

GPT-4o-mini reduced screening burden by 72-85% with excellent specificity, NPV, and sensitivity, suggesting it may be a reliable and cost-effective first-line screener for reviews.

LLMs have potential to make the review process more accessible to resource-limited and community researchers.



Authors Affiliations:
¹ Michael G. DeGroote School of Medicine, McMaster University.
² Niagara Health, Niagara Region, Ontario.

Corresponding Author: David Kanter Eivin
David.KanterEivin@medportal.ca

Reducing screening burden for systematic reviews: performance of GPT-4o-mini for title and abstract screening

David Kanter Eivin¹, Isabelle Lei¹, Julia Simone¹, Brock Browett¹, Thomas Barry¹, Sabrina Sikka¹, Calvin Armstrong¹, Kaitlin See¹, Sebastian Kolde¹, Dominic Di Stefano¹, Jolie Leung¹, Corrine Mitges², Craig Brick², Oliver Shaw², Suneel Upadhye², Stephenson Strobel².

INTRODUCTION

- Systematic reviews play an essential role in evidence-based medicine (EBM) by synthesizing high-quality evidence to guide clinical practice.
- For medical researchers conducting reviews, title and abstract screening incurs significant labor and time costs.
- Large Language Models (LLMs) offer fast, cost-effective predictions without needing training data, making them appealing to small teams and community researchers – who may not have the funding or manpower to conduct labor intensive reviews.
- Prior studies show promise in using LLMs for screening but highlight varying performance depending on the field studied, and subjectivity of inclusion criteria. They also use a variety of prompting strategies.
- Study Objective:** We evaluated GPT-4o-mini’s performance screening titles and abstracts for systematic reviews with both specific and more subjective criteria, clinically-focused and health systems research. We compared this performance to humans.

METHODS

- We analyzed 10,884 titles and abstracts from two ongoing reviews:
 - A 1260-article systematic review on red flag features of low back pain in the ED, with specific criteria, reviewed by emergency physicians (LBP).
 - A 9624-article scoping review on triage bias in ED triage, using broader criteria, reviewed by undergraduate medical students (TrB).
- GPT-4o-mini was provided the title, abstract, and inclusion/exclusion criteria, in addition to the below prompt.
- Queries were made to the instruction tuned model through the OpenAI API, with temperature = 0 for reproducibility.

System Prompt:

You are a researcher rigorously screening titles and abstracts of scientific papers for a systematic review which explores bias and prejudice in emergency department triage. Use the criteria below to evaluate the paper’s relevance. Assign a relevance score from 1 to 7 based on how likely the article is to meet all inclusion criteria without violating any exclusion criteria.

Scoring Guidelines:
1-2: Clearly not relevant. Explicitly violates exclusion criteria or has no relevance to the question.
3: Likely not relevant. Lacks key inclusion criteria but not entirely irrelevant.
4: Ambiguous. Insufficient detail in the title/abstract or conflicting signals, cannot determine either way.
5-6: Likely relevant. Meets most inclusion criteria but may have minor uncertainties or ambiguities.
7: Clearly relevant. Meets all inclusion criteria with no exclusion criteria violated.

Rules:
- Assign scores strictly based on the information provided in the title and abstract, and the main idea of the paper.
- If the title or abstract is "MISSING," evaluate based on the available content and assign an appropriate score.
- Only output a single numerical score (1-7) without additional text, explanation, or rationale.

- Articles were then classified based on the score predicted by GPT-4o-mini:
 - ≤2 = exclude (negative classification), ≥6 = include (positive classification) 3–5 = ambiguous (no classification made)
- Performance was calculated relative to a “gold standard” of double-screened articles, with conflicts resolved by consensus (LBP) or expert reviewer (TrB).

RESULTS

- Based on Likert-scale classification, **GPT-4o-mini classified 72.4% and 85.1% of articles with high accuracy** (99.12% and 99.46%) in the LBP and TrB reviews, respectively; the remainder were “ambiguous”.
 - F1 scores for classified articles were 0.81 and 0.75 for LBP and TrB, suggesting **good overall model performance**.

Review	% Classified	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 Score
Lower Back Pain (LBP) – High-Sensitivity	72.4%	99.12%	0.922	0.922	0.85	0.996	0.81
ED Triage Bias (TrB) – High Sensitivity	85.1%	99.46%	0.982	0.996	0.58	0.999	0.75

Table 1: Summary Predictive Statistics for Likert Score Classification using GPT-4o-mini.

- Likert score classification had AUC > 0.9 which is considered very strong and reinforcing a threshold of 5-6 for inclusion (**Figure 3**).
- Sensitivity (as shown in Figure 1) was good; however, given the importance of high-sensitivity in this use cost—even at the cost of PPV—we include a “high sensitivity” model where ambiguities are classified as included.
- Mean Likert scores correlated to human-rated relevance for included (4.6 and 5.1), conflict (2.9 and 3.5), and excluded (1.78 and 1.44) groups.
- Inter-rater agreement was 0.94 (κ = 0.54), which is comparable to GPT’s performance.

CONCLUSIONS

- GPT-4o-mini demonstrated promising results, reducing the screening burden by 72-85% with excellent specificity, NPV, and sensitivity.
- While variability in human adherence to criteria affected PPV, inter-rater concordance (κ = 0.54, 94% agreement) shows comparable performance to human reviewers.
- Subjective review of conflicts between LLMs and human-reviews reveals that the “gold standard” may not be golden; often the articles subtly meet exclusion criteria.
- These findings support the utility large language models as a valuable, cost-effective, and reliable tool for systematic review screening in emergency medicine, offering significant workload reduction with minimal compromise on accuracy.
- Further work is planned to determine what factors affect model performance, and what techniques can be employed to optimize reliability.

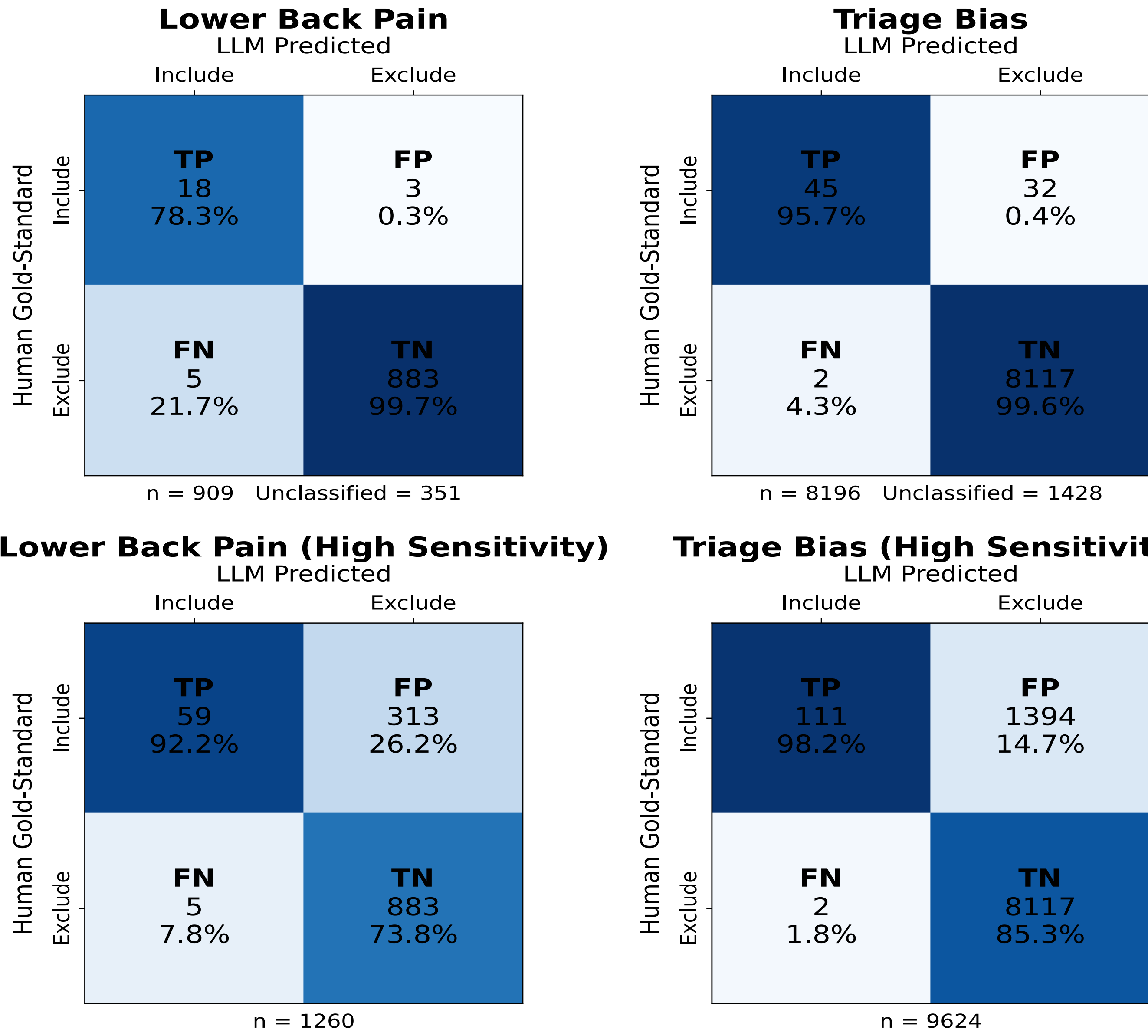


Figure 1: Confusion Matrices for Lower Back Pain and Triage Bias Reviews. In the initial variant, ambiguous (no classification made) are not considered as inclusions, leading to a lower-appearing TPR. In practice, unclassified articles need further review, leading to the matrix shown in the High Sensitivity variant.

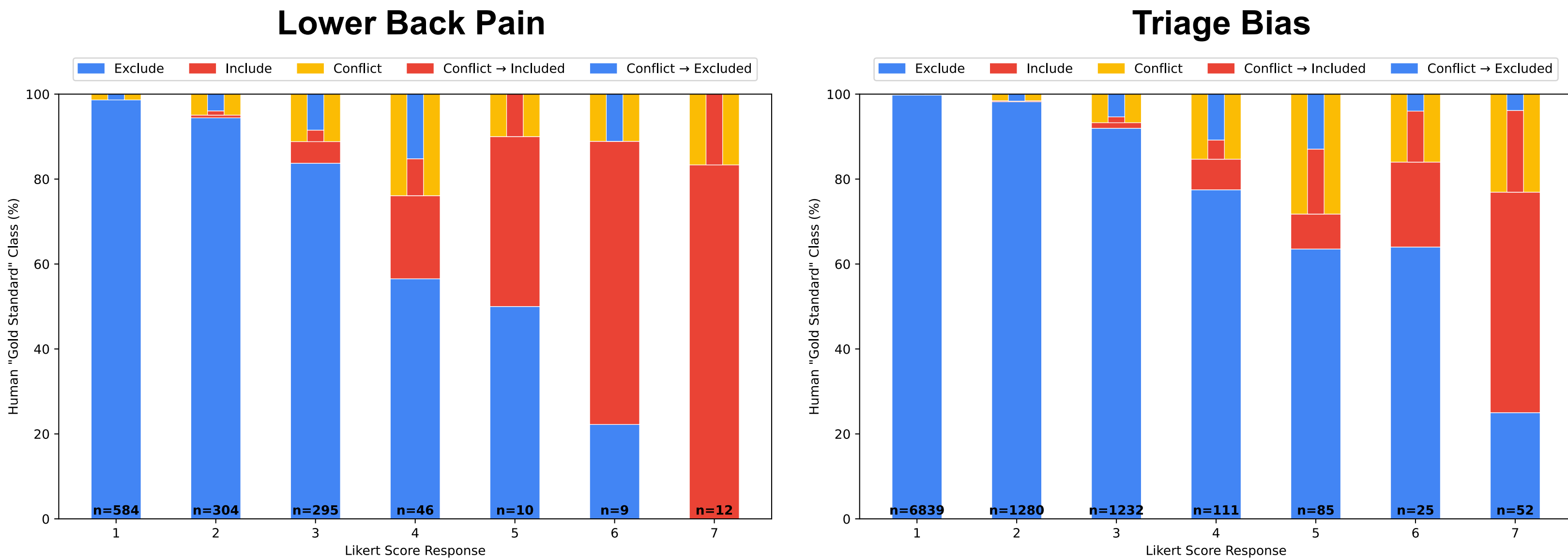


Figure 2: Proportions of Human Classifications for GPT-4o-mini predicted Likert score classes. The inner bar displays the final human classification of conflicts; thus, together the red and blue makeup the “gold standard”.

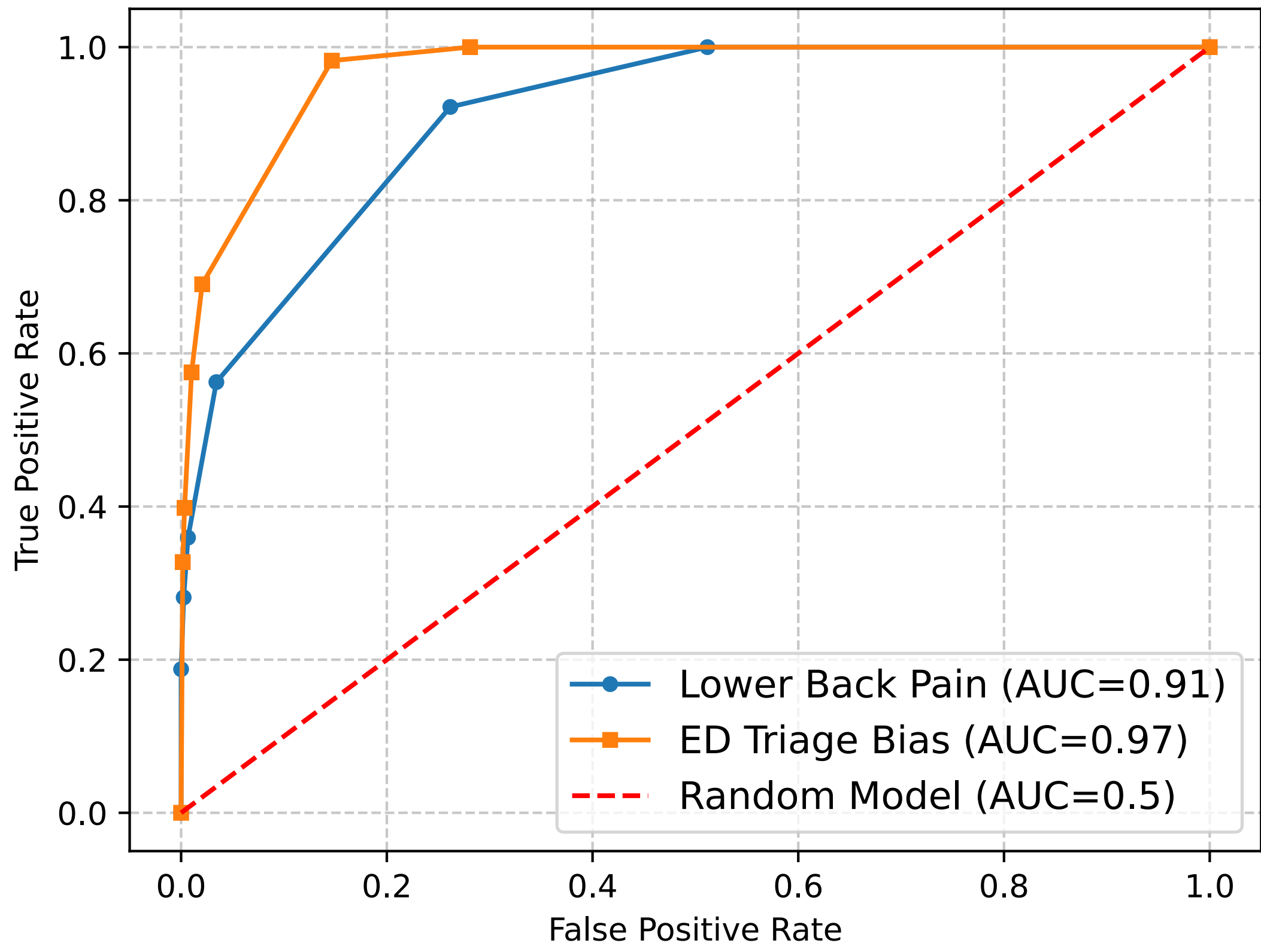


Figure 3: Receiver Operating Characteristics (ROC) Curves and Area Under the Curve for Likert Prediction Model. Each point represents the balance of TPR/FPR at each threshold of Likert-score for an article to be included. AUC scores above 0.9 are considered excellent.