Towards Fairer Health Recommendations:

Finding informative unbiased samples via Word Sense Disambiguation

Gavin Butts[1*], Pegah Emdad[2*], Jethro Lee[3*]

Shannon Song[2], Chiman Salavati[4], Wilmar Sosa Diaz[4], Roberto Montenegro[5], Scott Hale[6], Shiri Dori-Hacohen[4], Fabricio Murai[2]

* Equal Contribution
[1] Loyola Marymount University, [2] Worcester Polytechnic Institute, [3] Northeastern University, [4] University of Connecticut, [5] University of Washington School of Medicine, [6] Meedan & Oxford Institute

Table of contents

01

Introduction

Fairness and Health Recommender Systems

Introduction

Health Recommender Systems (HRS)

Impacts of Biased Data

Introduction

Personalized **health recommendations** deploying **machine learning** and **information retrieval**

Dependence of **Recommender systems' reliability** on the **quality** of their **training data**

Negative impact of **biased predictions** on **patient care** widening **health disparities**

Disambiguous Sample

Table 1: The term "white" in a racial vs. non-racial context

Race-Related	Not Race-Related
"5 Year Relative Survival: overall 84% for **white** women, 62% for black women, 95% for local disease, 69% regional disease (spread to lymph node), 17% for distant disease."	"**White** matter within the spinal cord contains the axons of neurons that are ascending and descending to transmit signals to and from the brain, respectively."

02

Related Work

Manual and Computational Bias Detection

Related Work

Manual Bias Detection

Computational Bias Detection

LLMs and Prompt Engineering

Related Work

Manual Bias Detection

Growing concerns over **biased AI models** in **healthcare recommender systems** due to their use in **high-stakes decisions**

Our Approach:

- Exploring AI models for debiasing medical text.
- Augmenting unbiased samples and evaluating a wider range of models, including LLMs
- Data refinement using WSD

Related Work

Computational Bias Detection

Our Approach:

- Apply **Large Language Models (LLMs)** for bias detection.
- Use **TinyLlama**, an efficient version of **Llama 2**, for bias classification.
- Implement **Word Sense Disambiguation (WSD)** to improve **data refinement** and enhance the set of **negative samples**.

Related Work

LLMs and Prompt Engineering

- **LLMs** perform **on par** with encoder-only models like **BERT** in NLP tasks **without fine-tuning**.

Prompting Techniques:

- **Zero-shot**, **Few-shot**, and **Chain of Thought (CoT)** prompting are crucial for improving **model quality** and **output accuracy**.

Our Approach:

- We are the first to evaluate **zero-** and **few-shot prompting** for detecting **bias in medical curricular content**.

03

Dataset

BRICC* Dataset for Bias Reduction in Curricular Content

*Salavati et al. (2024)

Data Labels

First-level: Identify Social Demographic

'Sex,' 'Gender,' 'Race,' 'Ethnicity,' 'Age,' and 'Geography'

Second-level: Identify Bias

'Biased,' 'Potentially Biased,' 'Non-Biased,' and 'Review'

Third-level: Identify Link Between Social Demographic and Medical Condition

Ex. 'Race-Disease'

Negative Samples

All Negative Samples

Extracted Negatives
XN

- Samples marked as biased without any other label

Labeled Negatives
LN

- Samples containing all labels

Extracted Negatives

Table 1: The term "white" in a racial vs. non-racial context

Race-Related	Not Race-Related
"5 Year Relative Survival: overall 84% for **white** women, 62% for black women, 95% for local disease, 69% regional disease (spread to lymph node), 17% for distant disease."	"**White** matter within the spinal cord contains the axons of neurons that are ascending and descending to transmit signals to and from the brain, respectively."

Uninformative Extracted Negative

Extracted Negatives

Table 1: The term "white" in a racial vs. non-racial context

Race-Related	Not Race-Related
"5 Year Relative Survival: overall 84% for **white** women, 62% for black women, 95% for local disease, 69% regional disease (spread to lymph node), 17% for distant disease."	"**White** matter within the spinal cord contains the axons of neurons that are ascending and descending to transmit signals to and from the brain, respectively."

1

0

4

Word Sense Disambiguation

Machine Learning for Word Sense Disambiguation and Classification

What is the most effective model for word sense disambiguation in a medical context? Can we produce accurate results?

Table 1: The term "white" in a racial vs. non-racial context

Race-Related	Not Race-Related
"5 Year Relative Survival: overall 84% for **white** women, 62% for black women, 95% for local disease, 69% regional disease (spread to lymph node), 17% for distant disease."	"**White** matter within the spinal cord contains the axons of neurons that are ascending and descending to transmit signals to and from the brain, respectively."

Word Sense Disambiguation (WSD) Experiments

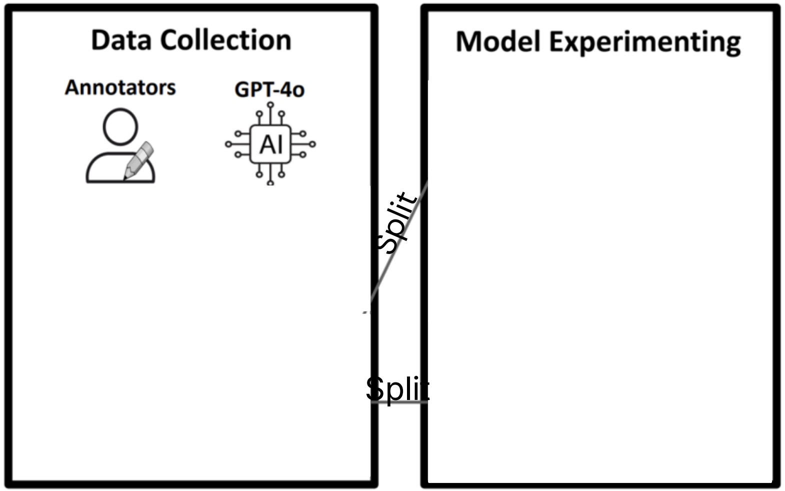


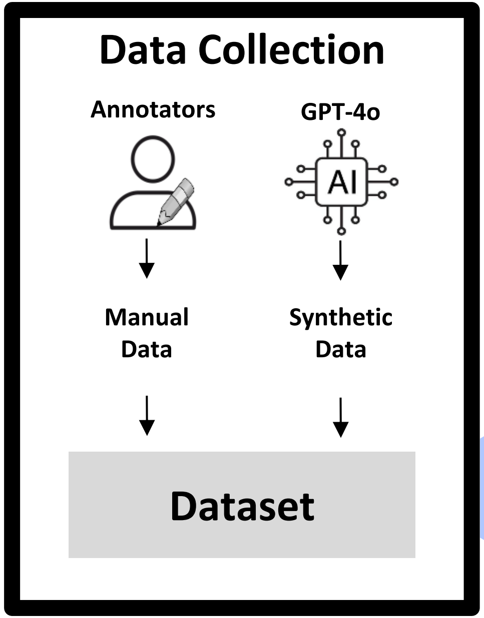
Figure 2: WSD training and evaluation. Excerpts manually labeled as race-related or not plus GPT-generated sentences are used to train and evaluate the WSD models.

Data for WSD

Extracted negatives randomly sampled

- Human expert labels data
  - Label: 1 if sample relates to a social demographic
  - Label: 0 otherwise
- Additional samples synthetically generated using GPT-4o
  - More accurate results?

Demographics of interest include Race and Ethnicity



WSD Problem Statement

Given a set of words $\mathcal{W}$ and a set of senses $\mathcal{S}_w = \{S_w^{(1)}, \ldots, S_w^{(k)}\}$ for each $w \in \mathcal{W}$

and a context (i.e. an ordered sequence of words)

$$x = (x_1, \ldots, x_{i-1}, w, x_{i+1}, \ldots x_n) \in \mathcal{X}$$

We are interested in determining if a term w is related to a sense s in an excerpt x

$$\textsc{IsRelated}(w, x, s_w) \in \{\textsc{true}, \textsc{false}\}$$

WSD Example

s = race/ethnicity

w = {'white', 'Black'...}

x

Black youth less likely to be diagnosed with MDD, Bipolar, or substance use disorder than **white** youth

$\textsc{IsRelated}(w, x, s_w)$= TRUE

White matter and Grey matter anatomy of the spinal cord.

x

$\textsc{IsRelated}(w, x, s_w)$= FALSE

Evaluation of WSD models

Table 2: Performance metrics for WSD on manually-annotated+GPT excerpts. Best result for each metric shown in bold. GlossBERT and GPT-4o are tied as the best models.

Metric	TF-IDF+ Logistic Reg.	ALBERT	Gloss BERT	GPT-3.5 Turbo	GPT-4o mini
Accuracy	0.839	0.926	**0.944**	0.925	**0.944**
Precision	0.816	0.935	**0.936**	0.916	**0.936**
Recall	0.839	0.977	**1.000**	**1.000**	**1.000**
F1 Score	0.817	0.956	**0.967**	0.956	**0.967**

Evaluation of WSD models

Table 3: Examples of WSD test cases and GlossBERT predicted probabilities for $y = 1$. Each excerpt has a term (bolded) listed among race/ethnicity keywords.

Input x (label y)	**Prediction**
Melanoma: increasing in incidence in the **white** population (CDC). ($y = 1$)	0.9998
2015 **American** Heart Association guidelines suggest treating patients presenting with systolic BP above 150-220 mmHg, but they do not offer a specific BP target. ($y = 0$)	0.9998
Calcific plaques are chalky **white** and arise from cardiac (aortic and mitral) valves. ($y = 0$)	0.0001

05

Bias Classification

Approach and Evaluation of Bias Detection Models

Bias Example

"They promote hair growth in the groin, axilla, chest and face, yet they also promote hair loss in the scalp in **men** who are genetically susceptible to androgenetic alopecia."

Label: Biased
Category: Gender Bias

Reasoning: "Use sex terms when speaking of populations, should be male instead of men. Also, include citation to support this assertion."

Bias Classification Problem Statement

- Formally, Salavati et al. define type-specific bias as a binary label $\text{BIAS}(x, t) \in \{\text{TRUE}, \text{FALSE}\}$ indicating whether excerpt x is biased with respect to a social identifier category t
- In the present work, we consider only the *general* definition of bias, regardless of which category t in a set $\mathcal{T}$ it belongs to:

$$\text{BIAS}(x, \mathcal{T}) = \text{TRUE} \iff \exists t \in \mathcal{T} \text{ s.t. } \text{BIAS}(x, t) = \text{TRUE}$$

Salavati et al. (2024)

Bias Classification Data

LN

Negatives labeled by human annotators

XN

Negatives extracted by use of social identifiers

XN*

Extracted Negatives filtered using WSD

Bias Classification Data Sets

LN

Human

annotated

dataset

LN+XN

Dataset

used by

Salavati et al.

LN+XN*

Refined

dataset by

Salavati et al

using WSD

Evaluation of Bias Detection models

Table 4: Performance metrics and 95%-CIs for RoBERTa, TinyLlama trained on dataset variants (LN+XN*, LN+XN, LN). Best results among each model variants (resp. across all models) and statistical ties shown are bolded (resp. underlined).

Metric	RoBERTa			TinyLlama		
	LN+XN*	LN+XN	LN	LN+XN*	LN+XN	LN
Precision	**0.613 ± 0.015**	**0.640 ± 0.021**	0.526 ± 0.029	**<u>0.675 ± 0.008</u>**	**<u>0.693 ± 0.028</u>**	0.536 ± 0.020
Recall	**0.692 ± 0.024**	0.667 ± 0.023	**<u>0.719 ± 0.026</u>**	**0.548 ± 0.030**	0.519 ± 0.029	**0.607 ± 0.035**
F1 Score	**<u>0.650 ± 0.013</u>**	**<u>0.652 ± 0.017</u>**	0.606 ± 0.017	**0.604 ± 0.021**	**0.593 ± 0.017**	**0.568 ± 0.016**
F2 Score	**<u>0.674 ± 0.019</u>**	**0.661 ± 0.016**	**0.669 ± 0.016**	**0.569 ± 0.027**	**0.546 ± 0.024**	**0.591 ± 0.025**
AUC	**<u>0.927 ± 0.003</u>**	**<u>0.930 ± 0.009</u>**	0.910 ± 0.008	**0.907 ± 0.005**	**0.903 ± 0.005**	0.871 ± 0.011

Evaluation of Bias Detection models

Table 5: Performance Metrics and 95%-CIs for Fine-Tuned Models against Baseline (*Salavati et al., 2024). Best results and statistical ties shown in bold.

Metric	RoBERTa	TinyLlama	Baseline*
Precision	0.613 ± 0.015	**0.675 ± 0.008**	0.504 ± 0.054
Recall	0.692 ± 0.024	0.548 ± 0.030	**0.812 ± 0.069**
F1 Score	**0.650 ± 0.014**	0.604 ± 0.021	0.615 ± 0.022
F2 Score	0.674 ± 0.019	0.569 ± 0.027	**0.717 ± 0.027**
AUC	**0.927 ± 0.003**	0.907 ± 0.005	**0.923 ± 0.004**

Evaluation of Bias Detection models

Table 6: Performance Metrics and 95%-CIs for Prompting GPT-4o mini. Best results for each metric shown in bold. AUC was ommitted as it cannot be computed for binary outputs.

Metric	Zero-Shot	Few-Shot
Precision	**0.367 ± 0.071**	0.259 ± 0.019
Recall	0.260 ± 0.029	**0.610 ± 0.026**
F1 Score	**0.303 ± 0.040**	**0.363 ± 0.023**
F2 Score	0.274 ± 0.032	**0.480 ± 0.025**

06

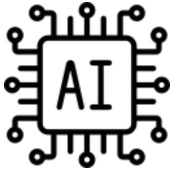
Conclusion

Key Findings and Future Implications

Conclusion

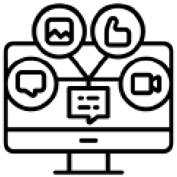
- **Health-related applications** and **recommender systems** are prone to biases
- Developed a framework to **detect and diagnose bias** in medical curricula by an emphasized focus on data over model
- **WSD** models were **highly effective** at distinguishing biased excerpts from non-biased ones
- While **prompt engineering** of **LLMs** can handle many tasks, they are **not well-suited** for health related bias classification

Discussion



Further explore the potential of ChatGPT-4o (or **other future OpenAI models**)

Use of **other bias categories** (E.g. geography)



Use case in **other domains** (Crucial role of tone in determining word meaning esp. in social media)

Challenges with LLMs (Computational cost, time constraints, accessibility issues)

Thank You.
Questions?

Link to paper