

Towards Fairer Health Recommendations: finding informative unbiased samples via Word Sense Disambiguation

Gavin Butts*
Loyola Marymount University
gbutts@lion.lmu.edu

Pegah Emdad*
Worcester Polytechnic Institute
pemdad@wpi.edu

Jethro Lee*
Northeastern University
lee.jet@northeastern.edu

Shannon Song
Worcester Polytechnic Institute
smsong@wpi.edu

Chiman Salavati
University of Connecticut
chiman.salavati@uconn.edu

Willmar Sosa Diaz
University of Connecticut
willmar.sosa_diaz@uconn.edu

Shiri Dori-Hacohen
University of Connecticut
shiridh@uconn.edu

Fabricio Murai
Worcester Polytechnic Institute
fmurai@wpi.edu

ABSTRACT

There have been growing concerns around high-stake applications that rely on models trained with biased data, which consequently produce biased predictions, often harming the most vulnerable. In particular, biased medical data could cause health-related applications and recommender systems to create outputs that jeopardize patient care and widen disparities in health outcomes. A recent framework titled *Fairness via AI* posits that, instead of attempting to correct model biases, researchers must focus on their root causes by using AI to debias data. Inspired by this framework, we tackle bias detection in medical curricula using NLP models, including LLMs, and evaluate them on a gold standard dataset containing 4,105 excerpts annotated by medical experts for bias from a large corpus. We build on previous work by coauthors which augments the set of negative samples with non-annotated text containing social identifier terms. However, some of these terms, especially those related to race and ethnicity, can carry different meanings (e.g., “white matter of spinal cord”). To address this issue, we propose the use of Word Sense Disambiguation models to refine dataset quality by removing irrelevant sentences. We then evaluate fine-tuned variations of BERT models as well as GPT models with zero- and few-shot prompting. We found LLMs, considered SOTA on many NLP tasks, unsuitable for bias detection, while fine-tuned BERT models generally perform well across all evaluated metrics.

KEYWORDS

medical text data, bias detection, LLMs, word sense disambiguation

ACM Reference Format:

Gavin Butts, Pegah Emdad, Jethro Lee, Shannon Song, Chiman Salavati, Willmar Sosa Diaz, Shiri Dori-Hacohen, and Fabricio Murai. 2024. Towards Fairer Health Recommendations: finding informative unbiased samples via Word Sense Disambiguation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (FAccRec '24)*. ACM, New York, NY, USA, 7 pages. <https://doi.org/XXXXXXX.XXXXXXX>

*Equal contribution.

1 INTRODUCTION

For decades, medicine has been marred by implicit and explicit biases that continue to negatively impact patient outcomes by perpetuating stereotypes and contributing to health disparities among social groups that face systemic oppression [8, 9]. Despite efforts to remediate and address these biases from their source, many medical schools still incorporate biased medical teachings during the preclinical years [12, 28]. Many educators continue to misuse race as a substitute for genetics or ancestry, or they use gender and sex terms incorrectly reinforcing the notion that sex and gender are binary or fixed rather than fluid, which can potentially alienate gender-nonconforming students and patients [1, 14, 15]. The current focus in AI research is primarily on identifying and exposing bias within AI systems, often without addressing the root causes of bias inherent in the data these systems are built upon. As long as structural inequalities exist in the real world, AI systems will perpetuate these biases [10]. By harnessing machine learning to analyze and detect these biases, we can advance equity in medical training and the fairness of AI models, leading to a more accurate and effective healthcare system.

Recently, Salavati et al. [25] introduced the BRICC (Bias Reduction in Curricular Content) dataset and proposed a systematic and scalable AI-based method for identifying potential bias in medical curricula. Given the steep cost of false negatives (i.e., classifying a biased sentence as unbiased), they emphasize that recall must be prioritized over precision. Moreover, due to the inherent difficulty of the task, one of the best approaches in this High-Recall Information Retrieval setting is the Technology Assisted Review (TAR) [7, 18], whereby a set of experts reviews the samples flagged by a model, as envisioned by Salavati et al. The paper also used a curated list of social identifiers to find additional, negative (i.e., non-biased) samples in unlabeled data. However, this latter approach suffered from social identifier terms that had ambiguous meanings, leading to lower quality of the training data, and negative samples that were too “easy” to classify as non-biased. For example, one social identifier used to filter for race-related data was “white”. In Table 1, simply searching for the keyword “white” will include both race-related and non-race-related text excerpts.

We believe that using these ambiguous meaning terms as-is leads to overestimating the true discernment power of the bias classifier

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.”

by making the problem too easy. It may be that the classifier is actually differentiating race-related from non-race-related terms, rather than biased from non-biased sentences.

For this reason, we propose a new framework to augment the sampling process for negative examples, using Word Sense Disambiguation (WSD) methods for data enhancement. We hypothesize that this would improve the distinction between biased and non-biased sentences in the bias classification process.

Our main contributions are as follows:

- (1) We enhance a framework for detecting bias in medical curriculum content, with a focus on improving data quality.
- (2) We leverage Word Sense Disambiguation (WSD) models in training bias detection classifiers by filtering out irrelevant samples from the data. Moreover, we use ChatGPT-4o to augment a set of manually labeled examples with synthetic samples to fine-tune and/or evaluate WSD models.
- (3) We fine-tune and evaluate various Transformer-based models including DistilBERT, RoBERTa, and BioBERT for the bias detection task. In addition, we use Large Language Models (LLMs), such as TinyLlama, for bias classification, evaluating zero-shot vs. few-shot prompting with GPT, to improve the performance of bias detection.
- (4) We present a comprehensive evaluation of the various models, highlighting the improvements achieved through the use of WSD and ChatGPT-generated sentences.

2 RELATED WORK

Health Recommender Systems (HRS). Recommender systems have become integral to the healthcare industry, providing personalized medical recommendations that enhance patient understanding of their medical condition and improve health outcomes [27]. These systems assist healthcare professionals in predicting and treating diseases by analyzing patient data to recommend personalized diets, exercise regimens, medications, diagnoses, and other health services [21, 24]. Despite numerous studies exploring various aspects of HRS, the literature for addressing various types of biases in such systems rooted in curricula contents is limited [25].

Debiasing medical corpora manually and via AI. Concerns over biased AI models and, particularly, recommender systems in healthcare applications have been gaining more attention due to their increased use in high-stake decisions [4]. In essence, their biases are rooted in implicit and explicit biases embedded in the data used for training them [13], which stem from various sources, including inherent biases in medical literature, the subjectivity of human annotators, and historical and systemic inequities present in healthcare systems [25]. Numerous recent studies have aimed to quantify and address this issue both manually and through AI.

Khan et al. [16] manually explored the systemic bias held by medical professionals when writing recommendation letters. On the other hand, Raza et al. [22] and Salavati et al. [25] aimed to detect bias in medical text using transformer-based language models. The former used a semi-autonomously labeled dataset covering diverse medical topics, whereas the latter employed a dataset manually labeled by medical experts focusing on biased information in medical curricular texts. Although both studies provide a comprehensive overview of bias detection, ensuring high-quality data remains an issue. While we also explore AI models for debiasing medical text data, we investigate better ways of augmenting the set of unbiased samples and consider a wider gamut of models, including LLMs.

Machine Learning for Bias Detection. Prior works have applied various BERT models for bias classification tasks. Tiderman et al. [26] used DistilBERT, a transformer-based distilled BERT model, to classify biased information in social media content. Similarly, Raza et al. [22] achieved the best results for bias classification in medical text through fine-tuning BERT, a simple encoder-only transformer. Building on the existing literature for bias detection in medical contexts, we additionally apply Large Language Models (LLMs) for this task. Specifically, we use TinyLlama [31], a computationally efficient variant of Llama 2, for bias classification. In addition, we consider additional strategies for constructing the set of negative samples— such as through the use of WSD for data refinement.

Use of LLMs for NLP tasks and prompt engineering. In NLP tasks, prompting LLMs have been shown to perform on par with encoder-only architectures, like BERT, without the need for fine-tuning [6]. It has been shown that prompting techniques, such as zero-shot, few-shot, or chain of thought (CoT), serve a key role in the quality and correctness of a model’s output [19]. These techniques have been used in many tasks, such as sentiment analysis [3], text classification [5], as well as for healthcare applications, such as question-answering, and as a clinical recommender system [20, 29]. Despite these initiatives, we are the first to evaluate zero- and few-shot prompting for detecting bias in medical curricular content.

3 DATASET

Our work builds on the BRICC dataset introduced by Salavati et al. [25], which consists of 509 PDF files and 12,647 pages of medical school instructional materials annotated by medical students and experts trained in identifying bias. Within the dataset, there are three tiers of coding. The first-level codes identify social identifiers within the excerpt. The second-level codes assess the presence or absence of bias in the excerpt, categorized into four distinct groups: ‘**biased**’, ‘**potentially biased**’, ‘**non-biased**’, and ‘**review**’. Additionally, third-level codes establish a link between a medical condition and one or more categories of social identifiers (e.g., race), specifying the type of identity and whether it was portrayed in a biased or unbiased manner. Each excerpt is then assigned one or more codes formatted as “TYPE-disease”, where TYPE represents one of 17 categories of social identifiers. Akin to the previous work, we focus on the most frequent types including sex, gender, race, ethnicity, age, and geography. Each category is associated with a list of keywords that can signify social identifiers.

Table 2: BRICC Dataset Characteristics

	Counts
Number of PDF Files	509
Total Number of Pages	12,647
Annotated Excerpts	4,105
Labeled Positives	1,116
Labeled Negatives (LN)	2,989
Extracted Negatives (XN)	4,391

Positive and negative samples. **Positive samples** are defined as those excerpts that contain either a ‘biased’, ‘potentially biased’, or ‘review’ and a selected “TYPE-disease”. **Negative samples** are subdivided into various types, as detailed in the previous work [25]. The negative types that we are most interested in are referred to as extracted negatives (XN). In this case, these are sentences from the corpus that, despite containing a category keyword, were deliberately excluded from the annotation process. Please refer to supplemental materials for a detailed explanation of negative types found in the BRICC dataset. Of the XN types, we filtered for those that contained at least one relevant keyword relating to our selected “TYPE-disease”. The distribution of positives and negatives across the dataset is displayed in Table 2.

We focus on XN samples in our experiments because the authors previously reported higher recall, 0.925, but at the expense of precision, 0.504, by using this negative set [25]. Despite this improved performance, we noticed that many of the keywords may lead to the inclusion of non-“TYPE-disease” related content. An example of this occurrence may be seen in Table 1. Hence, retaining only negative samples that relate to the social demographics of interest is a key computational task, which we address using WSD.

Labeling data for WSD. To gather samples suitable for training WSD models, we selected those XN excerpts that contained a keyword related to the selected social demographics categories: sex, gender, race, ethnicity, geography, and age. After obtaining a random sample of XN excerpts for each category, we had a human expert annotate whether the meaning of the keyword term was indeed related to that category or not. Based on the results of this annotation process, we decided to only focus on race keywords because they suffered the most from ambiguity. The other bias categories did not have a significant degree of ambiguity that required correction. These labeled excerpts were used to train our WSD models.

4 PRELIMINARIES

4.1 LLM Prompting

Reynolds et al. [23] suggest that zero-shot prompts could significantly outperform few-shot prompts. Their analysis highlights the need to consider the role of prompts in controlling and evaluating the performance of language models. Their study stated that since GPT-3 is often not learning from few-shot examples during the run time, this model can effectively be prompted without examples [23]. Additionally, Kojima et al. [17] demonstrate that chain of thought (CoT) prompting, a recent technique for eliciting complex multi-step reasoning through step-by-step answer examples, achieved state-of-the-art performances in arithmetics and symbolic reasoning tasks. They proposed Zero-shot-CoT, a zero-shot

template-based prompting using chain of thought reasoning, and highlighted its high performance.

4.2 Word Sense Disambiguation Task Definition

Generally speaking, word sense disambiguation (WSD) is the task of identifying the correct sense of a polysemous $w \in \mathcal{W}$ (a word with multiple meanings) in a given context x . Formally, given a set of words \mathcal{W} , a finite set of possible senses $\mathcal{S}_w = \{S_w^{(1)}, \dots, S_w^{(k)}\}$ for each $w \in \mathcal{W}$ and a context (ordered sequence of words) $x = (x_1, \dots, x_{i-1}, w, x_{i+1}, \dots, x_n) \in \mathcal{X}$, find a function $f : \mathcal{W} \times \mathcal{X} \rightarrow \mathcal{S}$, such that $f(w, x)$ is the correct sense of w in context x .

For this paper, we are interested in determining if a term w , listed as a possible social identifier for category t , is related to t in an excerpt x . To do so, we need to learn a function $\text{ISRELATED}(w, x, t) \in \{\text{TRUE}, \text{FALSE}\}$. For instance, the set of social identifiers for race is $\mathcal{S}_{\text{race}} = \{\text{‘white’}, \text{‘black’}, \dots\}$. Ideally, in one of the examples seen earlier, we want $\text{ISRELATED}(\text{‘white’}, \text{‘white matter within...’}, \text{‘race’}) = \text{FALSE}$.

4.3 Bias Detection Task Definition

In the context of medical education, we consider bias detection as a High Recall Information Retrieval task. It is the first step in a TAR system. This task consists of classifying a text excerpt x as unbiased ($\hat{y} = 0$) or potentially biased ($\hat{y} = 1$). In the latter case, the sample would be subsequently reviewed by a medical expert.

Bias may be related to one or more categories of social identifiers, including race, ethnicity, sex, gender, age, and geography. For instance, “*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.*” is labeled by medical excerpts as ‘**biased**’ with respect to *gender* (designated by the social identifier *men*). As explained, in the comment from one of the annotators: “*Use sex terms when speaking of populations, should be male instead of men. Also, include citation to support this assertion.*”

Formally, Salavati et al. [25] 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: $\text{BIAS}(x, \mathcal{T}) = \text{TRUE} \iff \exists t \in \mathcal{T} \text{ s.t. } \text{BIAS}(x, t) = \text{TRUE}$.

5 METHODOLOGY

In this section, we provide an overview of the proposed framework. Figure 1 (left) shows the data processing steps performed by Salavati et al. [25], which we leverage in the present work. As explained, in addition to labeled data, BRICC includes negative samples extracted from the non-annotated data (denoted as XN). Figure 1 (center) illustrates the application of WSD to filter out irrelevant samples, which results in the filtered XN set (XN*). This process is described in detail in Section 5.1. Last, Figure 1 (right) depicts the augmentation of labeled data with XN* for training different bias classifiers, whose performance we evaluate. Details are discussed in Section 5.2.

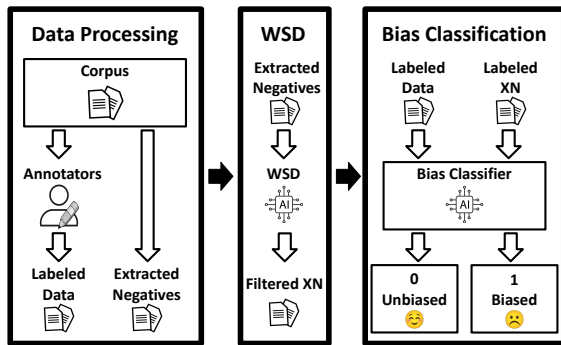


Figure 1: Workflow stages. (Left) **Data processing:** annotated excerpts are labeled as ‘biased’ (positive) or ‘non-biased’ (negative); XN: additional sentences extracted as negative examples. (Center) **Word Sense Disambiguation (WSD)** used for selecting from XN relevant negatives (XN*). (Right) **Training and evaluation of bias classifiers.**

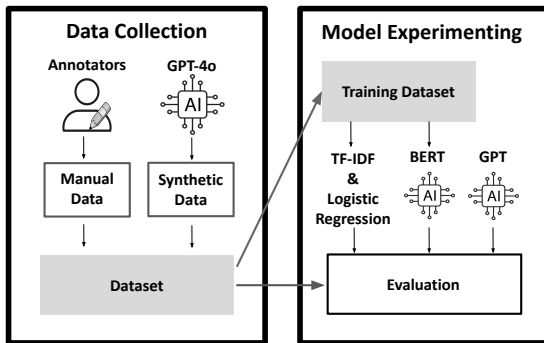


Figure 2: WSD training and evaluation. Excerpts manually labeled as race-related or not plus GPT-generated sentences are used to train and evaluate the WSD models.

5.1 Word Sense Disambiguation Experiments

We evaluate several models for WSD: a simple baseline, two fine-tuned variants of BERT, and two GPT models. For fine-tuning and evaluation, we combined our manual annotations with sentences generated by ChatGPT-4o, yielding 352 labeled excerpts.

Experimental Setup. We divide each of the two datasets (manually annotated excerpts and synthetic samples) independently with a 70-15-15 stratified split into training, validation, and test sets.

We investigated two ways of building the training set:

- Only manually-annotated excerpts;
- Both manually-annotated excerpts and synthetic samples.

WSD Models. We evaluate three models shown in the recent literature to perform well on WSD: ALBERT, GlossBERT, and GPT models. These models are compared to our baseline, a logistic regression with TF-IDF.

Fine-tuning and Prompting. We fine-tune all layers of the pre-trained ALBERT and GlossBERT models with a learning rate of 2×10^{-5}

and weight decay of 0.01 over 10 epochs, keeping the model that yielded the smallest validation loss.

Given the substantial empirical and theoretical evidence supporting the benefits of chain of thought (CoT) prompting in various LLM tasks [11, 30], we incorporate CoT into our zero-shot prompts. We opted to follow the prompt template presented in Kojima et al. [17], which was shown to produce the highest accuracy (i.e., “Let’s think step by step”). To prompt the model, we first specified the model’s role: “You are a helpful assistant that determines if the sentence is race or ethnicity related”. Then, we defined the task as: “Given the sentence ‘text’, think step by step: Is this sentence race or ethnicity related? Only output 1 or 0. If this sentence contains any terms relating to race or ethnicity, state 1. Otherwise, state 0.”

Metrics. We evaluate the models’ performance on the test set with respect to accuracy, precision, recall and F1 score.

5.2 Bias Classification Experiments

Using the BRICC dataset, we fine-tune binary classification neural language models and prompt pre-trained LLMs for bias classification. For fine-tuning, we consider encoder-only and decoder-only models (encoder-decoder models are often reserved for multi-modal tasks and causal language inference [2]). The fine-tuned models include RoBERTa, DistilBERT, BioBERT, and TinyLlama. Using prompt engineering, we additionally prompt GPT-4o mini.

Different sets of negatives. The datasets we will consider contain all positive samples, plus one of the following:

- Labeled negatives (LN),
- Labeled negatives plus extracted negatives filtered by keywords (LN + XN), and
- Labeled negatives plus extracted negatives filtered using word sense disambiguation (LN + XN*).

As described in Section 3, the set of extracted negatives (XN) is constructed by filtering data based on keywords that relate to either gender, sex, race, ethnicity, age, and/or geography. Then, we apply the best-performing WSD model to ensure these samples truly relate to the social demographics of interest, resulting in XN*. To assess performance variability as a function of the data splits, we split the dataset in K-folds for cross-validation and calculate average performance and confidence intervals.

Fine-tuning. To construct the models, we add a classification head to each language model and fully fine-tune each model along with the classification head. The models we utilize are: RoBERT, DistilBERT, and BioBERT, all encoder-only, and TinyLlama, a 1.1B decoder-only model derived from Meta’s Llama 2.

For each dataset we outlined, we do initial fine-tuning on RoBERTa, DistilBERT, BioBERT, and TinyLlama with a batch size of 8 and a learning rate of 2×10^{-5} . We use the validation set to tune the hyperparameters with grid search, leading to the final model.

Prompting. We evaluate the performance of GPT-4o mini by using zero- and few-shot prompting. To prompt this model, we first establish the model’s role: “You are a helpful assistant that determines if text is biased”. Then, we establish the task. We find that the best task description is: “Given text, determine if the text contains bias or no bias. The bias may target gender, sex, race, ethnicity, age, and/or

Table 3: Few-shot example inputs, outputs, and reasoning used for prompting GPT-4o mini for bias classification

Input x	Label, Comment
52 year old, married female with one daughter, employed as a school administrator with no prior psych history reports 2 month h/o [history of] sadness, subjective anxiety and intermittent trouble falling asleep.	Label: 1 , Use gender terms like woman for case studies
Once patient is on another treatment for her disorder, she no longer needs this medication.	Label: 0 , n/a (from XN set)
Recent meta-analysis suggested no difference in prevalence among countries, rate is 1-2% with increase during late adolescence.	Label: 1 , Term 'Late Adolescence' is an unclear time period

Table 4: 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

geography, include exclusive language, or make unsupported claims. The text may also contain no bias at all. If the text has bias state 1, if the text does not have bias state 0.”

Both zero- and few-shot prompting used the prompt above. For few-shot prompting, we tested various sets of examples from the dataset. Table 3 shows the subset that performed best.

Metrics. For the bias detection task, we also evaluate the models’ performance on the test set with respect to precision, recall and F1 score. In addition, we consider the F2 score and area under the ROC curve (AUC). F2 is similar to F1 but prioritizes recall over precision. Due to the class imbalance, AUC is more relevant than accuracy because it accounts for all possible threshold choices.

6 RESULTS

6.1 Evaluation of WSD models

Table 4 presents the evaluation results for the WSD models examined. The baseline achieved worse results than the other models for every metric. We find that GlossBERT outperforms ALBERT and that GPT-4o mini improves upon GPT-3.5 Turbo. Furthermore, both GlossBERT and GPT-4o are tied as the best models, both exhibiting a very high F1 Score (0.967). Using the cost as a tie-breaker between the two, we opted to use GlossBERT for the WSD task performed on the extracted negatives.

Table 5 illustrates examples from the test set, two of which were correctly identified and one that was not. While the first and third were correctly predicted with high confidence, the middle row was incorrectly classified with a “high confidence prediction”. For the few instances that GlossBERT incurred false positives, a closer inspection has revealed that those excerpts may be lacking enough context for this specific task. For example, in the middle row, the

Table 5: 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

Table 6: Performance Metrics and 95%-CIs for Fine-Tuned Models trained on LN+XN* data. RoBERTa yields the highest averages, but it is statistically tied with DistilBERT.

Metric	RoBERTa	DistilBERT	BioBERT
Precision	0.613 \pm 0.015	0.605 \pm 0.013	0.581 \pm 0.014
Recall	0.692 \pm 0.024	0.649 \pm 0.030	0.620 \pm 0.019
F1 Score	0.650 \pm 0.014	0.626 \pm 0.018	0.599 \pm 0.010
F2 Score	0.674 \pm 0.019	0.639 \pm 0.025	0.611 \pm 0.014
AUC	0.927 \pm 0.003	0.921 \pm 0.006	0.904 \pm 0.003

Table 7: Performance Metrics and 95%-CIs for Prompting GPT-4o mini. Best results for each metric shown in bold. AUC was omitted as it cannot be computed for binary outputs.

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

use of ‘American’ only indirectly relates to the ethnicity of the people that an organization serves.

We also investigate whether the synthetic samples generated by ChatGPT-4o were trivial, which would artificially inflate performance. When we evaluate the model results on only manually annotated excerpts, the performance of all models stays somewhat similar, except for GPT models, both of which achieve 100% accuracy. Therefore, the synthetic examples are at least as hard as the manually annotated excerpts for BERT models, justifying their use in our evaluation.

In addition, we evaluate the models’ performance when trained only on the manually-annotated data. In this case, there is a performance drop for the fine-tuned models (ALBERT declines from **0.926** to **0.852** and GlossBERT declines from **0.944** to **0.852** accuracy), indicating that the synthetic samples help the models to generalize better. Furthermore, GlossBERT remains tied as the best model, which supports our choice of using it for building the set of filtered extracted negatives in the bias detection task.

6.2 Evaluation of Bias Detection Models

Firstly, we compare the performance of the fine-tuned BERT variants on the bias detection task. Table 6 displays the models’ performance with respect to precision, recall, F1 and F2 score, and AUC.

Table 8: 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	0.675 ± 0.008	0.693 ± 0.028	0.536 ± 0.020
Recall	0.692 ± 0.024	0.667 ± 0.023	0.719 ± 0.026	0.548 ± 0.030	0.519 ± 0.029	0.607 ± 0.035
F1 Score	0.650 ± 0.013	0.652 ± 0.017	0.606 ± 0.017	0.604 ± 0.021	0.593 ± 0.017	0.568 ± 0.016
F2 Score	0.674 ± 0.019	0.661 ± 0.016	0.669 ± 0.016	0.569 ± 0.027	0.546 ± 0.024	0.591 ± 0.025
AUC	0.927 ± 0.003	0.930 ± 0.009	0.910 ± 0.008	0.907 ± 0.005	0.903 ± 0.005	0.871 ± 0.011

Table 9: 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

While RoBERTa and DistilBERT are statistically tied, BioBERT clearly performs worst among all BERT models. We select the RoBERTa model for further comparison due to the model’s higher mean evaluation metrics with lower standard deviations.

Secondly, we evaluate the performance of zero- and few-shot prompting with GPT-4o mini on bias detection. Table 7 shows the results obtained using the prompting techniques outlined in Section 5.2. Despite the substantial increase in recall seen with few-shot prompting, the low overall performance of GPT-4o mini deems it unsuitable for the bias detection task.

Next, we compare the best BERT model and a baseline from our prior work [25] with a fine-tuned TinyLlama. Table 9 shows the comparison results. Although TinyLlama achieves high precision, its lower recall causes it to be outperformed by RoBERTa and by the baseline with respect to both F1 and (especially) F2 scores. RoBERTa and the baseline are statistically tied with the highest AUCs (0.927 ± 0.003 and 0.923 ± 0.004), indicating that for either model the classification threshold can be tuned to find a trade-off between precision and recall suitable for the target application.

Last, we conduct an ablation test to assess the impact of WSD for data refinement by comparing the performance of RoBERTa and TinyLlama across various dataset configurations. The results in Table 8 show that the LN+XN* setting led to higher recall averages than LN+XN (despite not statistically significant) at a small cost in precision. LN achieves the highest recall, but at a steep cost in precision. Therefore, LN+XN* results in the highest F2 scores, indicating that it is the most adequate setting for TAR (Technology Assisted Review) purposes.

7 CONCLUSION

Despite recent strides in fairness, accountability, and transparency, health-related applications and recommender systems are still prone to biases amplified through data, which can perpetuate health disparities and affect patient care. To mitigate this issue, this paper

introduces a framework for detecting and diagnosing bias in the medical curriculum, focusing on the data guiding these models rather than on the models’ architecture. We use models trained and tested on instructional content annotated by medical experts for bias. We focus on bias related to sex/gender, race/ethnicity, age, and geography. Our method involves extracting non-annotated samples that contain a social identifier as negative samples for the bias classifier. For those extracted negatives, we employ word sense disambiguation to clean out any that have race/ethnicity-related terms but are not actually related to those categories.

Our findings demonstrate that while LLMs can handle many tasks, they are not well-suited for this one. Our zero- and few-shot prompting with GPT-4o mini underperformed compared to the baseline model from our previous work and scored significantly lower than the language models we tested. Similarly, using a domain-specific model like BioBERT showed no significant improvement. RoBERTa and TinyLlama were the best performers for bias detection, with RoBERTa matching the baseline and showing slight gains in precision and F1 score.

Our WSD models were highly effective at distinguishing biased excerpts from non-biased ones. ALBERT and GlossBERT nearly perfectly disambiguated sentences with race and ethnicity-related keywords. Although GPT models were comparable to BERT models, BERT consistently outperformed GPT in all metrics except recall. While this task focused on one bias category, these models could be adapted to other types with appropriate annotations. Applying WSD to bias detection in medical curricula yielded mixed results. The AUC for RoBERTa was similar to the baseline, but WSD improved both precision and F1 score.

This work could help identify potentially biased excerpts in medical curricula for review before they’re used to train models for future health-related applications and recommender systems, contributing to more equitable healthcare across all demographics.

8 DISCUSSION

Despite the encouraging results provided by our WSD and bias classification models, there are future directions we can take to enhance our project’s significance. First, in the WSD experiment, using ChatGPT-4o to generate more sentences noticeably increased the performance of our language models. Hence, it is likely that increasing the number of synthetic sentences can further enhance performance if the samples are diverse enough. LLMs often have a “temperature” parameter that can control the amount of randomness in the text generation. However, excessively high temperatures could also yield less coherent sentences.

We also want to consider how word sense disambiguation might be useful in the context of other social identifiers, such as geography (e.g., “American Heart Association” vs. “Native Americans”) and other domains where the tone of an excerpt is more important when evaluating word sense (e.g., social media).

Additionally, although LLMs like GPT models have significantly shown to be advanced in natural language processing, they also present a series of challenges [19]. Firstly, developing and training LLMs requires computational cost and can be time-consuming. So, they may be less accessible for smaller groups of researchers.

9 ACKNOWLEDGEMENTS

This material is based upon work supported in part by the National Science Foundation REU Site Grant 2349370 and the WPI STAR Program. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

- [1] Sarah E Ali-Khan, Tomasz Krakowski, Rabia Tahir, and Abdallah S Daar. The use of race, ethnicity and ancestry in human genetic research. *The HUGO journal*, 5:47–63, 2011.
- [2] Ahmad Asadi and Reza Safabakhsh. The encoder-decoder framework and its applications. *Deep learning: Concepts and architectures*, pages 133–167, 2020.
- [3] Kun Bu, Yuanchao Liu, and Xiaolong Ju. Efficient utilization of pre-trained models: A review of sentiment analysis via prompt learning. *Knowledge-Based Systems*, page 111148, 2023.
- [4] Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. Artificial intelligence, bias and clinical safety. *BMJ quality & safety*, 28(3):231–237, 2019.
- [5] Benjamin Clavié, Alexandru Ciceu, Frederick Naylor, Guillaume Soulié, and Thomas Brightwell. Large language models in the workplace: A case study on prompt engineering for job type classification. In *International Conference on Applications of Natural Language to Information Systems*, pages 3–17. Springer, 2023.
- [6] Giuseppe Colavito, Filippo Lanubile, Nicole Novielli, and Luigi Quaranta. Leveraging gpt-like llms to automate issue labeling. In *2024 IEEE/ACM 21st International Conference on Mining Software Repositories (MSR)*, pages 469–480. IEEE, 2024.
- [7] Gordon V Cormack and Maura R Grossman. Engineering quality and reliability in technology-assisted review. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 75–84, 2016.
- [8] Leonor Corsino, Kenyon Railey, Katherine Brooks, Daniel Ostrovsky, Sandro O Pinheiro, Alyson McGhan-Johnson, and Blanca Iris Padilla. The impact of racial bias in patient care and medical education: let’s focus on the educator. *MedEd-PORTAL*, 17:11183, 2021.
- [9] Erin Dehon, Nicole Weiss, Jonathan Jones, Whitney Faulconer, Elizabeth Hinton, and Sarah Sterling. A systematic review of the impact of physician implicit racial bias on clinical decision making. *Academic Emergency Medicine*, 24(8):895–904, 2017.
- [10] Shiri Dori-Hacohen, Roberto Montenegro, Fabricio Murai, Scott A Hale, Keen Sung, Michela Blain, and Jennifer Edwards-Johnson. Fairness via ai: Bias reduction in medical information. *arXiv preprint arXiv:2109.02202*, 2021.
- [11] Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: A theoretical perspective. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 70757–70798. Curran Associates, Inc., 2023.
- [12] Mark Halman, Lindsay Baker, and Stella Ng. Using critical consciousness to inform health professions education: A literature review. *Perspectives on medical education*, 6:12–20, 2017.
- [13] Debra Howcroft and Jill Rubery. ‘bias in, bias out’: gender equality and the future of work debate. *Labour & Industry: a journal of the social and economic relations of work*, 29(2):213–227, 2019.
- [14] Linda M Hunt, Nicole D Truesdell, and Meta J Kreiner. Genes, race, and culture in clinical care: racial profiling in the management of chronic illness. *Medical anthropology quarterly*, 27(2):253–271, 2013.
- [15] Reena Karani, Lara Varpio, Win May, Tanya Horsley, John Chenault, Karen Hughes Miller, and Bridget O’Brien. Commentary: racism and bias in health professions education: how educators, faculty developers, and researchers can make a difference. *Academic Medicine*, 92(11S):S1–S6, 2017.
- [16] Shawn Khan, Abirami Kirubarajan, Tahmina Shamsheri, Adam Clayton, and Geeta Mehta. Gender bias in reference letters for residency and academic medicine: a systematic review. *Postgraduate medical journal*, 99(1170):272–278, 2023.
- [17] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
- [18] Wojciech Kusa, Georgios Peikos, Moritz Staudinger, Aldo Lipani, and Allan Hanbury. Normalised precision at fixed recall for evaluating tar. In *The 10th ACM SIGIR/The 14th International Conference on the Theory of Information Retrieval*, 2024.
- [19] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.
- [20] Rajvardhan Patil, Thomas F Heston, and Vijay Bhuse. Prompt engineering in healthcare. *Electronics*, 13(15):2961, 2024.
- [21] Jhonny Pincay, Luis Terán, and Edy Portmann. Health recommender systems: a state-of-the-art review. In *2019 Sixth International Conference on eDemocracy & eGovernment (ICEDEG)*, pages 47–55. IEEE, 2019.
- [22] Shaina Raza, Muskan Garg, Deepak John Reji, Syed Raza Bashir, and Chen Ding. Nbias: A natural language processing framework for bias identification in text. *Expert Systems with Applications*, 237:121542, 2024.
- [23] Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pages 1–7, 2021.
- [24] Abhaya Kumar Sahoo, Chittaranjan Pradhan, Rabindra Kumar Barik, and Harishchandra Dubey. Deepreco: deep learning based health recommender system using collaborative filtering. *Computation*, 7(2):25, 2019.
- [25] Chiman Salavati, Shannon Song, Willmar Sosa Diaz, Scott A Hale, Roberto E Montenegro, Fabricio Murai, and Shiri Dori-Hacohen. Reducing biases towards minoritized populations in medical curricular content via artificial intelligence for fairer health outcomes. *arXiv preprint arXiv:2407.12680*, 2024.
- [26] Libby Tiderman, Juan Sanchez Mercedes, Fiona Romanoschi, and Fabricio Murai. Towards detecting cascades of biased medical claims on twitter. In *2023 IEEE MIT Undergraduate Research Technology Conference (URTC)*, pages 1–5, 2023.
- [27] Thi Ngoc Trang Tran, Alexander Felfernig, Christoph Trattner, and Andreas Holzinger. Recommender systems in the healthcare domain: state-of-the-art and research issues. *Journal of Intelligent Information Systems*, 57(1):171–201, 2021.
- [28] Jennifer Tsai, Laura Ucik, Nell Baldwin, Christopher Hasslinger, and Paul George. Race matters? examining and rethinking race portrayal in preclinical medical education. *Academic Medicine*, 91(7):916–920, 2016.
- [29] Jiaqi Wang, Enze Shi, Sigang Yu, Zihao Wu, Chong Ma, Haixing Dai, Qiushi Yang, Yanqing Kang, Jinru Wu, Huawen Hu, et al. Prompt engineering for healthcare: Methodologies and applications. *arXiv preprint arXiv:2304.14670*, 2023.
- [30] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022.
- [31] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.