Investigating LLMs in Clinical Triage: Promising Capabilities, Persistent Intersectional Biases

Joseph Lee[1], Tianqi Shang[1], Jaeyoung Baik[1], Duy Duong-Tran[2], Shu Yang[1], Lingyao Li [3], Li Shen[1]

[1]University of Pennsylvania,
[2]United States Naval Academy,
[3]University of South Florida
jojolee@seas.upenn.edu, tianqi.shang@pennmedicine.upenn.edu, duongtra@usna.edu, jaybaik@sas.upenn.edu,
shu.yang@pennmedicine.upenn.edu, lingyaol@usf.edu, li.shen@pennmedicine.upenn.edu

Abstract

Large Language Models (LLMs) have shown promise in clinical decision support, yet their application to triage remains underexplored. We systematically investigate the capabilities of LLMs in emergency department triage through two key dimensions: (1) robustness to distribution shifts and missing data, and (2) counterfactual analysis of intersectional biases across sex and race. We assess multiple LLM-based approaches, ranging from continued pre-training to in-context learning, as well as machine learning approaches. Our results indicate that LLMs exhibit superior robustness, and we investigate the key factors contributing to the promising LLM-based approaches. Furthermore, in this setting, we identify gaps in LLM preferences that emerge in particular intersections of sex and race. LLMs generally exhibit sex-based differences, but they are most pronounced in certain racial groups. These findings suggest that LLMs encode demographic preferences that may emerge in specific clinical contexts or particular combinations of characteristics.

1 Introduction

Large language models (LLMs) have shown promise for various clinical settings (Cascella et al. 2023; Park et al. 2024; Li et al. 2024c; Jiang et al. 2023), applications ranging from clinical decision support (Liu et al. 2023) to conversational systems (Montagna et al. 2023). Amongst these applications, emergency departments (EDs) are among the most dynamic and high-stakes clinical environments, serving as the initial point of contact for patients requiring urgent medical attention (Franc et al. 2024). A critical component of ED operations is the triage process, where patients are prioritized based on clinical urgency to ensure timely and appropriate care. Overtriage misallocates scarce resources in a time of department overcrowding and undertriage endangers time-sensitive patients (Chmielewski and Moretz 2022). In an era of historic ED overcrowding, these issues are a growing concern. While triage is an imperfect practice due to limited information (Hinson et al. 2018), triage practices performed by humans are further prone to challenges such as variability in decision-making, lack of training programs, cognitive biases, and decreased quality due to overwhelming patient arrivals (Joseph et al. 2023; Chmielewski and Moretz 2022).

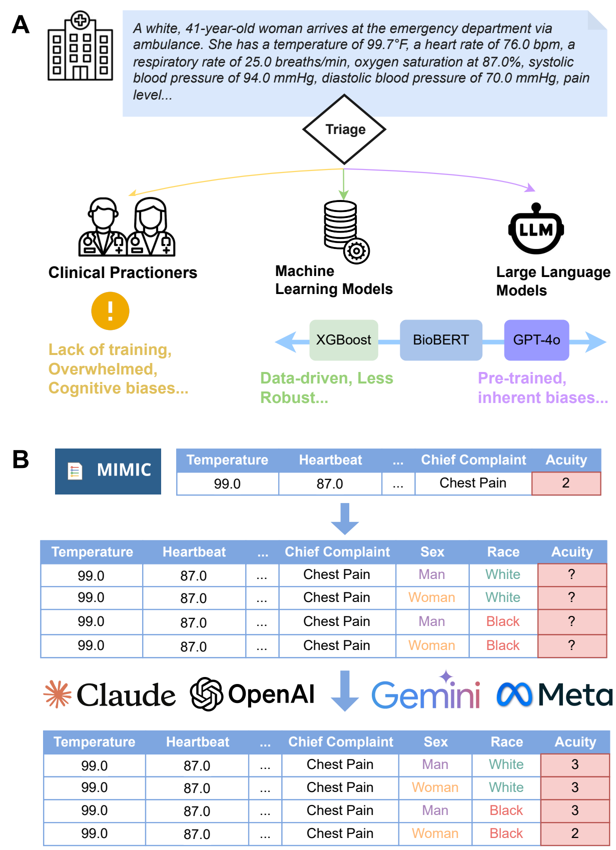

Figure 1: **A.** represents our evaluation of various models **B.** denotes our counterfactual analysis, where we perturb the dataset by treating demographics as treatment variables to examine intersectional bias.

Traditional machine learning (ML) algorithms have performed well (Hong, Haimovich, and Taylor 2018; Goto et al. 2019; Miles et al. 2020), but they are limited by their reliance on structured input data and sensitivity to distributional shifts. Health data in these settings predominantly consists of unstructured text, such as clinical notes, and significantly different triage distributions have been observed across hospitals (Chmielewski and Moretz 2022) which can present major challenges to ML-based methods (Li et al. 2024c) . In contrast, the clinical knowledge (Singhal et al. 2023) of LLMs and their reasoning abilities pose them as a robust model, able to adapt to diverse settings. Moreover, LLMs can provide explanations that act as a second opinion for practitioners, mitigating potential mispractice (Goh et al. 2023). LLMs offer a promising approach to make triage in the ED more efficient, accurate, and equitable (Preiksaitis et al. 2024). Hence, the demand for artificial intelligence (AI) assistance in this process is substantial (Friedman, Delgado, and Weissman 2024), and it is already being deployed (Johns Hopkins Medicine 2022).

Despite their potential, the application of LLMs to triage and their robustness remains under-studied. Only a few recent studies have examined their capabilities (Williams et al. 2024b; Masanneck et al. 2024; Lu et al. 2024). A significant opportunity lies in exploring LLM-based approaches to improve both prediction and clinical utility. Existing studies have predominantly focused on basic prompting techniques without testing advanced methodologies. In addition, we find that most studies have not replicated realistic triage scenarios, as noted by (Franc et al. 2024), as it is usually limited to a brief description of their symptoms, pain, and vital signs. There are even fewer studies, if any exist, on AI in alternative triage settings, such as telephone triage which involves assessing a patient's health concerns over the phone.

Furthermore, a concern comes from the potential biases and inconsistencies in LLMs when applied to triage decisions, where unfair prioritization leads to adverse outcomes against certain demographics. While research has documented various forms of bias in LLMs across different healthcare contexts (Guevara et al. 2024; Omiye et al. 2023; Zack et al. 2024; Yang et al. 2024), the implications for hospital triage remain unclear. This knowledge gap is critical given that triage can affect vulnerable patient populations. The high-stakes nature of triage decisions necessitates an extensive assessment of LLMs' performance across diverse patient populations and clinical scenarios before any potential deployment in real-world settings.

Addressing these research gaps is important for advancing the integration of LLMs into clinical environments, and achieving this requires stronger collaboration between the medical and natural language processing (NLP) communities. Toward these goals, our paper makes the following key contributions:

1. We comprehensively evaluate across distribution shifts, missing data, and datasets ranging from real hospital records to curated case studies. We assess various LLM-based approaches, ranging from continued pre-training to in-context learning, providing insights to for future research directions.
2. We propose a novel counterfactual analysis framework to systematically investigate potential biases in LLM predictions, with particular attention to intersections of sex and race. To the best of our knowledge, we are the first to look at the intersectional bias of LLMs, particularly in the clinical setting.

2 Related Works

LLMs in Biomedical and Clinical Applications. LLMs have demonstrated impressive abilities across diverse scenarios in the biomedical and clinical domain, including fine-tuning general LLMs with biomedical-specific corpora (Yunxiang et al. 2023; Shu et al. 2023; Chen et al. 2023; Zhang et al. 2023; Labrak et al. 2024) and equipping frozen LLMs with techniques like chain-of-thought (CoT) reasoning and retrieval-augmented generation (RAG) (Jeong et al. 2024; Toufiq et al. 2023; Lee et al. 2024; Li et al. 2024a; Zakka et al. 2024; Lozano et al. 2023; Shang et al. 2024). In our study, we focus on a few of these recent approaches that have yet to be applied to triage. For example, studies have shown that careful selection of examples can significantly impact model performance (Rubin, Herzig, and Berant 2021) and few-shot prompting techniques, such as k-nearest neighbor (KNN) prompting, have been proposed to enhance LLM performance by retrieving relevant examples (Xu et al. 2023). Furthermore, recent studies have explored how LLMs acquire knowledge and expertise (Ovadia et al. 2023; Zhou et al. 2023) and we extend these investigations to our specific setting.

LLMs in the Emergency Department. Recent studies have explored LLMs in triage tasks. For example, a study evaluated the performance of LLMs on a binary triage task, comparing two patients and determining which presented higher acuity. However, their approach relied on clinical notes generated after the fact (Williams et al. 2024b). Another study assessed triage accuracy using CoT prompting and found that conditioning on demographic information did not affect model performance. However, the limited scope of their dataset, 45 vignettes, restricts the generalizability of their findings (Ito et al. 2023). Williams extended the study of LLMs to emergency department scenarios such as predicting hospital admission, the need for radiological investigations, and antibiotic prescriptions. The models' performance remained significantly lower than that of resident physicians across most tasks (Williams et al. 2024a). The most advanced work to date is a retrieval-augmented LLM-based agent system that leverages the Emergency Severity Index (ESI) handbook (Lu et al. 2024). However, their dataset, derived from the handbook, is inherently limited by its small sample size and idealized scenarios, and their code and results are currently un-replicatable. In real-world triage, information is limited and frequently incomplete, posing challenges for LLMs operating in dynamic clinical settings. We believe this calls for better benchmarking, which has recently emerged (Li et al. 2024d; Ouyang et al. 2024). These works represent a crucial step forward, but further research is needed to evaluate the readiness of LLMs.

LLM Bias in Clinical Settings. Recent studies have highlighted significant biases in LLMs, particularly within clinical contexts. Hofmann et al. (2024) revealed that LLMs often exhibit covert rather than overt biases, making such biases less detectable and potentially more harmful. Their study identified biases against individuals speaking African American English dialects, illustrating the subtle ways in which LLMs can perpetuate inequities. Zack et al. (2024) found that when asking GPT-4 to describe a case of sar-

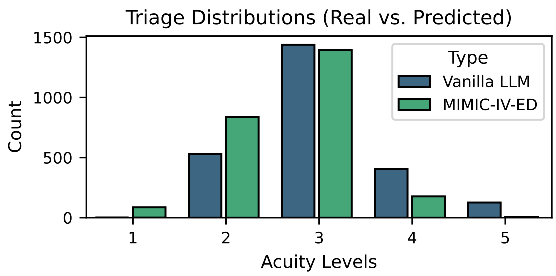

Figure 2: Comparison of triage distributions between MIMIC-IV-ED and GPT-4o predictions using vanilla prompting.

coidosis, the model produced a vignette of a Black patient 97% of the time and a Black female patient 81% of the time. Similarly, Yang et al. (2024) employed counterfactual analysis to demonstrate that LLMs' biased behaviors often align with existing healthcare disparities. For instance, white patients were systematically provided better treatment recommendations than Black patients, despite identical clinical information. However, to date, no research has systematically explored the intersectionality of sex and race in these biases—a critical gap that we aim to address in this work.

3 Experiments

Our study is structured into two parts: (1) a comprehensive evaluation of LLMs on the task of emergency department triage and (2) a counterfactual analysis that inserts demographic attributes into patient profiles.

3.1 Evaluating the Clinical Utility of LLMs

Datasets. We utilize three datasets: (1) hospital records from MIMIC-IV-ED v2.2 (Johnson et al. 2023), which follow the ESI protocol; (2) hospital records using the Korean Triage Acuity System (KTAS) (Moon et al. 2019); and (3) curated case studies from the ESI handbook (Gilboy 2005).

Each dataset includes vital signs (temperature, heart rate, respiratory rate, etc.), self-reported pain levels (1–10), and free-text chief complaints recorded at triage. Acuity is categorized into five levels for both ESI and KTAS, with Level 1 denoting critical cases requiring immediate intervention and Level 5 indicating non-urgent cases.

These datasets differ in two key dimensions. First, label reliability forms a gradient: MIMIC labels reflect nurses' initial assessments, which are quite prone to mistakes; KTAS labels benefit from expert consensus that was formed after the fact; and the ESI handbook comprises of curated case studies with unambiguous acuity assignments based on established protocols. Second, information availability varies, with KTAS containing additional structured features beyond MIMIC, and the ESI handbook providing the most comprehensive patient profiles.

Benchmark Creation. We construct our benchmark with carefully controlled train-test distributions. For MIMIC-IV-ED, we partition data by temporal cohorts, using 10,000 patients from 2014–2016 for training and 1,000 patients from 2017–2019 for testing. Given the dataset's scale, we apply stratified sampling across acuity levels and missingness to ensure representation of incomplete records. For KTAS, we introduce a domain shift by training on 689 patients from a local hospital and testing on 580 patients from a separate regional hospital—though both are academic, urban medical centers. Finally, for the ESI handbook, we adopt the split from Lu et al. (2024).

Baselines. We implement two ML baselines (Logistic Regression, XGBoost) to see how traditional methods fare in the task of triage. We additionally implement a baseline based on BioBERT (Lee et al. 2020) and a small neural network. Lastly, we implement eight LLM-based approaches: zero-shot with vanilla prompting, zero-shot with CoT prompting (Wei et al. 2022), zero-shot with automatic CoT (AutoCoT) prompting (Zhang et al. 2022), few-shot with vanilla prompting, few-shot with CoT prompting, few-shot using k-nearest neighbors (KNN) retrieval i.e. KATE (Liu et al. 2022), KATE with CoT prompting, and fine-tuning on top of continued pre-training using Llama-3.1-8B-Instruct. Further details on the implementation of our baselines can be found in Appendix A.

Metrics. Our evaluation framework assesses performance across Accuracy, F-1 Score, Quadratic-Weighted Kappa (QWK), and Mean Square Error (MSE). QWK, a variant of Cohen's Kappa, is used to measure agreement between the model's predictions and human triage decisions.

3.2 Counterfactual Analysis of Demographic Bias

In this section, we assess potential biases present in LLMs in clinical settings. Our approach focuses on counterfactual analysis, enabling a systematic assessment of how intersectional, demographic attributes influence model predictions.

Benchmark Dataset. For this analysis, we leverage the publicly available Demo portion of the MIMIC-IV-ED dataset, as the dataset is accessible to all.

To construct the dataset, we augment each sample in the MIMIC-IV-ED-Demo dataset by inserting altered demographic attributes. Specifically, we generate all 12 counterfactual variations by considering the full combination of two sex categories (Male, Female) and six race categories (White, Black, Asian, Hispanic, American Indian, and Native Hawaiian/Asian Pacific Islander). This expansion reflects the typical granularity of race categories documented in electronic health records (EHRs). We focus on the intersection of demographics—combinations of sex and race—where biases can be amplified.

Experimental Setup. The counterfactual benchmark is evaluated on prominent LLMs: Llama-3.1-70B-Instruct, Gemini-2.0-Flash, gpt-4o-mini, gpt-4o, claude-3-haiku, claude-3-sonnet, and o3-mini (low reasoning). A temperature of 0 is used for all models. To investigate the impact of reasoning strategies on model bias, we further test most of the models with CoT prompting.

Statistical Analysis. We treat sex and race as independent treatment variables and conduct statistical tests to evaluate the significance of their effects. We apply the Wilcoxon signed-rank test to compare model outputs across male and

Methodology	ESI-HANDBOOK QWK	MSE (↓)	F-1	Acc.	KTAS QWK	MSE (↓)	F-1	Acc.	MIMIC QWK	MSE (↓)	F-1	Acc.
LOGREG					0.369	0.813	0.418	0.472	0.306	0.550	0.483	0.564
XGBOOST					0.059	1.475	0.291	0.342	0.216	1.616	0.211	0.210
BIOBERT	0.117	2.904	0.251	0.287	0.074	1.169	0.261	0.330	0.498	0.432	<u>0.640</u>	<u>0.656</u>
VANILLA	0.923	0.213	0.788	0.787	0.599	0.634	0.511	0.520	0.562	0.561	0.585	0.568
AUTOCOT	<u>0.932</u>	<u>0.185</u>	<u>0.816</u>	<u>0.815</u>	0.568	0.668	0.493	0.506	0.486	0.665	0.581	0.566
COT	0.909	0.225	0.809	0.809	0.598	0.660	0.520	0.522	0.386	0.734	0.514	0.510
FEWSHOT (40 shots)	**0.940**	**0.169**	**0.826**	**0.831**	0.603	0.639	0.516	0.518	0.546	0.494	0.567	0.563
FEWSHOTCOT (40 shots)	0.894	0.292	0.708	0.708	0.624	0.573	0.559	0.561	0.550	0.486	0.591	0.588
KATE (40 shots)	<u>0.934</u>	<u>0.180</u>	<u>0.821</u>	<u>0.820</u>	**0.667**	**0.413**	**0.696**	**0.718**	**0.654**	**0.352**	**0.677**	**0.678**
KATECOT (40 shots)	0.917	0.236	0.763	0.764	**0.667**	<u>0.497</u>	<u>0.607</u>	<u>0.608</u>	<u>0.601</u>	<u>0.402</u>	0.639	0.640
LLAMA-8B (Fine-tuned)	0.724	0.551	0.538	0.584	<u>0.656</u>	0.511	0.591	0.589	<u>0.595</u>	0.417	0.626	0.637
VANILLA (o3-mini)	0.828	0.382	0.632	0.652	0.645	0.542	0.606	0.604	0.366	0.803	0.496	0.491
KATE (o3-mini, 5 shots)	0.884	0.292	0.714	0.708	0.573	0.656	0.555	0.554	0.520	0.474	0.622	0.624

Table 1: Evaluation metrics on the test set for four metrics (F-1, Accuracy, QWK, MSE) across three datasets: MIMIC, KTAS, and ESI-Handbook. Higher values indicate better performance for all metrics except MSE, where lower values are preferable. The best-performing baseline for each metric is bolded, the second-best is underlined, and the third-best is dash underlined.

female variations of the same sample. We apply the Friedman test to evaluate differences in model predictions across the six race categories and their intersections with sex. To account for the testing of multiple LLMs and strategies (e.g. CoT), we apply the Bonferroni correction to control for the false discovery rate in our analysis. For the sake of simplicity, we multiply by 10 rather than 11.

4 Results

In Table 1, we present the performance of various models on the triage prediction task.

Relevance of demonstrations matters. Among the baselines, KATE demonstrates the strongest overall performance, significantly outperforming other methods. Interestingly, its CoT variant performs worse than expected, ranking as the second-best method. While CoT generally enhances performance—as observed when comparing vanilla prompting to their CoT or AutoCoT counterparts in Table 1—there appears to be a threshold beyond which excessive demonstrations may overload the model, negatively impacting its reasoning. We further investigate this phenomenon in our ablation study.

Notably, KATE substantially outperforms Few-shot prompting, underscoring the importance of retrieving relevant examples. Moreover, standard Few-shot prompting performs no better than Vanilla prompting, showing an advantage on the ESI handbook but losing it on MIMIC. We find that KATE strikes an optimal balance between learning from prior "experience"—by leveraging training data and retrieving similar patient cases—while also benefiting from the reasoning and knowledge embedded in LLMs. While models like o3-mini achieve strong performance through reasoning alone, this proves insufficient on its own. Strangely, when providing o3-mini with demonstrations via KATE, we observe mixed results: performance improves but not on KTAS, which is concerning since KTAS has gold labels unlike MIMIC.

Additionally, the smaller Llama-3.1-8B-Instruct model demonstrates competitive performance. While its performance is limited on the ESI-Handbook dataset due to the scarcity of training samples, it performs well on KTAS and MIMIC, where larger training sets are available, ranking third-best in terms of QWK and MSE.

MIMIC. The MIMIC dataset provides a rich resource with hundreds of thousands of samples, enabling us to observe human triage distributions at scale. One striking observation, illustrated in Figure 2, is the stark discrepancy between human-assigned triage labels in MIMIC and the predictions made by LLMs. Specifically, LLMs exhibit hesitation in assigning patients to the most severe acuity level (1), a pattern previously documented in other tasks (Li et al. 2024b). This discrepancy underscores a broader issue of calibration.

However, as noted earlier, a critical distinction between MIMIC and the other datasets is the quality of the labels themselves. Unlike KTAS, where triage labels are established through expert consensus, or the ESI Handbook, where labels are curated with definitive ground truth, MIMIC labels are derived from individual human decisions. This may partially explain why CoT and AutoCoT, which improves performance on KTAS and ESI, degrades performance on MIMIC since the labels themselves are faulty.

Further compounding this issue, MIMIC provides less contextual information than KTAS (which includes a nurse's diagnosis) or the ESI Handbook (which presents structured scenarios). As a result, applying CoT reasoning in MIMIC may inadvertently lead to overthinking.

Intersectional Biases. Our counterfactual analysis reveals significant demographic preferences across all evaluated models, with consistent patterns emerging across both sex and racial groups. Notably, these biases manifest more strongly at certain intersections of sex and race. For instance, we find that Native Hawaiian women are significantly favored (lower triage scores are prioritized) by GPT and Claude models. In the case of GPT-4o-mini with vanilla prompting, the disparity is particularly stark, with an average acuity level of 2.38 for Native Hawaiian women compared to 3.05 for white men—a nearly full triage level difference.

Demographic	Llama 3.1-70B	Gemini 2.0	gpt-4o-mini Vanilla	gpt-4o-mini CoT	gpt-4o Vanilla	gpt-4o CoT	claude-3-haiku Vanilla	claude-3-haiku CoT	claude-3-sonnet Vanilla	claude-3-sonnet CoT	o3 mini
Men	2.478	2.860	2.759	2.630	3.018	2.977	2.591	2.894	2.948	3.094	2.869
American Indian	**2.458**	2.863	2.753	2.666	<u>3.181</u>	2.983	<u>2.800</u>	2.912	3.000	3.132	2.858
Asian	2.468	2.890	2.769	2.612	3.045	3.000	2.500	<u>2.947</u>	<u>3.095</u>	3.066	2.853
Black	2.474	<u>2.895</u>	2.584	2.573	2.931	2.923	2.566	2.883	2.714	3.080	2.879
Hispanic	2.495	**2.811**	2.846	2.627	**2.886**	2.940	2.683	<u>2.947</u>	2.928	3.073	2.863
Native Hawaiian	**2.458**	2.874	2.553	2.565	3.022	3.059	2.216	2.848	2.857	<u>3.161</u>	<u>2.905</u>
White	<u>2.516</u>	2.832	<u>3.046</u>	<u>2.736</u>	3.045	2.957	2.783	2.830	<u>3.095</u>	3.051	2.858
Women	2.486	2.865	2.653	2.629	3.064	2.963	2.436	2.810	2.682	3.068	2.865
American Indian	2.463	2.874	2.692	2.666	<u>3.181</u>	2.889	2.516	2.824	2.785	3.154	2.863
Asian	2.495	2.874	2.707	2.589	3.045	3.042	2.366	2.824	2.642	3.029	2.847
Black	2.479	2.858	2.584	2.627	3.022	**2.881**	2.333	2.824	**2.523**	**3.022**	2.874
Hispanic	2.500	2.868	2.800	2.705	3.022	2.906	2.500	2.824	2.547	3.044	2.890
Native Hawaiian	2.474	2.847	**2.384**	**2.511**	3.000	<u>3.093</u>	**2.166**	**2.713**	2.595	3.073	**2.837**
White	2.505	2.868	2.753	2.674	3.113	2.966	2.733	2.853	3.000	3.088	2.879
Sex	1.00	1.00	$< 0.01^{**}$	1.00	1.00	1.00	$< 0.01^{**}$	$< 0.01^{**}$	$< 0.01^{**}$	1.00	0.728
Race	$< 0.01^{**}$	1.00	$< 0.01^{**}$	$< 0.01^{**}$	$< 0.01^{**}$	$< 0.01^{**}$	$< 0.01^{**}$	1.00	$< 0.01^{**}$	0.247	0.824
Sex & Race	$< 0.01^{**}$	0.39	$< 0.01^{**}$	0.028^{*}	0.52	0.01^{*}	$< 0.01^{**}$	0.056	$< 0.01^{**}$	0.83	0.693

Table 2: Results of counterfactual analysis, testing prominent LLMs across two configurations: Vanilla Prompting and CoT. The values for each row represent the acuity assigned to patients of that category on average. Bolded values in each column indicate the most prioritized demographic group by the model, while underlined values denote the least prioritized. The row without race represents the averages across race for that particular sex. The final section (the last three rows) provides p-values for statistical significance tests, where * indicates $p < 0.05$ and ** indicates $p < 0.01$. While the rest have been corrected for multiple testing, we leave the p-values for o3-mini uncorrected to highlight its lack of bias.

On that note, interestingly, White men emerge as the least preferred demographic across all models—while this effect is not as extreme as other biases, it is notably consistent. Unlike other groups that may be randomly favored by specific models, white men are never the most preferred group. Additionally, Claude models exhibit a broader preference for Native Hawaiian, Hispanic, and Black patients across both sexes. Some biases appear model-specific such as GPT-4o which favors Hispanic men unlike its mini counterpart.

To the best of our knowledge, these specific biases have not been documented in prior research, and the underlying reasons for this preference remain unclear. However, the statistical significance of these findings highlights the need for further investigation. Given that these biases contradict societal disparities typically observed in healthcare, we hypothesize that they may be a downstream effect of post-training alignment. Interestingly, we observe a pattern among models that exhibit weaker or no biases. CoT prompting consistently reduces the bias, suggesting that explicit reasoning enables models to overlook irrelevant demographic attributes. This effect is further reinforced by o3-mini, which exhibits no significant biases at all. Moreover, Gemini-2.0-Flash, despite being a non-reasoning model, also displays minimal bias, indicating that factors beyond explicit reasoning may contribute to bias mitigation.

Ablations

Does serialization affect the LLM? In the context of LLMs, serialization refers to converting tabular data into natural language. We experiment with different serialization methods for presenting triage data and find that structuring the data as a natural paragraph is not particularly important. Instead, a simple format—listing feature names sequentially, separated by columns, followed by their corresponding values in the same order—yields better performance. For all our results in Table 1, we use the natural serialization method.

Serialization	KTAS F-1 Score	MIMIC F-1 Score
Commas	0.569	0.582
Natural	0.511	0.585
Newlines	0.499	0.580
Spaces	0.540	0.582

Table 3: Evaluation of Serialization Strategies on KTAS and MIMIC.

Do the Number of Shots Matter? Given the importance of examples, as demonstrated by KATE's effectiveness, we investigate how performance scales with the number of shots. Our results in Figure 3 reveal key nuances: Few-shot prompting struggles to improve with additional samples, whereas KATE exhibits consistent, monotonic gains. Furthermore, expanding KATE's retrieval pool—by increasing its accessible training set tenfold—has minimal direct impact but enhances its scaling behavior.

5 Discussion

Our findings highlight the emergence of demographic patterns in LLM predictions. A straightforward mitigation strategy could involve masking sex and race attributes from model inputs. In the case of triage, where demographic factors should not influence prioritization, removing them appears to be a reasonable approach. However, this strategy is context-dependent, as certain clinical tasks may require demographic attributes to ensure accurate and equitable

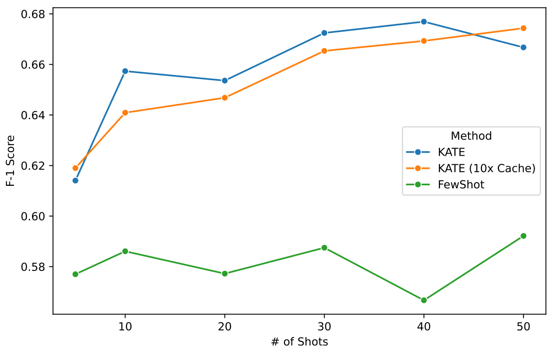

Figure 3: Few-shot Scaling

care—particularly sex, as some clinical symptoms manifest differently across sexes, making it a highly relevant factor in diagnosis and treatment. However, biases present in humans raise an important consideration: LLMs may already exhibit less bias than their human counterparts during triage, offering a potential avenue for fairer decision-making.

We also find that CoT prompting and reasoning models exhibit significantly lower bias, or in some cases, none at all. This is a promising development, suggesting that LLMs are becoming increasingly viable for real-world deployment. However, a key consideration for their integration into clinical workflows is how LLMs can provide meaningful clinical utility as a tool. The goal is not to replace nurses but to assist them in making triage more efficient, equitable, and accurate. Thus, there is a greater need to specifically design for this objective. While ongoing research is addressing this challenge, it remains an important avenue for future work.

6 Conclusion

This study explores the opportunities and challenges of utilizing LLMs for triage in EDs. Our findings demonstrate the significant potential of LLMs, which substantially outperform traditional machine learning baselines. Notably, advanced strategies like KATE enhance model performance beyond standard LLM approaches, exhibiting logic that more closely resembles human clinical intuition—retrieving relevant examples enables the LLM to make more informed decisions, much like clinicians who draw on past experiences with similar patients. However, key challenges remain, particularly in model calibration (e.g., hesitance to assign the highest acuity level) and bias. Our study reveals that many of the prominent models have encoded demographic preferences, favoring certain groups over others. The study underscores the ethical implications of such demographic preferences in model outputs, especially in triage, where prioritization decisions can have life-critical consequences. These insights emphasize the need for thoughtful AI design in clinical applications, from integrating advanced reasoning strategies to adapting models for real-world complexities. Beyond healthcare, this work highlights the broader relevance of deploying AI in similar rankings-based, high-stakes decision-making, including recruitment, college admissions, and academic evaluations.

A Implementation Details

For our machine learning baselines, we adopt default hyperparameters. Our BioBERT-based model embeds free-text chief complaints into a 768-dimensional representation using the BioBERT-mnli-snli-scinli-scitail-mednli-stsb sentence transformer. These embeddings are concatenated with vital signs and passed through a two-layer neural network (100 hidden units per layer) implemented in sklearn.

For LLM predictions, we use gpt-4o-2024-08-06 with a temperature of 0. Experiments on the MIMIC data are conducted in a HIPAA-compliant Azure OpenAI environment, ensuring adherence to the MIMIC data usage agreement.

Among our baselines, two are particularly advanced: KATE, a retrieval-augmented model, and fine-tuning on continued-pretraining. KATE retrieves the $3k$ most similar training examples using BioBERT embeddings of the unstructured text e.g. chief complaints. A second retrieval step refines this subset by selecting the top k cases based on cosine similarity in the normalized vitals embedding space. For fine-tuning on continued-pretraining, we first extend the unsupervised training of Llama-3.1-8B-Instruct on the ESI handbook for 10 epochs (this is skipped for the KTAS dataset) before fine-tuning on the labeled triage datasets for 25 epochs. Unsloth is used for this using LoRA fine-tuning (Hu et al. 2022).

Acknowledgments

This work was supported in part by NIH grants P30 AG073105, U01 AG066833, and U01 AG068057.

References

Cascella, M.; Montomoli, J.; Bellini, V.; and Bignami, E. 2023. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *Journal of medical systems*, 47(1): 33.

Chen, Z.; Cano, A. H.; Romanou, A.; Bonnet, A.; Matoba, K.; Salvi, F.; Pagliardini, M.; Fan, S.; Köpf, A.; Mohtashami, A.; et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Chmielewski, N.; and Moretz, J. 2022. ESI triage distribution in US emergency departments. *Advanced emergency nursing journal*, 44(1): 46–53.

Franc, J. M.; Hertelendy, A. J.; Cheng, L.; Hata, R.; and Verde, M. 2024. Accuracy of a Commercial Large Language Model (ChatGPT) to Perform Disaster Triage of Simulated Patients Using the Simple Triage and Rapid Treatment (START) Protocol: Gage Repeatability and Reproducibility Study. *Journal of Medical Internet Research*, 26: e55648.

Friedman, A. B.; Delgado, M. K.; and Weissman, G. E. 2024. Artificial Intelligence for Emergency Care Triage—Much Promise, but Still Much to Learn. *JAMA Network Open*, 7(5): e248857–e248857.

Gilboy, N. 2005. *The emergency severity index. version 4: Implementation handbook*. U.S. Dept. of Health and Human Services, Public Health Service, Agency for Healthcare Research and Quality.

Goh, E.; Bunning, B.; Khoong, E.; Gallo, R.; Milstein, A.; Centola, D.; and Chen, J. H. 2023. ChatGPT influence on medical decision-making, Bias, and equity: a randomized study of clinicians evaluating clinical vignettes. *Medrxiv*.

Goto, T.; Camargo, C. A.; Faridi, M. K.; Freishtat, R. J.; and Hasegawa, K. 2019. Machine learning–based prediction of clinical outcomes for children during emergency department triage. *JAMA network open*, 2(1): e186937–e186937.

Guevara, M.; Chen, S.; Thomas, S.; Chaunzwa, T. L.; Franco, I.; Kann, B. H.; Moningi, S.; Qian, J. M.; Goldstein, M.; Harper, S.; et al. 2024. Large language models to identify social determinants of health in electronic health records. *NPJ digital medicine*, 7(1): 6.

Hinson, J. S.; Martinez, D. A.; Schmitz, P. S.; Toerper, M.; Radu, D.; Scheulen, J.; Stewart de Ramirez, S. A.; and Levin, S. 2018. Accuracy of emergency department triage using the Emergency Severity Index and independent predictors of under-triage and over-triage in Brazil: a retrospective cohort analysis. *International journal of emergency medicine*, 11: 1–10.

Hofmann, V.; Kalluri, P. R.; Jurafsky, D.; and King, S. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028): 147–154.

Hong, W. S.; Haimovich, A. D.; and Taylor, R. A. 2018. Predicting hospital admission at emergency department triage using machine learning. *PloS one*, 13(7): e0201016.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Ito, N.; Kadomatsu, S.; Fujisawa, M.; Fukaguchi, K.; Ishizawa, R.; Kanda, N.; Kasugai, D.; Nakajima, M.; Goto, T.; and Tsugawa, Y. 2023. The accuracy and potential racial and ethnic biases of GPT-4 in the diagnosis and triage of health conditions: evaluation study. *JMIR Medical Education*, 9: e47532.

Jeong, M.; Sohn, J.; Sung, M.; and Kang, J. 2024. Improving Medical Reasoning through Retrieval and Self-Reflection with Retrieval-Augmented Large Language Models. *arXiv preprint arXiv:2401.15269*.

Jiang, L. Y.; Liu, X. C.; Nejatian, N. P.; Nasir-Moin, M.; Wang, D.; Abidin, A.; Eaton, K.; Riina, H. A.; Laufer, I.; Punjabi, P.; et al. 2023. Health system-scale language models are all-purpose prediction engines. *Nature*, 619(7969): 357–362.

Johns Hopkins Medicine. 2022. Insight Nov./Dec. 2022. *Insight*.

Johnson, A.; Bulgarelli, L.; Pollard, T.; Celi, L. A.; Mark, R.; and Horng, S. 2023. MIMIC-IV-ED (version 2.2).

Joseph, J. W.; Kennedy, M.; Landry, A. M.; Marsh, R. H.; Da'Marcus, E. B.; Im, D. E.; Chen, P. C.; Samuels-Kalow, M. E.; Nentwich, L. M.; Elhadad, N.; et al. 2023. Race and ethnicity and primary language in emergency department triage. *JAMA Network Open*, 6(10): e2337557–e2337557.

Labrak, Y.; Bazoge, A.; Morin, E.; Gourraud, P.-A.; Rouvier, M.; and Dufour, R. 2024. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. *arXiv preprint arXiv:2402.10373*.

Lee, J.; Yang, S.; Baik, J. Y.; Liu, X.; Tan, Z.; Li, D.; Wen, Z.; Hou, B.; Duong-Tran, D.; Chen, T.; and Shen, L. 2024. Knowledge-Driven Feature Selection and Engineering for Genotype Data with Large Language Models. arXiv:2410.01795.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Li, D.; Yang, S.; Tan, Z.; Baik, J. Y.; Yun, S.; Lee, J.; Chacko, A.; Hou, B.; Duong-Tran, D.; Ding, Y.; Liu, H.; Shen, L.; and Chen, T. 2024a. DALK: Dynamic Co-Augmentation of LLMs and KG to answer Alzheimer's Disease Questions with Scientific Literature. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2187–2205. Miami, Florida, USA: Association for Computational Linguistics.

Li, L.; Fan, L.; Atreja, S.; and Hemphill, L. 2024b. "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, 18(2): 1–36.

Li, L.; Zhou, J.; Gao, Z.; Hua, W.; Fan, L.; Yu, H.; Hagen, L.; Zhang, Y.; Assimes, T. L.; Hemphill, L.; et al. 2024c. A scoping review of using large language models (llms) to investigate electronic health records (ehrs). *arXiv preprint arXiv:2405.03066*.

Li, S. S.; Balachandran, V.; Feng, S.; Ilgen, J. S.; Pierson, E.; Koh, P. W.; and Tsvetkov, Y. 2024d. MediQ: Question-Asking LLMs and a Benchmark for Reliable Interactive Clinical Reasoning. *arXiv preprint arXiv:2406.00922*.

Liu, J.; Shen, D.; Zhang, Y.; Dolan, B.; Carin, L.; and Chen, W. 2022. What Makes Good In-Context Examples for GPT-3? In Agirre, E.; Apidianaki, M.; and Vulić, I., eds., *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, 100–114. Dublin, Ireland and Online: Association for Computational Linguistics.

Liu, S.; Wright, A. P.; Patterson, B. L.; Wanderer, J. P.; Turer, R. W.; Nelson, S. D.; McCoy, A. B.; Sittig, D. F.; and Wright, A. 2023. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *Journal of the American Medical Informatics Association*, 30(7): 1237–1245.

Lozano, A.; Fleming, S. L.; Chiang, C.-C.; and Shah, N. 2023. Clinfo. ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024*, 8–23. World Scientific.

Lu, M.; Ho, B.; Ren, D.; and Wang, X. 2024. TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 5747–5764.

Masanneck, L.; Schmidt, L.; Seifert, A.; Kölsche, T.; Huntemann, N.; Jansen, R.; Mehsin, M.; Bernhard, M.; Meuth, S. G.; Böhm, L.; et al. 2024. Triage Performance Across Large Language Models, ChatGPT, and Untrained Doctors in Emergency Medicine: Comparative Study. *Journal of Medical Internet Research*, 26: e53297.

Miles, J.; Turner, J.; Jacques, R.; Williams, J.; and Mason, S. 2020. Using machine-learning risk prediction models to triage the acuity of undifferentiated patients entering the emergency care system: a systematic review. *Diagnostic and prognostic research*, 4: 1–12.

Montagna, S.; Ferretti, S.; Klopfenstein, L. C.; Florio, A.; and Pengo, M. F. 2023. Data decentralisation of LLM-based chatbot systems in chronic disease self-management. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*, 205–212.

Moon, S.-H.; Shim, J. L.; Park, K.-S.; and Park, C.-S. 2019. Triage accuracy and causes of mistriage using the Korean Triage and Acuity Scale. *PloS one*, 14(9): e0216972.

Omiye, J. A.; Lester, J. C.; Spichak, S.; Rotemberg, V.; and Daneshjou, R. 2023. Large language models propagate race-based medicine. *NPJ Digital Medicine*, 6(1): 195.

Ouyang, Z.; Qiu, Y.; Wang, L.; De Melo, G.; Zhang, Y.; Wang, Y.; and He, L. 2024. CliMedBench: A Large-Scale Chinese Benchmark for Evaluating Medical Large Language Models in Clinical Scenarios. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 8428–8438. Miami, Florida, USA: Association for Computational Linguistics.

Ovadia, O.; Brief, M.; Mishaeli, M.; and Elisha, O. 2023. Fine-tuning or retrieval? comparing knowledge injection in llms. *arXiv preprint arXiv:2312.05934*.

Park, Y.-J.; Pillai, A.; Deng, J.; Guo, E.; Gupta, M.; Paget, M.; and Naugler, C. 2024. Assessing the research landscape and clinical utility of large language models: A scoping review. *BMC Medical Informatics and Decision Making*, 24(1): 72.

Preiksaitis, C.; Ashenburg, N.; Bunney, G.; Chu, A.; Kabeer, R.; Riley, F.; Ribeira, R.; and Rose, C. 2024. The Role of Large Language Models in Transforming Emergency Medicine: Scoping Review. *JMIR Medical Informatics*, 12: e53787.

Rubin, O.; Herzig, J.; and Berant, J. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.

Shang, T.; Yang, S.; He, W.; Zhai, T.; Li, D.; Hou, B.; Chen, T.; Moore, J. H.; Ritchie, M. D.; and Shen, L. 2024. Leveraging Social Determinants of Health in Alzheimer's Research Using LLM-Augmented Literature Mining and Knowledge Graphs. arXiv:2410.09080.

Shu, C.; Chen, B.; Liu, F.; Fu, Z.; Shareghi, E.; and Collier, N. 2023. Visual med-alpaca: A parameter-efficient biomedical llm with visual capabilities.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Toufiq, M.; Rinchai, D.; Bettacchioli, E.; Kabeer, B. S. A.; Khan, T.; Subba, B.; White, O.; Yurieva, M.; George, J.; Jourde-Chiche, N.; et al. 2023. Harnessing large language models (LLMs) for candidate gene prioritization and selection. *Journal of Translational Medicine*, 21(1): 728.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Williams, C. Y.; Miao, B. Y.; Kornblith, A. E.; and Butte, A. J. 2024a. Evaluating the use of large language models to provide clinical recommendations in the Emergency Department. *Nature Communications*, 15(1): 8236.

Williams, C. Y.; Zack, T.; Miao, B. Y.; Sushil, M.; Wang, M.; Kornblith, A. E.; and Butte, A. J. 2024b. Use of a large language model to assess clinical acuity of adults in the emergency department. *JAMA Network Open*, 7(5): e248895–e248895.

Xu, B.; Wang, Q.; Mao, Z.; Lyu, Y.; She, Q.; and Zhang, Y. 2023. k NN Prompting: Beyond-Context Learning with Calibration-Free Nearest Neighbor Inference. In *The Eleventh International Conference on Learning Representations*.

Yang, Y.; Liu, X.; Jin, Q.; Huang, F.; and Lu, Z. 2024. Unmasking and Quantifying Racial Bias of Large Language Models in Medical Report Generation. *ArXiv*.

Yunxiang, L.; Zihan, L.; Kai, Z.; Ruilong, D.; and You, Z. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.

Zack, T.; Lehman, E.; Suzgun, M.; Rodriguez, J. A.; Celi, L. A.; Gichoya, J.; Jurafsky, D.; Szolovits, P.; Bates, D. W.; Abdulnour, R.-E. E.; et al. 2024. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1): e12–e22.

Zakka, C.; Shad, R.; Chaurasia, A.; Dalal, A. R.; Kim, J. L.; Moor, M.; Fong, R.; Phillips, C.; Alexander, K.; Ashley, E.; et al. 2024. Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2): AIoa2300068.

Zhang, K.; Yu, J.; Yan, Z.; Liu, Y.; Adhikarla, E.; Fu, S.; Chen, X.; Chen, C.; Zhou, Y.; Li, X.; et al. 2023. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*.

Zhang, Z.; Zhang, A.; Li, M.; and Smola, A. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; et al. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36: 55006–55021.