ABIM Medical Bias Detection System

Development of an AI-Powered Bias Checker Agent for Clinical Vignettes and Healthcare Documentation

Project Report

January 2026

Abstract

This report documents the design and evaluation of a *Bias Checker Agent* to identify and classify bias in clinical vignettes, assessment content, and healthcare documentation in ABIM-like settings. The project progressed through two core experiments: (1) seven-class bias classification with fine-tuned transformer models on a synthetic dataset, and (2) a consolidated three-class formulation that improved separability and observed discrimination. We additionally summarize the deployed end-to-end pipeline using few-shot prompting with large language models as an explainability layer for healthcare AI audits.

Contents

1 Project Background and Motivation

AI systems used in healthcare—including clinical decision support, automated documentation, and educational assessment tools—can perpetuate and amplify inequities when trained on biased data or deployed without robust auditing. In ABIM-like environments (e.g., board examinations, certification content, and clinical vignette authoring), biased language can lead to unfair assessment and can reinforce disparities in clinical reasoning, diagnosis, and treatment.

Manual bias review is expensive, inconsistent, and may miss subtle patterns (e.g., stigmatizing framing, demographic assumptions, or different competency standards for trainees). This project therefore builds a dedicated agent that:

- detects bias in clinical text and assessment artifacts,
- classifies bias type(s) using a healthcare-focused taxonomy,
- highlights evidence spans and produces actionable rewrite recommendations,
- supports both batch auditing and interactive, human-in-the-loop review.

2 Bias Taxonomy

We grounded the work in a seven-category taxonomy designed for medical examination content and clinical documentation. The taxonomy includes demographic assumptions, stigmatizing language, assessment bias, algorithmic bias, documentation framing, and structural inequity patterns.

Table 1: Seven-category medical bias taxonomy used in Experiment 1.

Bias Category	Definition and Examples
No Bias	Neutral, clinically appropriate content. Uses evidence-based terminology, objective documentation, and patient-centered language without demographic assumptions or stigmatizing terms.
Demographic Bias	Assumptions or stereotypes based on race/ethnicity, gender, age, immigration status, socioeconomic status, or training background (e.g., IMG vs. US graduate). Examples: assuming lower adherence based on neighborhood, expecting IMGs to struggle with complex cases.
Clinical Stigma Bias	Stigmatizing language related to obesity, mental health, substance use, chronic pain, or insurance status. Examples: labeling patients as ”drug-seeking” or ”frequent flyer,” blaming obesity for all symptoms without differential diagnosis.
Assessment Bias	Biased evaluation of trainees or candidates based on communication style, accent, or cultural norms rather than competence. Examples: penalizing shared decision-making, applying different ”professionalism” standards based on background.
Algorithmic Bias	Systematic unfairness in AI/automated systems using problematic proxies or amplifying historical inequities. Examples: using healthcare spending as proxy for need (underserving Black patients), automated scoring that penalizes non-native English speakers.
Documentation Bias	Biased framing in clinical notes using labels like ”non-compliant,” ”difficult,” or ”frequent flyer” without context. Emphasizing social details in stigmatizing ways or embedding stereotypes in clinical text.
Structural Bias	Systemic inequities in healthcare policies, resource allocation, or training structures. Examples: directing resources away from safety-net hospitals, certification processes triggered more often for certain demographics.

3 Methodology Overview

3.1 Experimental Design

The research proceeded in iterative phases:

- **Experiment 1:** Generate and model a seven-class dataset to test feasibility and identify failure modes.
- **Experiment 2:** Consolidate overlapping categories into three classes and re-train to evaluate improved discrimination.
- **Experiment 3 (Deployment):** Implement an end-to-end few-shot prompting pipeline to improve generalization and add explainability.

3.2 Synthetic Dataset Generation

Because publicly available, large-scale datasets labeled for medical bias are limited, we generated a synthetic dataset of 3,500 clinically realistic samples. Each class included both *exam vignettes* and *feedback snippets* to reflect ABIM-like content sources.

Table 2: Synthetic dataset distribution across bias categories and source types (N=3,500).

Bias Category	Exam Vignettes	Feedback Snippets	Total
No Bias	250	250	500
Demographic Bias	250	250	500
Clinical Stigma Bias	250	250	500
Assessment Bias	250	250	500
Algorithmic Bias	250	250	500
Documentation Bias	250	250	500
Structural Bias	250	250	500
Total	1,750	1,750	3,500

4 Experiment 1: Seven-Class Classification (Transformer Fine-Tuning)

4.1 Setup

Two transformer architectures were compared:

- **RoBERTa-base** (general-domain transformer) [2]
- **Bio-ClinicalBERT** (clinical-domain transformer) [3]

Both models were fine-tuned using LoRA adapters [4] with regularization (dropout), standard optimization (AdamW), and multiple epochs.

4.2 Key Observation: Semantic Overlap Between Bias Types

A central finding in Experiment 1 was that the seven-class formulation introduced **semantic overlap** that caused systematic confusion:

- **Demographic vs. Algorithmic**: both frequently reference race/ethnicity and population-level disparities.
- **Clinical Stigma vs. Documentation**: stigmatizing language often appears through documentation framing.
- **Structural vs. Demographic**: structural inequities often manifest via demographic proxies.
- **Assessment vs. Stigma**: evaluation bias can use stigmatizing language patterns.

4.3 Quantifying Overlap via Similarity Analysis

To quantify overlap, within-class TF–IDF cosine similarity was computed. Low values and a narrow range across categories suggest overlapping vocabulary and limited separability.

Table 3: Within-class TF–IDF cosine similarity (higher indicates more internally consistent language).

Bias Category	Internal Similarity Score
No Bias	0.0988
Demographic Bias	0.0978
Clinical Stigma Bias	0.1063
Assessment Bias	0.1031
Algorithmic Bias	0.1463
Documentation Bias	0.1304
Structural Bias	0.1330

Interpretation. The similarity analysis supports the qualitative observation that a highly granular taxonomy can be *conceptually correct* but *operationally ambiguous* for automated classification without additional supervision signals or richer real-world labels.

5 Experiment 2: Consolidated Three-Class Classification

5.1 Rationale and Mapping

Based on Experiment 1 overlap patterns, we consolidated seven categories into three broader classes to increase separability while preserving practical meaning for auditing workflows.

Table 4: Consolidation mapping from seven classes to three classes for Experiment 2.

Consolidated Category	Original Categories Merged
No Bias	no_bias (unchanged)
Demographic Bias	demographic_bias + structural_bias + algorithmic_bias
Clinical Stigma Bias	clinical_stigma_bias + documentation_bias + assessment_bias

5.2 Results: Model Comparison

The consolidated task produced substantially improved and more stable discrimination. RoBERTa converged rapidly and achieved high macro-F1, while Bio-ClinicalBERT underperformed on this bias-detection task (suggesting that general linguistic signals mattered more than domain terminology for these labels).

Table 5: Macro F1 by epoch for RoBERTa vs. Bio-ClinicalBERT on the consolidated 3-class task.

Model	Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5
ClinicalBERT	0.307	0.508	0.609	0.668	0.630
RoBERTa	0.340	0.740	0.972	0.972	0.979

5.3 Core Observations (Experiment 1 vs. Experiment 2)

- **Improved class separability:** Consolidation reduced label ambiguity and improved discrimination.

- **Unexpected model behavior:** RoBERTa outperformed Bio-ClinicalBERT on bias detection, indicating bias cues are largely linguistic rather than clinical-jargon dependent.
- **Synthetic overfitting risk:** Near-perfect scores suggested possible shortcut learning (keywords strongly correlated with labels), motivating the shift to a generalizable approach in deployment.

5.4 Evaluation Visualizations

Figure 1 summarizes evaluation outputs (ROC curves, confusion matrix, and aggregate metrics).

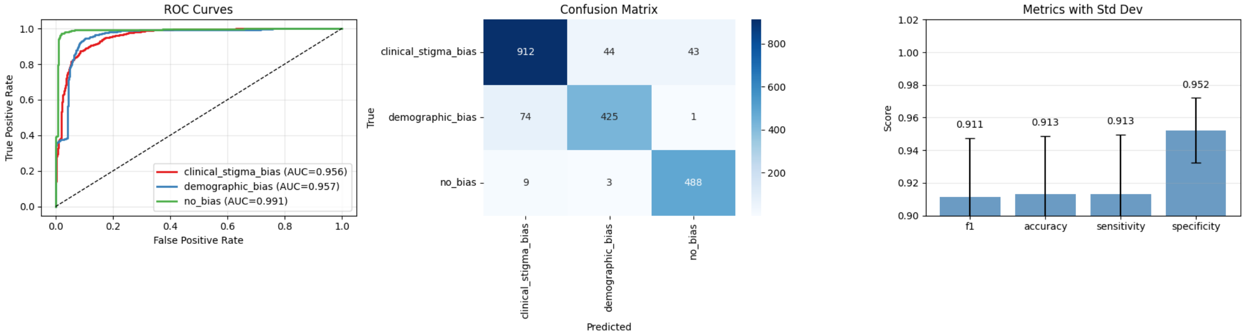
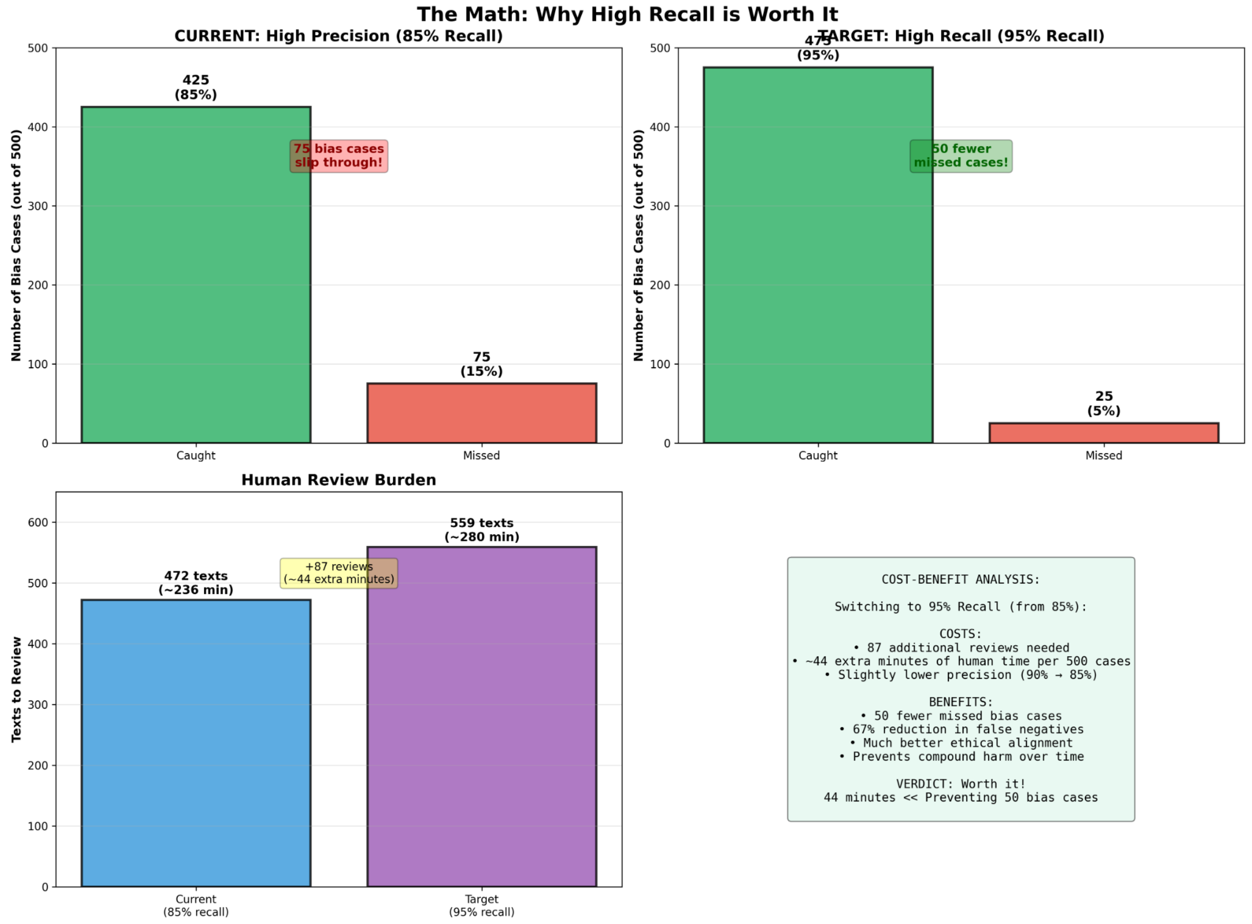


Figure 1: Experiment 2 evaluation outputs: ROC curves by class, confusion matrix, and aggregate metrics with standard deviation.

6 Operational Decision: Why High Recall Matters for Bias Auditing

Bias auditing is often **recall-sensitive**: missed bias cases can propagate harm downstream (e.g., biased exam items, biased feedback, or biased clinical documentation). The cost-benefit analysis below illustrates why increasing recall (even at some precision cost) can be worthwhile in a human-in-the-loop review workflow.



COST-BENEFIT ANALYSIS:

Switching to 95% Recall (from 85%):

COSTS:
- 87 additional reviews needed
- ~44 extra minutes of human time per 500 cases
- Slightly lower precision (90% → 85%)

BENEFITS:
- 50 fewer missed bias cases
- 67% reduction in false negatives
- Much better ethical alignment
- Prevents compound harm over time

VERDICT: Worth it!
44 minutes << Preventing 50 bias cases

Figure 2: Cost–benefit illustration: increasing recall reduces missed bias cases at the expense of additional human review.

7 Experiment 3: Few-Shot Prompting and End-to-End Pipeline (Deployment)

7.1 Motivation

Due to concerns about synthetic-data shortcut learning and the need for transparent explanations, the project implemented a few-shot prompting approach using large language models [5]. The few-shot system:

- generalizes to content beyond the synthetic distribution,
- returns structured JSON outputs for downstream UI rendering,
- provides evidence spans and rewrite recommendations (explainability layer).

7.2 Few-Shot Example Design

The prompting strategy included curated examples representing major bias patterns and an intersectional case.

Table 6: Few-shot examples used for in-context learning in Experiment 3.

Ex #	Primary Bias	Content Summary
1	Demographic (racial)	45yo African American male with chest pain; substance abuse screening prioritized based on ”patient population”
2	Clinical Stigma (pain + weight)	Patient with BMI 38 described as ”exaggerating pain”; pain attributed to ”weight and poor lifestyle choices”
3	No Bias (clean)	58yo individual with cough; comprehensive workup, collaborative treatment planning, scheduled follow-up
4	Assessment (diagnostic anchoring)	Department chief's fibromyalgia diagnosis maintained despite contradicting inflammatory markers and positive ANA
5	Intersectional (age + diagnostic + mental health)	Elderly patient labeled as having ”early dementia” based on confusion without cognitive testing

7.3 Pipeline Summary

The production pipeline follows:

1. **Input**: clinical vignette, feedback snippet, or documentation excerpt.

2. **Prompt assembly**: system prompt + few-shot examples + user content.

3. **Inference**: LLM generates structured JSON with bias types, confidence, evidence, and recommendations.

4. **Presentation**: results rendered as cards in a web UI for reviewer action.

7.4 Security Note (Key Management)

During implementation, API keys must be managed securely (e.g., environment variables, secrets managers). Keys should never be committed to notebooks or source files.

8 Conclusion

This project demonstrates a practical pathway for building a healthcare-oriented bias auditing agent:

- A comprehensive taxonomy enables structured audits, but overly granular class schemes can reduce model reliability due to semantic overlap.

- Consolidating overlapping classes improved discrimination and interpretability for real workflows.

- Few-shot prompting added generalization and explainability, enabling deployment as an end-to-end bias checking pipeline.

References

References

[1] Obermeyer, Z., et al. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.

[2] Liu, Y., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*.

[3] Alsentzer, E., et al. (2019). Publicly available clinical BERT embeddings. *arXiv:1904.03323*.

[4] Hu, E. J., et al. (2021). LoRA: Low-rank adaptation of large language models. *arXiv:2106.09685*.

[5] Brown, T., et al. (2020). Language models are few-shot learners. *NeurIPS*.

[6] FitzGerald, C., & Hurst, S. (2017). Implicit bias in healthcare professionals: a systematic review. *BMC Medical Ethics*, 18(1), 19.

[7] Gianfrancesco, M. A., et al. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine*, 178(11), 1544–1547.