detect change over time

Department of Surgery

UNIVERSITY OF WISCONSIN
SCHOOL OF MEDICINE AND PUBLIC HEALTH

Exceptional People. Extraordinary Results.