A Counterfactual LLM Framework for Detecting Human Biases: A Case Study of Sex/Gender in Emergency Triage

Ariel Guerra-Adames[1,2,*], Marta Avalos-Fernandez[2], Océane Dorémus[1], Leo Anthony Celi[3,4,5], Cédric Gil-Jardiné[1,6], Emmanuel Lagarde[1]

[1]AHeaD Team, Bordeaux Population Health Research Center, Université de Bordeaux, Bordeaux, F-33000, France.
[2]SISTM Team, Bordeaux Population Health Research Center, Université de Bordeaux, Bordeaux, F-33000, France.
[3]Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, USA.
[4]Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, USA.
[5]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, USA.
[6]Emergency Department, CHU de Bordeaux, Bordeaux, F-33000, France.

Corresponding author(s). E-mail(s): ariel.guerra-adames@u-bordeaux.fr;

Abstract

We present a novel, domain-agnostic counterfactual approach that uses Large Language Models (LLMs) to quantify gender disparities in human clinical decision-making. The method trains an LLM to emulate observed decisions, then evaluates counterfactual pairs in which only gender is flipped, estimating directional disparities while holding all other clinical factors constant. We study emergency triage, validating the approach on more than 150,000 admissions to the Bordeaux University Hospital (France) and replicating results on a subset of MIMIC-IV across a different language, population, and healthcare system. In the Bordeaux cohort, otherwise identical presentations were approximately 2.1% more likely to receive a lower-severity triage score when presented as female rather than male; scaled to national emergency volumes in France, this corresponds to more than 200,000 lower-severity assignments per year. Modality-specific analyses indicate that both explicit tabular gender indicators and implicit textual gender cues contribute to the disparity. Beyond emergency care, the approach supports bias audits in other settings (e.g., hiring, academic, and justice decisions), providing a scalable tool to detect and address inequities in real-world decision-making.

1 Introduction

Emergency triage represents one of the most critical decision points in healthcare delivery, where trained professionals must rapidly assess patient acuity and allocate limited resources under significant time pressure. Modern emergency departments employ standardized triage systems, such as the 5-level Emergency Severity Index (ESI) in the United States or the French Emergency Nurses Classification in Hospitals (FRENCH), to stratify patients based on clinical urgency [1]. These systems aim to ensure that patients with the most critical conditions receive immediate care, with triage

levels ranging from 1 (requiring immediate resuscitation) to 5 (non-urgent). The stakes of these decisions are profound: undertriage can delay critical interventions and increase mortality risk, while overtriage can overwhelm emergency resources and compromise care for other patients [2].

Despite standardization efforts, mounting evidence suggests that emergency triage decisions are susceptible to systematic biases related to patient demographics. Multiple studies have documented disparities in triage assignment based on sex (biological and physiological characteristics of males and females) or gender (socially constructed characteristics) [3–12], race [13–18], age [19], and socioeconomic status [20–22]. Gender bias, in particular, has been observed across diverse clinical presentations. Women presenting with acute coronary syndrome are more likely to receive lower triage acuity scores compared to men [23]. Similar patterns have been documented for stroke and abdominal pain [24], where women experience longer wait times and higher rates of undertriage. These disparities persist even after controlling for clinical factors, suggesting the influence of implicit biases in clinical decision-making[25].

It is important to emphasize that, to date, there is no systematic "gold standard" for evaluating triage decisions that would allow one to determine, for each patient, whether they have been over- or under-triaged. While several studies illustrate the usefulness of audit in improving triage quality [26, 27] and provide examples of how such audits can be conducted [28], they do not establish universal reference standards. Patient outcomes may offer some indication of the appropriateness of a triage decision and the association with patient's sensitive attributes [29], but they cannot account for the contextual information available to the triage nurse at the time of assessment.

The emergence of artificial intelligence, particularly Large Language Models (LLMs), offers unprecedented chances to both understand and address these biases. Crucially, while LLMs have demonstrated remarkable capabilities in clinical tasks [30–33], they also notoriously inherit and reproduce many unintended human behaviors, including the systematic biases present in their training data [34–36]. This tendency, traditionally viewed as a significant limitation, presents a unique opportunity: **LLMs can serve as "bias mirrors,"** providing a scalable tool for auditing human decisions. Rather than deploying LLMs to replace human decision-makers, we propose leveraging their ability to faithfully reproduce human behavior—including its biases—as an analytical tool to detect and quantify these disparities in clinical practice. Their ability to process both structured and unstructured clinical data makes them particularly well-suited for emergency triage, where decisions integrate vital signs, chief complaints, and clinical narratives. By training an LLM to accurately emulate existing human triage decisions, we aim to create a high-fidelity model that captures both the clinical reasoning and the biases embedded in current practice.

In a previous study, we developed a proof-of-concept of a counterfactual framework based on LLMs to detect bias in emergency triage decisions [37]. This approach (illustrated in Figure 1) consists of (1) fine-tuning an LLM to accurately reproduce human triage decisions, (2) generating counterfactual patient pairs where only gender-related information is altered while maintaining clinical factors constant and (3) comparing triage predictions between original and counterfactual presentations using the fine-tuned triage LLM.

This initial work raised several key questions: How can systematic disparities be quantified through appropriate bias metrics? How can we isolate the causal effect of patient gender or sex on triage outcomes while controlling for clinically relevant variables? Once bias is isolated and quantified, what is its potential impact on patient care? Finally, are the observed results reproducible across different datasets, languages, countries and healthcare systems? In this new study, we address these questions by developing appropriate metrics and applying the framework to two distinct datasets: a French dataset protected under European (GDPR) and national (CNIL) regulations, and a semi-publicly accessible U.S. dataset.

2 Conceptual Framework

Our proposed framework consists of three interconnected components designed to detect and quantify biases from sensitive attributes in emergency triage: (1) a fine-tuned LLM-based triage model that replicates human triage decision-making, (2) an LLM-based counterfactual sample generator that systematically modifies sensitive attributes while preserving clinical content, and (3) a set of triage-specific inequality metrics for bias quantification. The framework takes as input a dataset of patient admissions from an emergency department (ED), where each record contains an ordinal triage score assigned by trained clinical personnel, free-text triage notes documenting the clinical assessment and circumstances, and structured variables capturing vital parameters and contextual factors (e.g.,

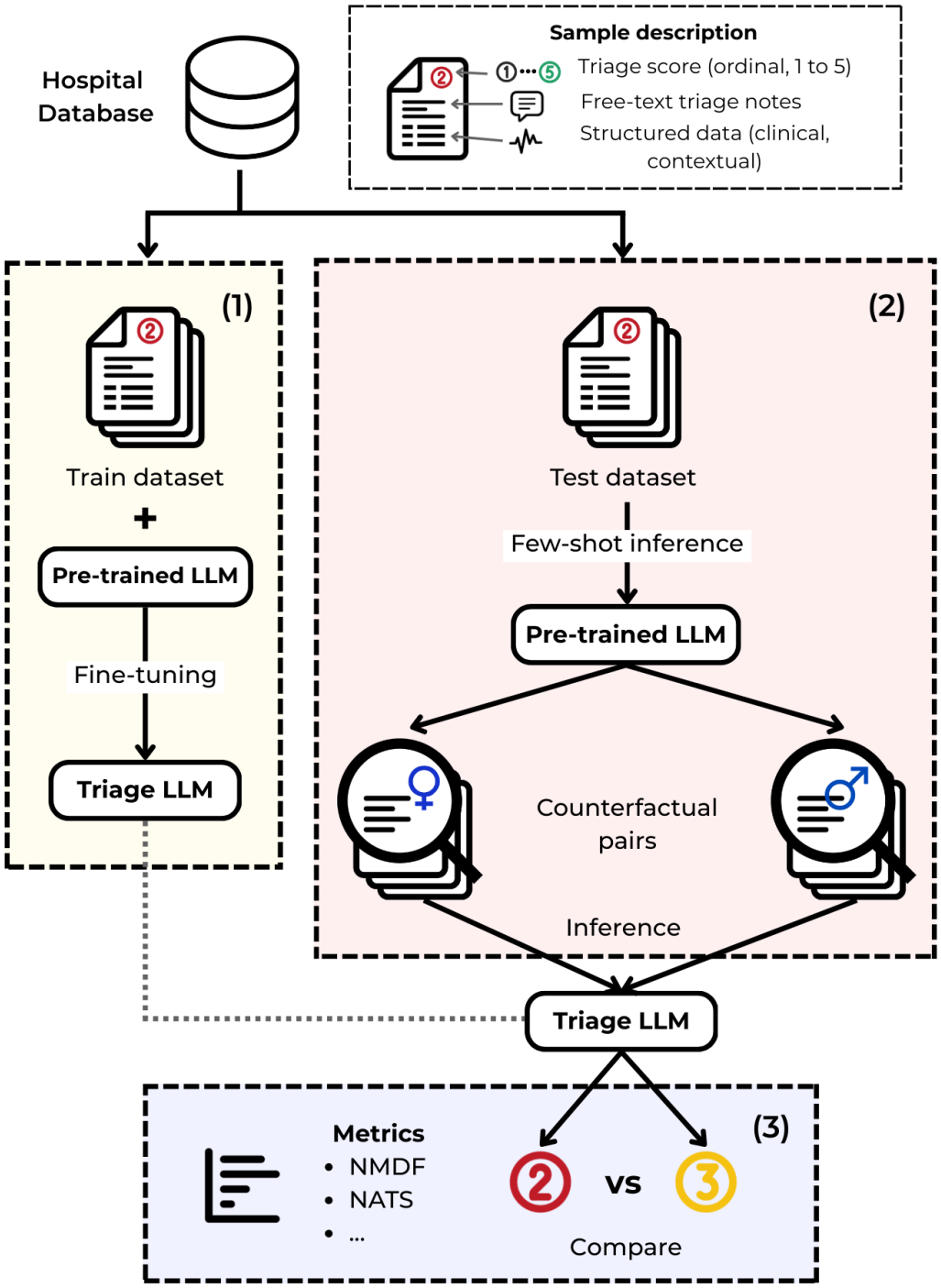

Fig. 1: Overview of the counterfactual framework for gender bias detection in emergency triage: (1) LLM fine-tuning for the task of emergency triage, (2) generation of counterfactual pairs according to a sensitive variable like sex/gender, (3) comparison of triage predictions by the fine-tuned LLM on counterfactual pairs.

patient demographics, nurse identification, temporal factors). The dataset is partitioned into training and test sets of equal size (50%), assuring that the distribution of triage scores is the same on both partitions.

2.1 Fine-tuning LLMs for predicting emergency triage

We frame the task of learning emergency triage from patient data as an ordinal regression problem. The goal is to train a model that predicts the correct triage category.

Input Space

Let $\mathcal{X}$ be the input space of patient records. Each patient record $x \in \mathcal{X}$ is a composite of two modalities: structured tabular data and unstructured textual data: $x = (x_{\text{tab}}, x_{\text{text}})$ where:

- $x_{\text{tab}} \in \mathbb{R}^{d_{\text{tab}}}$ represents d_{tab}-dimensional tabular data, including vital parameters (e.g., heart rate, blood pressure, blood oxygen saturation, body temperature), sociodemographic information (e.g., age, administrative sex of patient), and contextual information (e.g. time of the day, saturation of the ED at the time of patient presentation).
- x_{text} represents textual data such as patient-reported symptoms leading to the ED, medical history, and surrounding context, which are sequences of tokens from a vocabulary $\mathcal{V}$.

Output Space

The output space $\mathcal{Y}$ consists of ordinal triage scores. Let $y \in \mathcal{Y} = \{1, 2, 3, 4, 5\}$ be the true triage score assigned by human emergency triage nurses, where 1 denotes the highest urgency (most severe) and 5 denotes the lowest urgency (least severe).

Model Training

We fine-tune a pretrained LLM $f_\theta : \mathcal{X} \to \mathcal{Y}$ to act as a 5-way classifier over $\mathcal{Y} = \{1, 2, 3, 4, 5\}$. Each patient record $x = (x_{\text{tab}}, x_{\text{text}})$ is rendered as a prompt x_p that linearizes x_{tab}, appends x_{text}, and instructs the model to output a triage score. We cast classification as constrained causal language modeling: we reserve label tokens $T = \{t_1, \ldots, t_5\}$ mapped one-to-one to $\mathcal{Y}$ and train the model to generate the correct label token as the next token. Given the training set $\mathcal{D}_{\text{train}} = \{(x_{p,i}, y_i)\}_{i=1}^{N_{\text{train}}}$, we optimize

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N_{\text{train}}} \log p_\theta(t_{y_i} \mid x_{p,i}) \,.$$

After fine-tuning we obtain parameters $\hat{\theta}$ and the classifier $\hat{f} := f_{\hat{\theta}}$. At inference, for input x we form x_p and predict

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_{\hat{\theta}}(t_y \mid x_p) \,,$$

restricting decoding to the label set T.

2.2 Generating counterfactual pairs

Let each patient record be $x = (x_{\text{tab}}, x_{\text{text}})$, where the tabular features decompose as

$$x_{\text{tab}} = (x_{\text{clin}}, x_{\text{sex}}), \qquad x_{\text{sex}} \in \{0, 1\} \ (0 = \text{male},\ 1 = \text{female}).$$

We construct a counterfactual x' by flipping the recorded sex in the tabular part and rewriting gendered expressions in the text while preserving all other clinical content intact. This design follows causal-reasoning principles for fairness that aim to avoid discrimination by analyzing counterfactual worlds [38].

Tabular counterfactual

We define

$$x'_{\text{sex}} = 1 - x_{\text{sex}}, \qquad x'_{\text{tab}} = (x_{\text{clin}}, x'_{\text{sex}}),$$

i.e., all non-sex tabular variables are held fixed.

Textual counterfactual

Let the note be a token sequence $x_{\text{text}} = (t_1, \ldots, t_n)$. We define a rewriting function

$$g : \mathcal{V} \times \{0, 1\} \to \mathcal{V}$$

that minimally edits explicit and implicit gendered mentions (e.g., pronouns, gendered nouns, kinship terms, honorifics) to be consistent with a target sex, yielding

$$x'_{\text{text}} = g(x_{\text{text}};\, x'_{\text{sex}}).$$

The function g is applied only to the text and is implemented via few-shot prompting of the LLM; the tabular flip $x'_{\text{sex}} = 1 - x_{\text{sex}}$ is deterministic and does not use g.

Counterfactual pair

The full counterfactual record is

$$x' = (x'_{\text{tab}},\, x'_{\text{text}}).$$

Given any instance $x_i = (x_{\text{clin},i}, x_{\text{sex},i}, x_{\text{text},i})$, we form the paired counterfactual

$$x'_i = \big((x_{\text{clin},i},\, 1 - x_{\text{sex},i}),\; g(x_{\text{text},i};\, 1 - x_{\text{sex},i})\big).$$

Implementation with few-shot learning

We implement g using a separate pre-trained LLM in a few-shot learning configuration. The model is prompted with manually curated exemplars that demonstrate how to flip the structured gender field while leaving all other tabular values unchanged, and how to edit the free text to update gendered terms and pronouns consistently with s_{cf} without altering clinical entities, temporal qualifiers, negations, or any non-gender clinical content.

We apply g only to records for which a gender-consistent counterfactual can be produced without changing clinical meaning; when a record contains inherently gender-specific content that cannot be made consistent without altering medical semantics (e.g., pregnancy-specific content), we exclude it from counterfactualization.

Paired test set

Let $\mathcal{D}_{\text{test}}$ be the original test set. Applying g yields a paired test set

$$\mathcal{D}_{\text{pairs}} = \big\{(x_i, x'_i, s_{\text{orig},i}, s_{\text{cf},i})\big\}_{i=1}^{N_{\text{pairs}}}, \qquad N_{\text{pairs}} = \big|\mathcal{D}_{\text{pairs}}\big|.$$

For each pair, we obtain model predictions

$$\hat{y}_i = f_\theta(x_i), \qquad \hat{y}'_i = f_\theta(x'_i),$$

which are used in all downstream bias metrics.

2.3 Quantifying gender biases

We analyze the predictions $\hat{y}_i = f_\theta(x_i)$ and $\hat{y}'_i = f_\theta(x'_i)$ for each counterfactual pair (x_i, x'_i), where x'_i is obtained by applying the gender-flipping mapping g to x_i (see previous subsection). Let $\text{gender}(x) \in \text{male}, \text{female}$ denote the patient gender encoded in x. $f_\theta(x)$ denotes the discrete predicted class in $1, \ldots, 5$, used as an ordered scalar for inequalities.

Pairwise Disagreement Rate (PDR)

The PDR quantifies the proportion of counterfactual pairs for which the model's predicted triage score changes after applying the gender-flip transformation g, i.e., the fraction of (x, x') with $f_\theta(x) \neq f_\theta(x')$, holding all other information fixed:

$$\text{PDR} \;=\; \mathbb{P}(f_\theta(x) \neq f_\theta(x')) \;=\; \mathbb{E}[\mathbb{I}(f_\theta(x) \neq f_\theta(x'))]\,.$$

Given a paired test set $\mathcal{D}_{\text{pairs}} = \{(x_i, x'_i)\}_{i=1}^{N_{\text{pairs}}}$, we estimate

$$\widehat{\text{PDR}} \;=\; \frac{1}{N_{\text{pairs}}} \sum_{i=1}^{N_{\text{pairs}}} \mathbb{I}(f_\theta(x_i) \neq f_\theta(x'_i))\,.$$

Directional change probabilities

We treat the predicted classes $f_\theta(x) \in \{1, \ldots, 5\}$ as ordered severity levels (higher = more severe). We quantify the direction of change in predicted severity when flipping gender.

For $s_1, s_2 \in \{\text{male}, \text{female}\}$ with $s_1 \neq s_2$, define

$$\mathbb{P}^{\uparrow}_{s_1 \to s_2} \;=\; \mathbb{P}(f_\theta(x') > f_\theta(x) \mid \text{gender}(x) = s_1,\; \text{gender}(x') = s_2)\,,$$

$$\mathbb{P}^{\downarrow}_{s_1\to s_2} \;=\; \mathbb{P}(f_\theta(x') < f_\theta(x) \mid \text{gender}(x) = s_1,\ \text{gender}(x') = s_2)\,.$$

Let $\mathcal{D}_{\text{pairs},\, s_1\to s_2} = \{(x_i, x_i') : \text{gender}(x_i) = s_1,\ \text{gender}(x_i') = s_2\}$ with size $|\mathcal{D}_{\text{pairs},\, s_1\to s_2}|$. The empirical estimates are

$$\widehat{\mathbb{P}}^{\uparrow}_{s_1\to s_2} = \frac{1}{|\mathcal{D}_{\text{pairs},\, s_1\to s_2}|} \sum_{(x_i,x_i')\in\mathcal{D}_{\text{pairs},\, s_1\to s_2}} \mathbb{I}(f_\theta(x_i') > f_\theta(x_i))\,,$$

$$\widehat{\mathbb{P}}^{\downarrow}_{s_1\to s_2} = \frac{1}{|\mathcal{D}_{\text{pairs},\, s_1\to s_2}|} \sum_{(x_i,x_i')\in\mathcal{D}_{\text{pairs},\, s_1\to s_2}} \mathbb{I}(f_\theta(x_i') < f_\theta(x_i))\,.$$

Net Mean Disadvantage (NMD)

To compute the NMD, we combine the original and counterfactual pairs and split them by gender. We then calculate the average difference in predicted triage scores between females and males. A positive NMD indicates that, on average, females (originals and counterfactuals combined) receive higher scores (i.e., less severe triage). Negative NMD indicates that, on average, females receive lower scores (i.e., more severe triage) than males.

For each pair i, let p_i^{male} and p_i^{female} denote the two gender presentations (original or counterfactual). Define

$$\text{NMDF} \;=\; \mathbb{E}\big[f_\theta\big(p^{\text{female}}\big) - f_\theta\big(p^{\text{male}}\big)\big]\,, \quad \widehat{\text{NMDF}} \;=\; \frac{1}{N_{\text{pairs}}} \sum_{i=1}^{N_{\text{pairs}}} \Big(f_\theta(p_i^{\text{female}}) - f_\theta(p_i^{\text{male}})\Big).$$

Net Asymmetric Triage Shift (NATS)

Let $\mathbb{P}^{\uparrow}_{s_1\to s_2}$ and $\mathbb{P}^{\downarrow}_{s_1\to s_2}$ be the upward (less severe) and downward (more severe) directional change probabilities defined above. We define

$$\text{NATS}(+) \;=\; \mathbb{P}^{\uparrow}_{\text{female}\to\text{male}} \;-\; \mathbb{P}^{\uparrow}_{\text{male}\to\text{female}}, \qquad \text{NATS}(-) \;=\; \mathbb{P}^{\downarrow}_{\text{female}\to\text{male}} \;-\; \mathbb{P}^{\downarrow}_{\text{male}\to\text{female}}.$$

$\text{NATS}(+) > 0$ indicates a greater propensity for more severe shifts when flipping female→male than male→female; $\text{NATS}(-) > 0$ indicates a greater propensity for left severe shifts when flipping female→male than male→female. Both indices lie in $[-1, 1]$. All NATS values (full, text-isolated, tabular-isolated) are computed on the same index set of counterfactual pairs to ensure comparability.

Directional Triage Skew (DTS)

Within each transformation direction, we summarize the net tendency toward higher vs. lower predicted severity using the directional change probabilities defined above:

$$\text{DTS}_{F|M} \;=\; \mathbb{P}^{\uparrow}_{\text{male}\to\text{female}} \;-\; \mathbb{P}^{\downarrow}_{\text{male}\to\text{female}}, \qquad \text{DTS}_{M|F} \;=\; \mathbb{P}^{\uparrow}_{\text{female}\to\text{male}} \;-\; \mathbb{P}^{\downarrow}_{\text{female}\to\text{male}}.$$

Estimation follows directly:

$$\widehat{\text{DTS}}_{F|M} = \widehat{\mathbb{P}}^{\uparrow}_{\text{male}\to\text{female}} - \widehat{\mathbb{P}}^{\downarrow}_{\text{male}\to\text{female}}, \qquad \widehat{\text{DTS}}_{M|F} = \widehat{\mathbb{P}}^{\uparrow}_{\text{female}\to\text{male}} - \widehat{\mathbb{P}}^{\downarrow}_{\text{female}\to\text{male}}.$$

By construction, $\text{DTS} \in [-1, 1]$. Positive values indicate a net upward (less severe) shift in predicted class within that direction; negative values indicate a net downward (more severe) shift; values near zero indicate balance between upward and downward changes. All DTS estimates are computed on the same index set of counterfactual pairs as other metrics to ensure comparability.

2.4 Decomposing bias: stratified (modality-isolated) analysis

We seek to quantify how much of the asymmetry in predicted severity shifts arises from (i) the free-text notes versus (ii) the tabular sex field. To do so, we recompute the directional change probabilities, NATS(±), and DTS when only one modality is allowed to change, using the same index set of pairs as in the full setting to ensure comparability.

For each record $x_i = (x_{\text{tab},i}, x_{\text{text},i})$ with

$$x_{\text{tab},i} = (x_{\text{clin},i}, x_{\text{sex},i}), \quad x_{\text{sex},i} \in \{0,1\} \ (0 = \text{male}, 1 = \text{female}),$$

let the counterfactual sex bit be $x'_{\text{sex},i} = 1 - x_{\text{sex},i}$. The text-rewriting function g applies only to the note and minimally edits gendered mentions to target a given sex (defined previously). We keep a fixed index set $\mathcal{I} = \{1, \ldots, N_{\text{pairs}}\}$ across all analyses.

Text-isolated analysis (text-iso)

We remove the tabular sex field and flip only textual gender cues:

$$x_i^{\text{text-iso}} = \big(x_{\text{tab},i}^{\text{neut}},\, x_{\text{text},i}\big), \qquad x_i'^{\text{text-iso}} = \big(x_{\text{tab},i}^{\text{neut}},\, g(x_{\text{text},i};\, x'_{\text{sex},i})\big),$$

where $x_{\text{tab},i}^{\text{neut}} = (x_{\text{clin},i})$ drops the sex field while keeping all other tabular variables fixed. Using predictions on the pairs $\{(x_i^{\text{text-iso}}, x_i'^{\text{text-iso}})\}_{i\in\mathcal{I}}$, we compute the directional change probabilities and then

$$\text{NATS}(+)_{\text{text-iso}} = \mathbb{P}^{\uparrow}_{\text{f}\to\text{m, text-iso}} - \mathbb{P}^{\uparrow}_{\text{m}\to\text{f, text-iso}}, \qquad \text{NATS}(-)_{\text{text-iso}} = \mathbb{P}^{\downarrow}_{\text{f}\to\text{m, text-iso}} - \mathbb{P}^{\downarrow}_{\text{m}\to\text{f, text-iso}},$$

$$\text{DTS}_{F|M,\text{text-iso}} = \mathbb{P}^{\uparrow}_{\text{m}\to\text{f, text-iso}} - \mathbb{P}^{\downarrow}_{\text{m}\to\text{f, text-iso}}, \qquad \text{DTS}_{M|F,\text{text-iso}} = \mathbb{P}^{\uparrow}_{\text{f}\to\text{m, text-iso}} - \mathbb{P}^{\downarrow}_{\text{f}\to\text{m, text-iso}}.$$

Tabular-isolated analysis tab-iso

We remove the free-text and flip only the tabular sex bit. Let $\varnothing$ denote the empty-text placeholder:

$$x_i^{\text{tab-iso}} = \big((x_{\text{clin},i},\, x_{\text{sex},i}),\, \varnothing\big), \qquad x_i'^{\text{tab-iso}} = \big((x_{\text{clin},i},\, 1 - x_{\text{sex},i}),\, \varnothing\big).$$

Using predictions on $\{(x_i^{\text{tab-iso}}, x_i'^{\text{tab-iso}})\}_{i\in\mathcal{I}}$, we compute the corresponding directional probabilities and then

$$\text{NATS}(+)_{\text{tab-iso}} = \mathbb{P}^{\uparrow}_{\text{f}\to\text{m, tab-iso}} - \mathbb{P}^{\uparrow}_{\text{m}\to\text{f, tab-iso}}, \qquad \text{NATS}(-)_{\text{tab-iso}} = \mathbb{P}^{\downarrow}_{\text{f}\to\text{m, tab-iso}} - \mathbb{P}^{\downarrow}_{\text{m}\to\text{f, tab-iso}},$$

$$\text{DTS}_{F|M,\text{tab-iso}} = \mathbb{P}^{\uparrow}_{\text{m}\to\text{f, tab-iso}} - \mathbb{P}^{\downarrow}_{\text{m}\to\text{f, tab-iso}}, \qquad \text{DTS}_{M|F,\text{tab-iso}} = \mathbb{P}^{\uparrow}_{\text{f}\to\text{m, tab-iso}} - \mathbb{P}^{\downarrow}_{\text{f}\to\text{m, tab-iso}}.$$

All modality-isolated metrics use the same pair index set $\mathcal{I}$ as the full analysis to preserve one-to-one comparability. Empirical estimates are obtained by replacing probabilities with their sample analogs computed from model predictions in each setting.

2.5 Impact of pre-training on gender bias

When we observe asymmetric changes in predicted severity across gender counterfactuals, these effects may originate from (i) the Bordeaux University Hospital ED dataset (e.g., human or data-collection artifacts) or (ii) biases absorbed during general-domain pre-training of the LLM. Because the pre-training corpora are external to our control, they may inject biases that influence triage predictions independently of our training data. We therefore ask: **does the model already exhibit gender-related asymmetries before being fine-tuned on our data?**

We consider two predictors evaluated under the same framework and on the same paired test set: $f^{(\text{FT})}$, a model pre-trained on general-domain text and then fine-tuned for triage prediction on the ED training split., and $f^{(\text{FS})}$, a model trained from scratch (random initialization) on the same ED training split, without general-domain pre-training.

For any model $m \in \{\text{FT}, \text{FS}\}$, all previously defined metrics are computed by replacing f_θ with $f^{(m)}$. We evaluate in three settings using the same index set $\mathcal{I}$: full (both modalities changed), text-isolated, and tabular-isolated (as defined in the stratified analysis).

If directional change probabilities and derived indices are already non-zero for the FS model, this suggests dataset- or labeling-related asymmetries. Additional asymmetry present in FT beyond FS implicates biases acquired during general-domain pre-training.

3 Results

3.1 Description of Datasets

Our study utilizes two distinct, large-scale ED datasets to train our models and evaluate gender bias: a primary dataset from the Bordeaux University Hospital (*CHU de Bordeaux*), in France, and a subset of the publicly-available MIMIC-IV database in the United States. Both institutions function as tertiary referral centers with full trauma capabilities, ensuring comparable scope of emergency care despite differences in language, patient populations, and documentation standards. However, structural differences in healthcare financing likely influence patient visit patterns, particularly the proportion of non-urgent presentations (ESI/CIMU level 5). In France's public system, a non-admitted ED visit is billed at a fixed national rate (known as the *Forfait Patient Urgences*). This fee is covered by the national health insurance for all individuals affiliated with it, with any remaining balance typically reimbursed by complementary insurance, resulting in minimal or no out-of-pocket costs. In contrast, in the United States, ED visit charges vary by insurance plan, deductible status, and services provided, even in emergencies. These differences may partly explain the lower prevalence of level-5 triages in the MIMIC-IV dataset compared to the CHU de Bordeaux dataset (see Table 1 for a detailed distribution of triage levels). Unlike the Bordeaux University Hospital ED dataset, the MIMIC-IV database includes a substantial proportion of patients assigned a triage score of 1 and, conversely, very few patients assigned a score of 5.

Bordeaux University Hospital ED

The primary dataset consists of retrospective data from 464,989 de-identified adult ED visits at the Bordeaux University Hospital in France, between January 2013 and December 2021. For each visit, we extracted structured data, including patient age and sex and an unstructured free-text triage notes written in French by the triage nurse who assigns the triage score for every patient. The ground-truth triage level was assigned according to the 5-level triage scale used at the hospital.

The dataset underwent systematic pre-filtering to ensure data quality. First, we excluded all records from before January 2016 to ensure all admissions have an associated score which adheres to the the 5-level triage scale introduced in 2015 at the Bordeaux University Hospital. We then removed visits with missing data related to triage level, patient sex, nurse gender, or vital parameters. To enable meaningful counterfactual analysis, we excluded cases with sex-specific medical conditions by filtering out specific ICD-10 chief complaint categories (e.g., recent pelvic or genital pain, urogenital problems, genital bleeding) and performing regular expression searches in clinical notes for sex-specific terms such as pregnancy, prostate, menstruation, and gynecological conditions. Finally, we excluded triage level 1 (resuscitation) cases due to their life-threatening nature and significant class imbalance. This systematic filtering process reduced the dataset from 464,989 to 144,888 visits. The filtered dataset was finally divided into equally-sized train and test partitions of 72,444 samples.

MIMIC-IV

For external validation, we used the MIMIC-IV (v2.2) database [39, 40], which contains de-identified data from ED stays at the Beth Israel Deaconess Medical Center in Boston, MA, between 2008 and 2019. We processed the data by merging the MIMIC-IV-Note and ED modules to obtain both clinical notes and acuity scores for each patient. The text for every patient was truncated to include only information that would be available to a triage nurse at the moment of patient arrival, ensuring temporal consistency with the triage decision-making process. We extracted patient demographics and the English-language free-text chief complaint. To maintain the same four-level triage formulation used in our primary dataset, we excluded patients with a triage score of 5. After applying identical exclusion criteria, the final dataset comprised 76,432 patient visits.

3.2 Model Performance

Triage prediction task

To select the most suitable architecture for our bias analysis, we evaluated a range of models on the emergency triage task using the full test partition of the Bordeaux University Hospital ED dataset ($n = 72{,}444$). Performance was primarily assessed using quadratically weighted Cohen's kappa (κ_w), a standard measure for triage agreement among human experts [41–43], alongside precision, recall, and F1-score.

Characteristic	Bor. Univ. Hospital ED	MIMIC-IV
Location	Bordeaux, France	Boston, MA, USA
Time Period	2016–2021	2008–2019
Language	French	English
Triage Scale	CIMU (5-level)	ESI (5-level)
Total Visits	144,888	76,432
Age (years), mean ± SD	46.2 ± 22.3	60.2 ± 18.2
Male:Female Ratio	1:0.865	1:1.265
Triage Level Distribution: Level 1 (Most severe)	(Filtered)	11.8%
Triage Level Distribution: Level 2	17.5%	48.3%
Triage Level Distribution: Level 3	41.2%	39.5%
Triage Level Distribution: Level 4	35.1%	0.4%
Triage Level Distribution: Level 5 (Least severe)	6.2%	(Filtered)

Table 1: Characteristics of the filtered Bordeaux University Hospital and MIMIC-IV Datasets

Model variants

We compared classical machine learning and deep-learning architectures, from feature-based baselines to medium- and large-scale LLMs:

- RF/TF-IDF baseline: Random Forest model trained on TF-IDF word vectors.
- CHUBert (3.3×10^8 params): a large RoBERTa-style model pre-trained with masked language modeling on the Bordeaux University Hospital train split, then fine-tuned with a classification head for triage prediction.
- MLP w/ CamemBERT embeddings (3.3×10^8 params): a multi-layer perceptron with a single hidden layer of 200 units, taking CamemBERT-derived embeddings as input.
- MistralNeMo 12B: a 12B-parameter LLM fine-tuned for the triage task.
- gpt-oss-20b: a 20B-parameter LLM fine-tuned for the triage task.
- MedGemma 27B: a 27B-parameter medical LLM fine-tuned for the triage task.

All deep-learning models were fine-tuned until training loss stabilized (typically 3–5 epochs). Additional training details are provided in the Methods.

Figure 2 summarizes the relationship between model size and κ_w; Table 2 reports the full set of metrics.

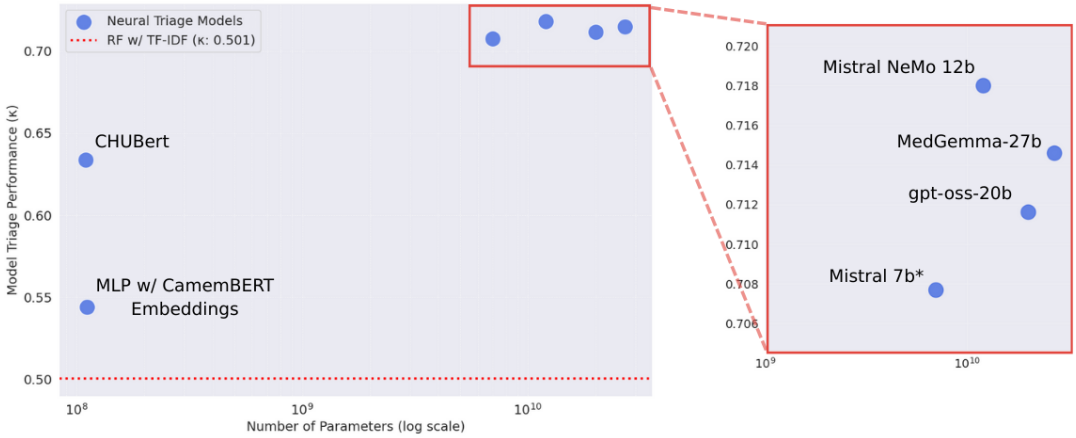

Fig. 2: Performance of triage models on the Bordeaux University Hospital ED test set, evaluated with quadratically weighted Cohen's κ. Models are ordered by parameter size. The red dashed horizontal line indicates the performance of the RF/TF-IDF baseline.

Across the evaluated architectures, we observed a clear performance gradient from classical approaches to larger LLMs. The RF/TF-IDF baseline performed the weakest ($\kappa_w = 0.5006$), while transformer-based models, even at moderate parameter scales such as CHUBert (3.3×10^8), yielded substantial gains ($\kappa_w = 0.6339$). Among medium-to-large LLMs, MistralNeMo 12B achieved the highest κ_w (0.7180) and recall (0.6656), with MedGemma 27B a close second in κ_w (0.7146) despite

being more than twice as large. The best precision (0.6692) and F1-score (0.6633) were obtained by Mistral 7B; these results originate from our prior study [37] and are included here for reference. Interestingly, scaling beyond 12B parameters did not yield substantial improvements in κ_w, suggesting diminishing returns for this task at higher model capacities.

Model	Size	κ_w	Precision	Recall	F1 Score
RF w/ TF-IDF	–	0.5006	0.6103	0.5699	0.5346
MLP w/ BERT Embeddings	3.3×10^8	0.5439	0.5716	0.5733	0.5559
CHUBert	3.3×10^8	0.6339	0.6133	0.6133	0.6035
Mistral NeMo 12B	1.2×10^{10}	**0.7180**	0.6640	0.6656	0.6615
gpt-oss-20b	2.1×10^{10}	0.7116	0.6667	**0.6677**	**0.6648**
MedGemma 27B	2.7×10^{10}	0.7146	0.6585	0.6582	0.6580
Mistral 7B [37]	7×10^9	0.7077	**0.6692**	0.6676	0.6633

Table 2: Performance metrics on the test set. For each metric, the best performance is highlighted in bold. The Mistral 7B row reports results from our previous work using similar data [37].

Counterfactual Pairs Generation Task

We manually assessed the quality of the LLM-generated counterfactual pairs by having two independent annotators review a stratified random sample of 500 transformations from the Bordeaux University Hospital ED dataset and 250 from the MIMIC-IV dataset. Each transformation was classified into one of three categories: (1) correct and complete transformation, where all gender references were accurately modified without changing the rest of the content in the clinical notes, (2) incomplete transformation, where the main gender reference was changed but some secondary gendered language or pronouns were missed, or where medical details were altered and (3) failed transformations, where the transformation changed clinical information other than gender references or added hallucinated information. For the Bordeaux University Hospital ED subset (n=500), annotators agreed on 494 samples and disagreed on 6, which were labeled either as correct or incomplete by annotators. For the MIMIC-IV subset (n=250), annotators agreed on 248 samples and disagreed on 2, which were also labeled either as correct or incomplete. No samples were labeled as failed in either dataset.

3.3 Detection of gender biases

We conducted the bias analysis on the full test partition of the Bordeaux University Hospital ED dataset (40,082 originally male and 32,362 originally female records), under three transformation conditions: full, text-isolated, and tabular-isolated (Table 3). For conceptual reproducibility, we repeat the full-transformation analysis on the MIMIC-IV test partition (21,342 originally male and 16,874 originally female patient records). Metrics that are proportions (e.g., DTS, NATS) are reported as percentages. Throughout, we describe changes in terms of higher vs. lower predicted triage scores, without normative terminology.

We foreground the Directional Triage Skew (DTS), which, within each transformation direction, measures the net balance between increases and decreases in the predicted triage score. In the full setting on the Bordeaux University Hospital ED dataset, we obtained $\text{DTS}_{M|F} = -0.8\%$ [−1.1, −0.5] and $\text{DTS}_{F|M} = 1.3\%$ [1.0, 1.6]. Thus, when transforming samples from female→male, decreases in the predicted score occur more often than increases by 0.8 percentage points; when transforming samples from male→female, increases occur more often than decreases by 1.3 points. Taken together, these skews imply that, for otherwise identical presentations, the female variant is 2.1 percentage points more likely than the male variant to receive a higher predicted triage score (less severe). The Pairwise Disagreement Rate (PDR) was 8.4% [8.2, 8.6], indicating that the transformation changes the assigned class in a non-trivial minority of cases.

Modality-isolated analyses show complementary patterns. Under text-isolated transformations, both skews were substantially negative, $\text{DTS}_{M|F} = -10.3\%$ [−10.8, −9.9] and $\text{DTS}_{F|M} = -12.3\%$ [−12.8, −11.9], meaning that editing only textual gender cues tends to produce more decreases than increases in the predicted score in both directions. In contrast, under tabular-isolated transformations, both skews were positive and similar in magnitude, $\text{DTS}_{M|F} = +3.8\%$ [3.2, 4.3] and $\text{DTS}_{F|F} = +3.8\%$ [3.2, 4.4], meaning that flipping only the tabular sex field tends to produce more increases than decreases in the predicted score in both directions. These modality-specific inversions

suggest that explicit tabular sex markers and implicit textual gender cues influence predictions in opposite directions, with text edits prompting more frequent decreases and tabular flips prompting more frequent increases in the predicted triage score.

The Net Mean Disadvantage for Females (NMDF) metric aligns directionally with these results. In the Bordeaux University Hospital ED dataset, NMDF = 0.011 [0.009, 0.013] and, in the tabular-isolated setting, NMDF = 0.006 [0.001, 0.010]. Because NMDF is a mean difference on the ordinal triage scale (not a proportion), absolute values are small; the positive sign indicates that, on average, the female variant receives a slightly higher (less severe) triage score than the male variant for the same clinical content.

Under text isolation, NATS(−) = −0.8% [−1.0, −0.5] and NATS(+) = 1.2% [0.7, 1.7] meaning that higher-score changes are more frequent for male→female transformations than female→male, while lower-score changes are more frequent for female→male than male→female transformations; under tabular isolation, NATS(+) = −0.8% [−1.4, −0.3] and NATS(−) = −0.8% [−1.4, −0.2], indicating that, conditional on a change, both higher-score and lower-score shifts are more frequent when flipping male→female than female→male, consistent with the positive DTS in both directions when only the sex field is flipped and no textual information is used for triage.

Net asymmetric indices are consistent with the aforementioned skews. In the Bordeaux University Hospital ED full setting, NATS(−) = −1.1% [−1.4, −0.8] and NATS(+) = 1.0% [0.7, 1.3], indicating a greater propensity for upward (less severe) scores when transforming males→females than females→males, and a greater propensity for downward (more severe) scores when transforming females→males than males→females, respectively.

Conceptual reproduction

On the MIMIC-IV dataset, the pattern supports external validity: NMDF = 0.022 [0.016, 0.027], $\mathrm{DTS}_{M|F}$ = −2.46% [−3.32, −1.61] and $\mathrm{DTS}_{F|M}$ = +1.96% [1.19, 2.67], implying that, for otherwise identical presentations, the female variant is 4.4 percentage points more likely than the male variant to receive a higher (less severe) predicted triage score. The PDR was 28.16% [27.7, 28.6], indicating a higher overall rate of score changes when transformations are applied. Net asymmetries were NATS(−) = −1.49% [−2.24, −0.76] and NATS(+) = +2.93% [2.21, 3.66], again indicating more downward (less severe) changes for male→female transformations than female→male and more upward (more severe) changes for female→male than male→female.

The combination $\mathrm{DTS}_{F|M} > 0$ and $\mathrm{DTS}_{M|F} < 0$ in both datasets quantifies a consistent asymmetry whereby female presentations are more likely than male presentations to receive higher predicted triage scores for the same clinical content. Figure 4 visualizes these directional imbalances, and Table 3 reports all estimates with 95% bootstrap confidence intervals.

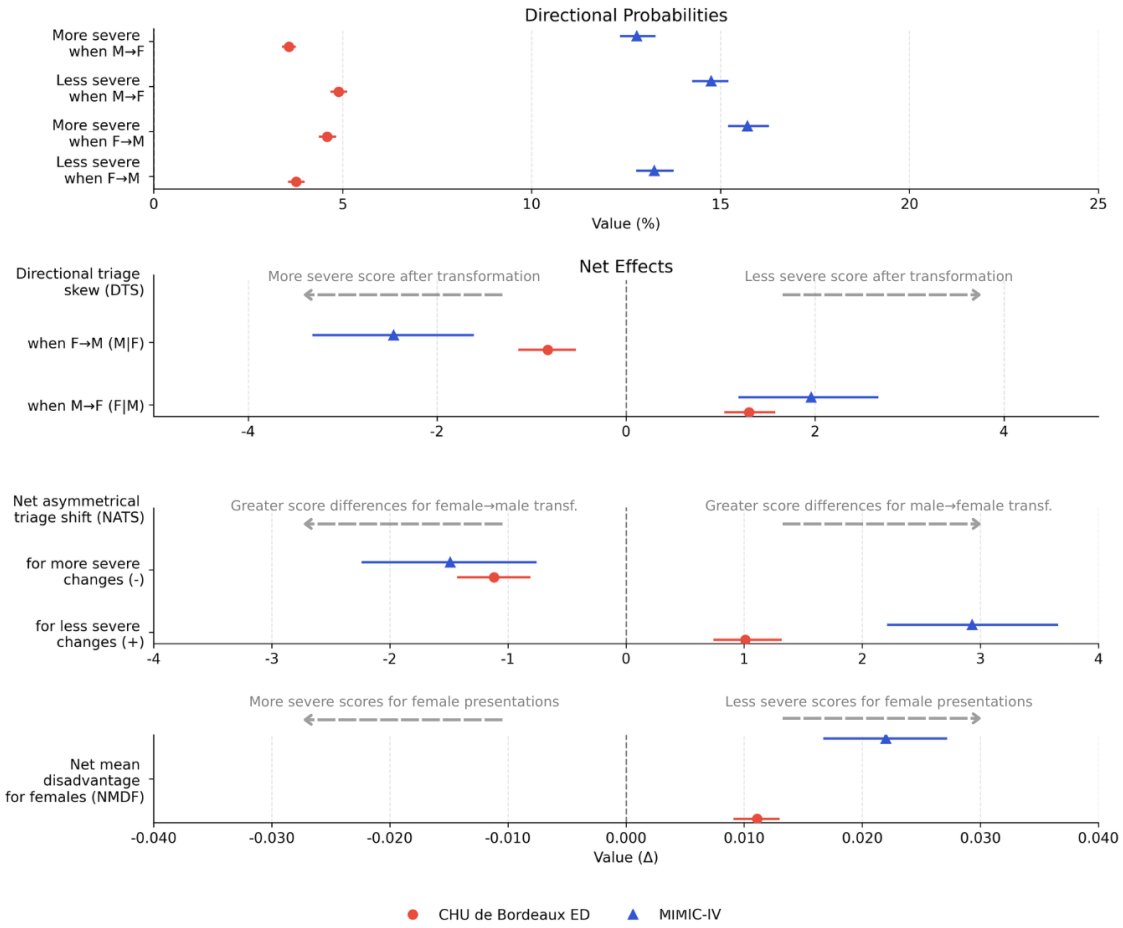

Fig. 3: Bias metrics for the CHU de Bordeaux ED and MIMIC-IV datasets.

Stratifying the counterfactual agreement by triage level reveals a clear acuity dependence in both datasets. Effects are negligible at the most critical level and data-sparse at the trivial end (MIMIC Tri 4). The largest divergence occurs at intermediate acuity (Tri 3), where transforming female→male is more often judged "more severe" than male→female (MIMIC: OR = 0.81, 95% CI [0.75, 0.88]; Bordeaux Univ. Hosp. ED: OR = 0.62, 95% CI [0.55, 0.69]). Full percentage breakdowns and odds ratios by level are reported in Appendix C.

	Bordeaux University Hospital ED			MIMIC-IV
Metric	**Full**	**Text-Iso**	**Tab-Iso**	**Full**
Pairwise Disagreement Rate (PDR)	8.4% [8.2%–8.6%]	18.7% [18.4%–19.0%]	31.9% [31.6%–32.3%]	28.16% [27.70%–28.61%]
Directional Probabilities				
More severe when $M \to F$ ($\mathbb{P}^{\downarrow}_{\mathrm{M}\to\mathrm{F}}$)	3.8% [3.6%–4.0%]	3.3% [3.1%–3.5%]	17.4% [17.0%–17.9%]	12.78% [12.34%–13.28%]
Less severe when $M \to F$ ($\mathbb{P}^{\uparrow}_{\mathrm{M}\to\mathrm{F}}$)	3.6% [3.4%–3.8%]	14.4% [14.1%–14.7%]	14.5% [14.1%–14.8%]	14.75% [14.25%–15.21%]
More severe when $F \to M$ ($\mathbb{P}^{\downarrow}_{\mathrm{F}\to\mathrm{M}}$)	4.9% [4.7%–5.1%]	4.1% [3.9%–4.3%]	18.2% [17.8%–18.6%]	15.71% [15.20%–16.28%]
Less severe when $F \to M$ ($\mathbb{P}^{\uparrow}_{\mathrm{F}\to\mathrm{M}}$)	4.6% [4.4%–4.8%]	15.6% [15.3%–16.0%]	13.6% [13.2%–14.0%]	13.25% [12.77%–13.76%]
Net Effects				
Directional Triage Skew for F→M ($\mathrm{DTS}_{M\mid F}$)	−0.8% [−1.1%–−0.5%]	−10.3% [−10.8%–−9.9%]	3.8% [3.2%–4.3%]	−2.46% [−3.32%–−1.61%]
Directional Triage Skew for M→F ($\mathrm{DTS}_{F\mid M}$)	1.3% [1.0%–1.6%]	−12.3% [−12.8%–−11.9%]	3.8% [3.2%–4.4%]	1.96% [1.19%–2.67%]
Downward Net Asymmetric Triage Shift (NATS(−))	−1.1% [−1.4%–−0.8%]	−0.8% [−1.0%–−0.5%]	−0.8% [−1.4%–−0.2%]	−1.49% [−2.24%–−0.76%]
Upward Net Asymmetric Triage Shift (NATS(+))	1.0% [0.7%–1.3%]	1.2% [0.7%–1.7%]	−0.8% [−1.4%–−0.3%]	2.93% [2.21%–3.66%]
Net Mean Disadvantage for Females (NMDF)	0.011 [0.009–0.013]	−0.0024 [−0.0056–0.0008]	0.0055 [0.0011–0.01]	0.022 [0.0167–0.0272]

Table 3: Side-by-side comparison of gender bias metrics between the Bordeaux University Hospital and the MIMIC-IV datasets. CHU results are reported for three transformation conditions (Full, Text-isolated, Tabular-isolated); MIMIC-IV reports the Full transformation only. Brackets show 95% confidence intervals from 1,000 bootstrap iterations.

Are pre-trained models more biased?

To directly evaluate the impact of pre-training on gender bias in emergency triage, we compared two model variants on the Bordeaux University Hospital ED test set: an LLM that has been pre-trained on a diverse corpus (Mistral NeMo 12B) and subsequently fine-tuned for triage prediction ("fine-tuned from pre-trained"), and a BERT-based architecture which was trained only on data from the train partition of the Bordeaux University Hospital ED dataset ("pre-trained from scratch"). Figure 4 visualizes this comparison across two metrics: NATS(-) and NATS(+).

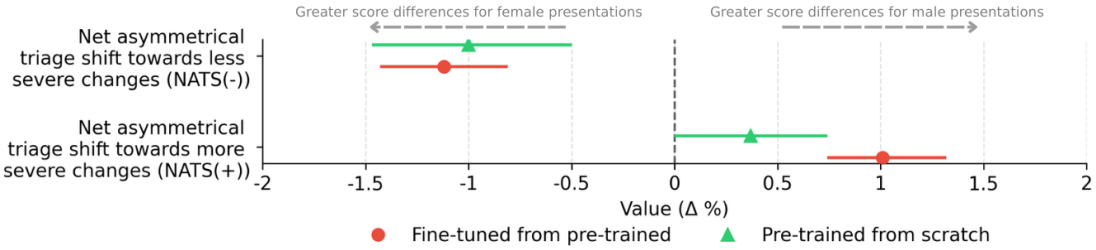

Fig. 4: Bias metrics with and without pre-training.

For differences in less severe changes (NATS(-)), both models are essentially identical: $\text{NATS(-)}_{\text{Fine-tuned}} = -1.12\%$ $[-1.43, -0.81]$ versus $\text{NATS(-)}_{\text{Pre-trained}} = -1.00\%$ $[-1.47, -0.50]$, with overlapping intervals and a negligible difference (≈ 0.12 percentage points). For differences in more severe changes (NATS(+)), the fine-tuned model is somewhat larger: $\text{NATS(+)}_{\text{Fine-tuned}} = +1.01\%$ $[+0.74, +1.32]$ versus $\text{NATS(+)}_{\text{Pre-trained}} = +0.37\%$ $[+0.00, +0.74]$; this suggests a modest increase but is borderline given that the pre-trained model's lower bound touches zero.

This small sensitivity evaluation suggest that while general-domain pre-training may slightly amplify net gender bias in some scenarios, the primary driver of gendered triage disparities is not the pre-training procedure itself, but the underlying structure of gender cues in clinical input data and the language patterns learned during fine-tuning.

4 Discussion

This study presents the first application of LLMs in a counterfactual analysis to detect gender bias in emergency triage decisions. Our findings reveal consistent patterns of gender-based disparities in triage assignment across two large, independent emergency department datasets from different countries and healthcare systems. The detection of these biases through an LLM trained to emulate human decisions provides compelling evidence that such disparities are embedded in current clinical practice.

Key Findings and Clinical Implications

The most striking finding is a consistent directional skew in how identical presentations are triaged after gender flips, succinctly summarized by the Directional Triage Skew (DTS). In the full Bordeaux University Hospital ED setting, $\text{DTS}_{M|F} = +1.3\%$ $[1.0, 1.6]$ and $\text{DTS}_{F|M} = -0.8\%$ $[-1.1, -0.5]$, implying that, for the same clinical conditions, women are about 2.1% more likely to be assigned less severe triage scores than men. This asymmetry suggests that triage decisions incorporate gender-related heuristics that disadvantage female patients.

The clinical implications are profound. In emergency medicine, where minutes can determine outcomes, systematic undertriage of women could contribute to delayed diagnosis and treatment, potentially explaining documented disparities in outcomes for conditions like acute coronary syndrome and stroke [44]. To illustrate the potential scale, if roughly half of emergency visits are by women, a 2.1% differential corresponds to about 220,000 additional undertriage assignments for women per year in France (20.9 million visits $\times$ 50% $\times$ 2.1%), and about 1.63 million per year in the United States (155.4 million $\times$ 50% $\times$ 2.1%). Consistent with this, we also observe that female presentations receive an average triage score 0.011–0.021 points higher (less urgent) than identical

male presentations, which can translate into meaningful differences in wait times and resource allocation at scale. These findings underscore ethical considerations for responsible ML use in clinical pathways and the necessity of governance around fairness auditing [45]

Methodological Contributions

Our counterfactual framework advances the field of bias detection in several ways. First, by training an LLM to accurately replicate human triage decisions, we create a high-fidelity model of current clinical practice. This approach sidesteps the challenge of defining "correct" triage assignments, instead focusing on detecting inconsistencies in how patient gender influences decisions. Second, our modality-specific analysis reveals that bias operates through multiple channels. The high sensitivity when transforming only tabular gender indicators suggests that explicit gender markers strongly influence triage, while the different patterns observed with text-isolated transformations indicate that implicit gender cues in clinical narratives also contribute to bias, albeit through different mechanisms. Third, the use of asymmetric bias metrics (NATS) provides insights beyond simple parity measures. Traditional fairness metrics might miss these directional effects, which reveal not just that outcomes differ by gender, but that the model (and by extension, human decision-makers) responds asymmetrically to gender transformations.

Limitations and Future Directions

Several limitations warrant consideration. First, there is an inherent "multimodality gap" between the information available to our models and the full set of cues used in real triage encounters. Our analysis relies on recorded clinical data, whereas human triage integrates additional modalities such as speech prosody, tone, visible distress, agitation, and overall demeanor. These non-verbal signals can influence perceived urgency but are not captured in our inputs, so our estimates reflect biases in documented information rather than the complete sensory context of triage. Second, while our LLM accurately emulates human triage decisions, it necessarily inherits any biases present in the training data. Our analysis detects relative differences in how gender influences triage but cannot determine absolute "correct" triage levels. Third, our binary gender framework, while enabling clear causal analysis, does not capture the full spectrum of gender identity and its intersection with other demographic factors. Fourth, while our observed effects demonstrate meaningful patterns of bias, there remains the possibility that the LLM itself may *amplify or distort* certain biases present in the training data differently than humans; the complex, non-linear nature of transformer architectures could magnify subtle patterns in ways that do not directly correspond to the original human decision-making processes. We are currently investigating this phenomenon, including through latent space exploration, to better understand *how* the model reproduces these biases and to quantify any potential amplification or distortion effects.

Future work should extend this framework to examine intersectional biases, particularly the interaction of gender with race, age, and socioeconomic status. Additionally, prospective studies could evaluate whether awareness of these biases, facilitated by our detection framework, leads to more equitable triage practices.

Toward Equitable Emergency Care

Our findings should not be interpreted as evidence of conscious bias. Rather, they likely reflect the influence of implicit biases and systemic factors that shape clinical decision-making. The consistency of these patterns across healthcare systems suggests that such biases are deeply embedded in medical training, clinical culture, and decision-making frameworks. Women's symptoms are often described as "atypical", while their vital signs and other clinical measures are frequently assessed against reference standards that are not sufficiently adapted to them. This work demonstrates how AI can serve as a powerful tool for healthcare equity, not by replacing human judgment, but by helping us understand and address its limitations. By providing healthcare systems with scalable methods to audit their practices for bias, we can work toward emergency departments where triage decisions are based solely on clinical need, regardless of patient demographics.

Broader applicability

While this case study focuses on emergency triage, the proposed counterfactual-pairing framework is domain-agnostic and can be applied wherever bias testing and auditing is conducted. Given access to decision inputs (structured and/or textual) and a well-specified transformation g for the protected

attribute(s), the same methodology can audit human resources decisions (e.g., screening, hiring, promotion), academic decisions (e.g., admissions, grading, fellowship selection), and justice-related decisions (e.g., risk assessment, pretrial release, sentencing). The approach does not require defining an absolute ground truth; instead, it detects directional asymmetries in how outcomes change under protected-attribute flips, making it directly compatible with existing discrimination testing practices. Moreover, our cross-country and cross-language replication suggests the framework is portable across institutions, languages, and populations, provided the domain-specific realization of g and neutralization protocols are adapted to the local data schema and documentation norms.

5 Methods

5.1 Data Collection and Preprocessing

5.1.1 *Bordeaux University Hospital ED* Dataset

The raw dataset used for this study comprises 520,621 admissions to the Adult Emergency Department of the Bordeaux University Hospital between January 2013 and December 2021. Each data point contain tabular variables from each admission, including an admission identifier, date and time, estimated saturation of the ED at the time of the admission, patient age and gender, chief complaint, history of present illness, past medical history, vital signs (heart rate, respiratory rate, blood pressure, blood oxygen saturation, systolic and diastolic blood pressure, temperature and Glasgow coma score), and the associated triage score. Additionally, information related to the nurses who performed triage was also included in the analysis, such as the gender of the triage nurse, number of years of experience at the date of triage, and whether they received specialized triage training. Figure 5 shows the distribution of triage scores, patient age and patient gender in the raw dataset.

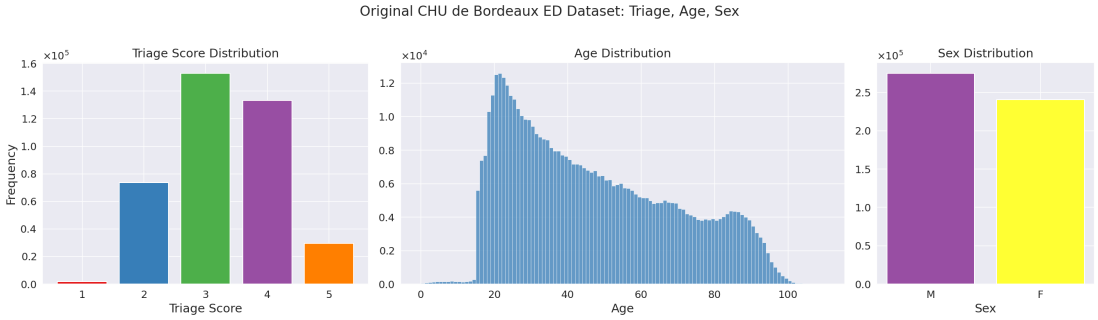

Fig. 5: Distributions of triage score, patient age and patient gender from the raw dataset (before filtering).

Dataset filtering

An initial filtering was performed to exclude all records from before January 2016, ensuring normalization to the 5-level triage scale introduced in 2015. Samples that were missing patient gender, triage nurse gender, age, triage score, chief complaint or triage notes, and samples related to patients under 18 years of age were also excluded.

Gender-specific admission motives or pre-existing conditions were filtered out by examining the past medical history, chief complaint, and history of present illness of each patient. This was done in a sequential manner, first, by excluding admissions with gender-specific chief complaints[1], such as "Other specified conditions associated with female genital organs and menstrual cycle" or "Other specified disorders of male genital organs". The remaining samples were then screened for exclusion using a regular expression search for words associated to gender-specific conditions, such as *uter-* or *prostat-*, in their history of present illness and past medical history. The full list of gender-specific chief complaints and keywords can be found in the Appendix.

Samples with triage score of 1 (resuscitation, immediate care needed) were excluded due to the life-threatening nature of the condition of patients, potentially diminishing bias influence, but also

[1]Chief complaints in the Bordeaux University Hospital Pellegrin ED are categorical and associated to a potential ICD-10-coded diagnosis.

representing a significant class imbalance. The succession of filtering processes resulted in a sample size of 144,888. This filtered version of the dataset was then randomly divided into train and test partitions of equal size (50%), respectively. All filtered tabular and free-text data was then arranged into singular strings containing all available information for each patient and their respective variable names.

5.1.2 MIMIC-IV Dataset

The complete MIMIC-IV database contains data from about 540,000 hospitalizations of patients admitted to the emergency department or intensive care unit of the Beth Israel Deaconess Medical Center in Boston, Massachusetts, between 2008 and 2019. The database is composed of different modules containing different information of different units, such as the emergency department (ED) module, or the clinical notes (Note) module.

Dataset linkage and filtering

In order to assemble a dataset which resembles the Bordeaux University Hospital ED dataset, we selected the ED, Hospitalization (Hosp) and Note modules of the MIMIC-IV database. More specifically, we selected variables from the *triage* and *edstays* tables from the ED module, the *admissions*, *patients* and *providers* tables from the Hosp module and the *discharge* table from the Note module. The first filtering step for this dataset was the linkage between all the aforementioned tables, only keeping samples for patients for which linkage across the three modules was possible.

We excluded samples with a triage score of five ($\sim 0.2\%$ of linked samples) in order to replicate the four-level triage task from the Bordeaux University Hospital ED dataset, and potentially avoiding a significant class-imbalance problem. As with the Bordeaux dataset, samples were screened using regular expressions containing gender-specific words which could not be reasonably transformed into counterfactual pairs. Additionally, since the free-text notes in the MIMIC-IV database contain the complete patient discharge notes, we performed a second regular expression search for any sections in the texts which would not be relevant at triage time.

The remaining sample, composed of about 76,432 admissions, was divided into train and test partitions of equal size.

5.2 Triage Model Training and Inference

Model Architectures

A number of Transformed-based language models were considered for the task of automatic triage, mainly because of their capacity to encode and process large amounts of unstructured texts in multiple languages. These models were selected according on their size and performance balance based on existing benchmarks, ranging from a traditional encoder-only models to state-of-the-art large Mixture of Experts (MoE) models.

In order to establish a baseline, a Random Forest classifier was fitted using TF-IDF vectorization on the clinical notes. The optimal hyperparameter combination for this model was found using 3-fold cross-validation on a sub-partition of the train set. Similarly, a multi-layer perceptron (MLP) with one hidden layer was trained from scratch on word embeddings produced with the CamemBERT large model.

To assess the potential impact of pre-existing biases in pre-trained LLMs, we construct a RoBERTa-style encoder entirely from scratch on the Bordeaux University Hospital ED training partition and then adapt it to triage via supervised fine-tuning. First, a byte-level BPE tokenizer is trained on all ED free text (vocabulary size 52,000, minimum frequency 2) with standard RoBERTa special tokens. The masked-language model is initialized randomly with a fresh `RobertaConfig` aligned to the tokenizer (`vocab_size` = 52,000, `max_position_embeddings` = 1,024, `num_hidden_layers` = 12, `num_attention_heads` = 12, `type_vocab_size` = 1; dropout = 0.1 for both hidden and attention probabilities), yielding a `RobertaForMaskedLM` encoder-decoder whose parameters are learned from scratch. For domain adaptation to triage, we attach a 4-way softmax classification head (`num_labels` = 4) to a RoBERTa encoder initialized from the MLM checkpoint and fine-tune on triage-labeled ED texts from the same train partition.

To explore the scale-performance trade-off of state-of-the-art LLMs, we fine-tuned three LLMs on the train partition of the dataset: Mistral NeMo 12B [46], MedGemma 27B [47], and gpt-oss-20b [48].

Training and Inference Procedures

Fine-tuning of all LLMs was performed using the quantized model weights loaded through the Unsloth framework for optimized training. Each model was configured with 4-bit quantization using bfloat16 precision to reduce memory requirements while maintaining training stability.

To enable efficient fine-tuning of the large language model, we used Low-Rank Adaptation (LoRA) as a parameter-efficient fine-tuning approach. The LoRA configuration utilized a rank of 16 with an alpha value of 32 and a dropout rate of 0.1. The adaptation targeted all attention and feed-forward projection layers, as well as the gate, up, and down projections of the transformer architecture.

The training data was formatted using instruction-following templates following the recommended chat format for each model to enable supervised instruction tuning. Each training sample was structured with a system instruction identifying the model as an emergency triage assistant, followed by the patient's clinical information enclosed within clearly marked boundaries, and concluding with the expected triage score response.

For evaluation, test samples were formatted using the same instruction template structure but truncated before the target answer to enable open-ended generation. Inference was performed using a text generation pipeline configured to generate a maximum of 2 new tokens, sufficient to produce only the numerical triage score. The pipeline utilized multi-GPU distribution with load balancing across available GPUs to optimize inference speed. The padding token was set to the end-of-sequence token to ensure consistent tokenization across all samples. Model predictions were extracted from the generated text and parsed as integers corresponding to triage acuity levels ranging from 1 to 4 or 2 to 5, depending on the dataset used.

5.3 Counterfactual Pair Generation

Three variables from the test portion of the dataset (gender of patient, history of present illness, and past medical history) were selected and assembled into free-text strings. The Mistral Small 24B v3.2 model was used in a 7-shot learning configuration for automatic counterfactual pair generation. Seven manually-annotated examples of counterfactual pairs were included in the model instructions, after the system prompt. The model was instructed to change all references to patient gender/gender to the opposite gender while maintaining all other clinical information unchanged. The model was instructed to respond in comma-separated format (gender of patient, history of present illness, past medical history), allowing storage of transformed data following the same structure as original data. The model, along with its prompt, was served locally using vLLM [49].

While the Mistral small model was selected for its balance between size and performance at the time of the execution of experiments, we believe any state-of-the-art reasoning model of similar or greater size could be used for this task.

5.4 Hardware Specifications

All fine-tuning and counterfactual pair generation tasks were executed on a server running Ubuntu 22.04 LTS with an AMD EPYC 7713 Processor and four NVIDIA A100 80GB GPUs.

5.5 Regulatory Approval

This study adhered to the MR-004 reference methodology (research not involving the human person, health studies and assessments for reused data) as outlined by the French National Commission on Informatics and Liberty (CNIL), within the Bordeaux University Hospital.

6 Conclusion

This study establishes a novel paradigm for detecting and quantifying bias in clinical decision-making through counterfactual analysis with LLMs. By training an LLM to accurately emulate human triage decisions and then systematically evaluating its behavior on gender-swapped patient presentations, we have uncovered consistent patterns of gender bias that persist across different healthcare systems, languages, and patient populations. The consistent asymmetries observed in both French and American emergency departments provides compelling evidence that current triage practices systematically disadvantage female patients in subtle but measurable ways.

Our findings carry immediate practical implications. Healthcare systems can now deploy this framework to audit their own triage practices, identifying specific clinical presentations or contexts where gender bias is most pronounced. The modality-specific analysis further enables targeted interventions, for instance, modifying electronic health record interfaces to minimize the salience of demographic information during triage assessment, or developing training programs that address implicit biases in interpreting clinical narratives. Importantly, our approach scales efficiently and can be extended to detect biases related to other demographic factors and their intersections.

Crucially, a framework such as this one serves as a diagnostic tool that can point towards concrete recommendations for debiasing, which must ultimately be developed through expert clinical consensus. To this end, the results of this study are currently being shared and discussed with clinicians at the Bordeaux University Hospital to help refine training protocols for triage nurses. The broader ambition is to leverage these findings to collaborate with national bodies, such as the French National College of emergency physicians (SFMU), to inform and produce national recommendations for more equitable emergency care.

Looking forward, this work demonstrates the transformative potential of AI not as a replacement for clinical judgment, but as a mirror that reflects and quantifies the biases embedded in current practice. As healthcare systems increasingly commit to equity as a core value, tools that can objectively measure progress toward that goal become essential. By making visible the invisible influences of gender on clinical decisions, we take a crucial step toward emergency departments where a patient's urgency for care is determined solely by their clinical need. The question now is not whether these biases exist, but how quickly and effectively we can act to eliminate them.

Data Availability

Version 2.2 of the MIMIC-IV dataset is publicly available at this address upon approval of a user's profile. The Bordeaux University Hospital ED dataset cannot be made publicly available due to current patient privacy regulations in France but may be accessible through formal collaboration agreements.

Code Availability

The code for model training, counterfactual generation, and bias analysis will be made publicly available upon publication of the article.

Funding

The research efforts of A.G.A. are supported by a doctoral grant from the Digital Public Health Graduate Program of the University of Bordeaux (CD EUR DPH). L.A.C. is funded by the National Institute of Health through DS-I Africa U54 TW012043-01 and Bridge2AI OT2OD032701, the National Science Foundation through ITEST #2148451, and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: RS-2024-00403047). Data collection and processing is made possible with funding from the TARPON project of the Health Data Hub, France.

Acknowledgments

We sincerely thank the triage nurses who participated in this study. While their daily dedication to patient care is a given, it is their contribution to this research that made it possible, and we hope its outcomes will in turn help improve their work.

Author Contributions. A.G.A., M.A.F., C.G.J., and E.L. conceptualized the study. A.G.A. performed all experiments, data manipulation, and data analysis, and drafted the original manuscript. O.D. and L.A.C. contributed to methodology development. E.L. and M.A.F. handled project direction, and supervised the research. C.G.J. and L.A.C. provided expertise in emergency medicine. All authors (A.G.A., M.A.F., O.D., L.A.C., C.G.J. and E.L.) reviewed and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

Competing Interests. None

References

[1] Zachariasse, J. M., van der Hagen, V., Seiger, N., Mackway-Jones, K., van Veen, M. & Moll, H. A. Performance of triage systems in emergency care: a systematic review and meta-analysis. *BMJ Open* **9**, e026471 (2019).

[2] Hinson, J. S. et al. Triage performance in emergency medicine: a systematic review. *Ann. Emerg. Med.* **74**, 140–152 (2019).

[3] Coisy, F. et al. Do emergency medicine health care workers rate triage level of chest pain differently based upon appearance in simulated patients? *Eur. J. Emerg. Med.* **31**, 188–194 (2024).

[4] Han, N., Jeong, S., Lee, S.-Y. & Kong, S. Y. Gender and age bias in the evaluation of suicide attempt behavior in an emergency department. *Community Ment. Health J.* **59**, 1521–1531 (2023).

[5] Banco, D. et al. Sex and race differences in the evaluation and treatment of young adults presenting to the emergency department with chest pain. *J. Am. Heart Assoc.* **11**, e024199 (2022).

[6] Onal, E. G. et al. Comparison of emergency department throughput and process times between male and female patients: A retrospective cohort investigation by the reducing disparities increasing equity in emergency medicine study group. *J. Am. Coll. Emerg. Physicians Open* **3**, e12792 (2022).

[7] Lopez, R. et al. Sex-based differences in timely emergency department evaluations for patients with drug poisoning. *Public Health* **199**, 57–64 (2021).

[8] Preciado, S. M. et al. Evaluating sex disparities in the emergency department management of patients with suspected acute coronary syndrome. *Ann. Emerg. Med.* **77**, 416–424 (2021).

[9] Vigil, J. M. et al. Patient ethnicity affects triage assessments and patient prioritization in US Department of Veterans Affairs emergency departments. *Medicine* **95**, e3191 (2016).

[10] Chen, E. H. et al. Gender disparity in analgesic treatment of emergency department patients with acute abdominal pain. *Acad. Emerg. Med.* **15**, 414–418 (2008).

[11] Croskerry, P. & Sinclair, D. Emergency medicine: a practice prone to error? *Can. J. Emerg. Med.* **3**, 271–276 (2001).

[12] Arslanian-Engoren, C. Gender and age bias in triage decisions. *J. Emerg. Nurs.* **26**, 117–124 (2000).

[13] Schrader, C. D. & Lewis, L. M. Racial disparity in emergency department triage. *J. Emerg. Med.* **44**, 511–518 (2013).

[14] López, L., Wilper, A. P., Cervantes, M. C., Betancourt, J. R. & Green, A. R. Racial and sex differences in emergency department triage assessment and test ordering for chest pain, 1997–2006. *Acad. Emerg. Med.* **17**, 801–808 (2010).

[15] Lin, P. et al. Disparities in emergency department prioritization and rooming of patients with similar triage acuity score. *Acad. Emerg. Med.* **29**, 1320–1328 (2022).

[16] Peitzman, C., Carreras Tartak, J., Samuels-Kalow, M., Raja, A. & Macias-Konstantopoulos, W. Racial differences in triage for emergency department patients with subjective chief complaints. *West. J. Emerg. Med.* **24**, 888–893 (2023).

[17] Morel, S. Inequality and discrimination in access to urgent care in France: ethnographies of three healthcare structures and their audiences. *Soc. Sci. Med.* **232**, 25–32 (2019).

[18] Wu, Y., Xirasagar, S., Nan, Z., Heidari, K. & Sen, S. Racial disparities in utilization of emergency medical services and related impact on poststroke disability. *Med. Care* **61**, 796–804 (2023).

[19] Platts-Mills, T. F. et al. Accuracy of the Emergency Severity Index triage instrument for identifying elder emergency department patients receiving an immediate life-saving intervention. *Acad. Emerg. Med.* **17**, 238–243 (2010).

[20] Kangovi, S., Barg, F. K., Carter, T., Long, J. A., Shannon, R. & Grande, D. Understanding why patients of low socioeconomic status prefer hospitals over ambulatory care. *Health Aff.* **32**, 1196–1203 (2013).

[21] Turner, A. J., Francetic, I., Watkinson, R., Gillibrand, S. & Sutton, M. Socioeconomic inequality in access to timely and appropriate care in emergency departments. *J. Health Econ.* **85**, 102668 (2022).

[22] Verma, S., Wilson, F., Wang, H., Smith, L. & Tak, H. J. Impact of community socioeconomic characteristics on emergency medical service delays in responding to fatal vehicle crashes. *AJPM Focus* **2**, 100129 (2023).

[23] Mnatzaganian, G., Braitberg, G., Hiller, J. E., Kuhn, L. & Chapman, R. Sex differences in in-hospital mortality following a first acute myocardial infarction: symptomatology, delayed presentation, and hospital setting. *BMC Cardiovasc. Disord.* **16**, 109 (2016).

[24] Amy, Y. X. et al. Sex differences in presentation and outcome after an acute transient or minor neurologic event. *JAMA Neurol.* **76**, 962–968 (2019).

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

[26] Ouellet, S. et al. Strategies to improve the quality of nurse triage in emergency departments: A systematic review. *Int. Emerg. Nurs.* **81**, 101639 (2025).

[27] Zaboli, A. et al. Daily triage audit can improve nurses' triage stratification: A pre-post study. *J. Adv. Nurs.* **79**, 605–615 (2023).

[28] Pilleron, B. et al. Nurses' moral judgements during emergency department triage - A prospective mixed multicenter study. *Int. Emerg. Nurs.* **75**, 101479 (2024).

[29] Liu, X. et al. Evaluating prognostic bias of critical illness severity scores based on age, sex, and primary language in the United States: A retrospective multicenter study. *Crit. Care Explor.* **6**, e1033 (2024).

[30] Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).

[31] Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).

[32] Patel, S. B. & Lam, K. ChatGPT: the future of discharge summaries? *Lancet Digit. Health* **5**, e107–e108 (2023).

[33] Shen, D., Wu, G. & Suk, H. I. Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* **23**, 1–27 (2021).

[34] Mahajan, A., Obermeyer, Z., Daneshjou, R., Lester, J. & Powell, D. Cognitive bias in clinical large language models. *npj Digit. Med.* **8**, 428 (2025).

[35] Yang, Y., Liu, X., Jin, Q., Huang, F. & Lu, Z. Unmasking and quantifying racial bias of large language models in medical report generation. *Commun. Med.* **4**, 176 (2024).

[36] Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V. & Daneshjou, R. Large language models propagate race-based medicine. *npj Digit. Med.* **6**, 195 (2023).

[37] Guerra-Adames, A., Avalos-Fernandez, M., Doremus, O., Gil-Jardiné, C. & Lagarde, E. Uncovering judgment biases in emergency triage: a public health approach based on large language models. In *Proceedings of Machine Learning for Health* 420–439 (PMLR, 2025).

[38] Kusner, M. J., Loftus, J., Russell, C. & Silva, R. Counterfactual fairness. *Adv. Neural Inf. Process. Syst.* **30** (2017).

[39] Johnson, A. E. W. et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **10**, 1 (2023).

[40] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R. & Stanley, H. E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. **101** (23), pp.**e215–e220** (2000).

[41] Suamchaiyaphum, K., Jones, A. R. & Polancich, S. The accuracy of triage classification using Emergency Severity Index. *Int. Emerg. Nurs.* **77**, 101537 (2024).

[42] Pourasghar, F., Tabrizi, J. S., Sarbakhsh, P. & Daemi, A. Kappa agreement of emergency department triage scales; a systematic review and meta-analysis. *J. Clin. Res. Gov.* **3**, 124–133 (2014).

[43] Lalla, R., Sammy, I., Paul, J., Nunes, P., Ramcharitar Maharaj, V. & Robertson, P. Assessing the validity of the Triage Risk Screening Tool in a third world setting. *J. Int. Med. Res.* **46**, 557–563 (2018).

[44] Mehilli, J. & Presbitero, P. Coronary artery disease and acute coronary syndrome in women. *Heart* **106**, 487–492 (2020).

[45] Char, D. S., Shah, N. H. & Magnus, D. Implementing machine learning in health care—addressing ethical challenges. *N. Engl. J. Med.* **378**, 981–983 (2018).

[46] Mistral AI. Mistral NeMo: A State-of-the-Art 12B Language Model with 128k Context Length. https://mistral.ai/news/mistral-nemo (2024).

[47] Sellergren, A. et al. MedGemma Technical Report. Preprint at https://arxiv.org/abs/2507.05201 (2025).

[48] OpenAI. gpt-oss-20b and gpt-oss-120b Model Card. https://cdn.openai.com/pdf/419b6906-9da6-406c-a19d-1bb078ac7637/oai_gpt-oss_model_card.pdf (2025).

[49] Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C.H., Gonzalez, J., Zhang, H. & Stoica, I. Efficient memory management for large language model serving with paged attention. *Proceedings of the 29th symposium on operating systems principles*. **611–626**, (2023)

Appendix A Exclusion criteria for strongly gendered admission motives and past medical history

All patients admitted to the ED at Bordeaux University Hospital are assigned a chief complaint class from a list of 189 possible motives, roughly corresponding to ICD-10 diagnoses. Some classes are clearly indicative of patient sex, involving sex-specific pathologies or organs. As a first filtering step, we removed samples classified under any of the following classes:

- Recent pelvic or genital pain
- Female/male urogenital problem without pain
- Female/male genital bleeding
- Foreign body in the genitourinary tract

This pre-filtering was insufficient, as past medical history and history of present illness fields sometimes contain additional sex-specific references. We performed regular expression searches in both free-text columns to identify records containing any of 33 French-language sex-specific medical terms or term stems. These included references to reproductive anatomy (e.g., *uter, ovair, prostat, testicul*), sex-specific conditions (e.g., *grossesse, endométriose, andropause*), procedures (e.g., *hystérectomie, ivg, cesarienne*), medical specialties (e.g., *Gyné, Obsté*), and sex-specific neoplasms (e.g., *cancer du sein, K sein*). Records matching any of these terms were excluded from counterfactual analysis. The complete list of search terms with English translations is provided in Supplementary Table A1.

Category	**French terms (regex stems)**	**English translation/description**
Female reproductive anatomy	uter, utér, ovair, ovarien, vagin, vulv, mamma, fallope	uterus, uterine, ovary, ovarian, vaginal, vulva, mammary, fallopian
Male reproductive anatomy	prostat, testicul, peni, pénien, scrot, séminale	prostate, testicular, penile, scrotal, seminal
Pregnancy and reproductive conditions	grossesse, enceinte, fausse couche, ivg, curetage	pregnancy, pregnant, miscarriage, voluntary interruption of pregnancy, curettage
Female-specific conditions	menopause, ménorrhée, règles, endométriose, endometriose	menopause, menstruation, menstrual periods, endometriosis
Male-specific conditions	andropause	andropause
Gynecological procedures	hystérectomie, hysterectomie, ligature trompes, cesarienne, cezarienne	hysterectomy, tubal ligation, cesarean section
Medical specialties	Gyné, Obsté, gyneco, obste	gynecology, obstetrics
Sex-specific neoplasms	cancer du sein, K sein	breast cancer (K is French medical shorthand for cancer)

Table A1: **French-language regular expression terms used for sex-specific content exclusion.** All terms were searched as case-insensitive substrings in past medical history and history of present illness fields.

A.1 MIMIC-IV

We applied an analogous filtering procedure to the MIMIC-IV dataset to exclude records with sex-specific content that could not be meaningfully transformed in counterfactual analysis. Regular expression searches were performed on both the chief complaint and free-text clinical note fields using English-language equivalents of the French terms employed in the Bordeaux dataset. These English search terms encompassed the same conceptual categories: reproductive anatomy (e.g., *uter, ovar, prostat, testicul*), sex-specific conditions (e.g., *pregnan, menopaus, endometrios*), procedures (e.g., *hysterectom, cesare, tubal ligation*), and sex-specific neoplasms (e.g., *breast cancer*). Records containing any of these terms were excluded from the MIMIC-IV counterfactual analysis cohort. The complete list of English search terms is provided in Supplementary Table A2.

Category	English terms (regex stems)	Description
Female reproductive anatomy	uter, uterin, ovar, ovarian, vagin, vulv, mammar, fallop	uterus, uterine, ovary, ovarian, vaginal, vulva, mammary, fallopian
Male reproductive anatomy	prostat, testic, penil, scrot, seminal	prostate, testicular, penile, scrotal, seminal
Pregnancy and reproductive conditions	pregnan, miscarr, abortion, curettage	pregnancy, pregnant, miscarriage, abortion, curettage
Female-specific conditions	menopaus, menstruat, menses, endometrios	menopause, menstruation, menstrual periods, endometriosis
Male-specific conditions	andropaus	andropause
Gynecological procedures	hysterectom, tubal ligation, cesare, c-section	hysterectomy, tubal ligation, cesarean section
Medical specialties	gyneco, obstet, ob/gyn	gynecology, obstetrics
Sex-specific neoplasms	ovarian cancer, prostate cancer	sex-specific malignancies

Table A2: **English-language regular expression terms used for sex-specific content exclusion in MIMIC-IV.** All terms were searched as case-insensitive substrings in chief complaint and clinical note fields.

Appendix B Large Language Model Prompt Templates

B.1 Emergency Triage Prediction Task

B.1.1 Bordeaux University Hospital ED (French)

The following prompt template was used to fine-tune and evaluate Mistral NeMo 12B for triage prediction on French clinical notes. The model was instructed to classify patients into triage levels 2–5 based on the clinical history provided:

French Triage Prompt Template

```
<s>[INST] Vous êtes un système qui aide à effectuer le triage des patients dans le
service des urgences d'un hôpital. Compte tenu de l'anamnèse (qui contient diverses
informations relatives au patient(e) et son contexte), affectez un score de triage
de 2 à 5 (soit 2, soit 3, soit 4, soit 5) pour le patient, ou 2 est plus grave et 5
moins grave. Juste le numéro, pas d'explications avec.
#### Début de l'anamnèse ####
{patient_clinical_history}
#### Fin de l'anamnèse ####
Score de triage: [/INST]
```

English translation: "You are a system that helps perform triage of patients in a hospital emergency department. Given the anamnesis (which contains various information related to the patient and their context), assign a triage score from 2 to 5 (either 2, 3, 4, or 5) for the patient, where 2 is more severe and 5 is less severe. Just the number, no explanations with it."

B.1.2 MIMIC-IV (English)

For the MIMIC-IV dataset, we used an analogous English-language prompt template adapted to the 4-level triage scale (ESI levels 1–4) used in the dataset:

English Triage Prompt Template

```
<s>[INST] You are an assistant that helps perform emergency triage at an emergency
department of a hospital. Taking into account each patient's clinical history (and
other information related to the patient's condition and their context), assign a
triage score from 1 to 4 (either 1, 2, 3 or 4) where 1 is the most critical and 4 is
the least critical. Only the number, no explanations with
#### Start of the patient's file ####
{patient_clinical_history}
#### End of the patient's file ####
Triage score: [/INST]
```

B.2 Counterfactual Pair Generation Task

B.2.1 Bordeaux University Hospital ED (French)

For generating counterfactual pairs from French clinical notes, we employed a few-shot learning approach. The generation prompt template structure was:

French Counterfactual Generation Prompt Structure

System prompt:

```
Tu es un assistant qui va transformer des anamnèses médicales en remplaçant les
références au sexe dans le texte par le sexe opposé.
```

User message:

```
Tu vas recevoir l'anamnèse d'un(e) patient(e). Ta tâche est de remplacer les
références au sexe par le sexe opposé. Par exemple, "patient" devient "patiente",
"homme" devient "femme", etc. N'effectue aucune traduction et ne corrige aucune faute
d'orthographe.

[10 manually curated exemplars demonstrating gender transformations]

#### Anamnèse à changer ####
Anamnèse originale:
Sexe patient : {sex}, Anamnèse : {clinical_note}, Antécédents : {medical_history}
```

English translation:

System prompt: "You are an assistant who will transform medical anamneses by replacing sex references in the text with the opposite sex."

User message: "You will receive the anamnesis of a patient. Your task is to replace sex references with the opposite sex. For example, 'patient' (masculine) becomes 'patiente' (feminine), 'man' becomes 'woman', etc. Do not perform any translation and do not correct any spelling errors."

The prompt included 10 manually curated exemplars demonstrating various gender transformation scenarios, including gendered nouns, pronouns and grammatical agreement markers.

B.2.2 MIMIC-IV (English)

For generating counterfactual pairs from English clinical notes in MIMIC-IV, we employed an analogous few-shot learning approach. The generation prompt template structure was:

English Counterfactual Generation Prompt Structure

System prompt:

```
You are an assistant who will transform medical notes by replacing sex references in
the text with the opposite sex.
```

User message:

```
You will receive a patient's clinical note. Your task is to replace sex references
with the opposite sex. For example, "he" becomes "she", "man" becomes "woman", etc.
Do not perform any other modifications to the clinical content and do not correct any
spelling errors.

[10 manually curated exemplars demonstrating gender transformations]

#### Note to transform ####
Original note:
Patient sex: {sex}, Chief complaint: {chief_complaint}, Clinical note: {clinical_note}
```

The prompt included 10 manually curated exemplars demonstrating various gender transformation scenarios in English clinical documentation, including pronoun changes (he/she, his/her, him/her) and gendered nouns (man/woman, gentleman/lady, boy/girl).

Appendix C Gender bias differences across triage levels

This sub-analysis aims to evaluate whether the results observed in our counterfactual study are influenced by the original severity level (i.e., the original triage level). By stratifying our analysis by triage score, we can determine if gender bias varies across different urgency levels.

C.1 Bordeaux University Hospital ED

We analyzed gender bias by comparing counterfactual predictions (M→F and F→M) across triage levels. Table C3 shows the proportion of cases where the counterfactual was judged less severe ($cf > orig$) or more severe ($cf < orig$) than the original, stratified by sex and triage level.

Triage Score	$cf > orig$ (M)	$cf < orig$ (M)	$cf > orig$ (F)	$cf < orig$ (F)	Δ M-F $cf >$	Δ M-F $cf <$
2	9.9%	0.0%	5.4%	0.0%	+4.5	0.0
3	6.1%	3.4%	4.9%	5.4%	+1.2	-2.0
4	1.8%	4.6%	1.7%	5.7%	+0.1	-1.1
5	0.0%	12.7%	0.0%	9.0%	0.0	+3.7

Table C3: Percentage of discordant predictions by triage level and sex (Bordeaux)

To quantify the strength of gender effects, we calculated odds ratios (OR) for the probability that the counterfactual was judged more severe than the original ($cf < orig$). Table C4 presents these results.

Triage Sc.	$cf < orig$ (M)	$cf \geq orig$ (M)	$cf < orig$ (F)	$cf \geq orig$ (F)	OR (M/F)	IC95%	p-val
2	0	7,438	0	5,858	—	—	—
3	530	15,115	769	13,578	0.62	[0.55–0.69]	$< 10^{-15}$
4	712	14,758	646	10,713	0.80	[0.72–0.89]	$< 10^{-5}$
5	194	1,335	72	726	1.47	[1.10–1.95]	0.0086

Table C4: Odds ratios for increased severity judgment ($cf < orig$) by triage level (Bordeaux)

At Triage 2 (most urgent after filtering), female counterfactuals are judged less severe (Table C3). At Triage 3 (intermediate situations), a clear and highly significant effect favors F→M being judged more severe than M→F (OR 0.62, 95% CI [0.55–0.69], $p < 10^{-15}$) (Table C4): male sex as counterfactual increases the probability of being judged more severe compared to female. Triage 4 shows

the same direction as Triage 3, with a more moderate but significant effect (OR 0.80, $p < 10^{-5}$). At Triage 5 (least urgent), the pattern reverses: M→F is more often judged "more severe" than F→M (OR 1.47, 95% CI [1.10–1.95], $p = 0.0086$).

When studying the gender effect through our counterfactual approach with Bordeaux University Hospital data, we observe overall a higher probability for women to be judged less severe. When examining in detail by triage level, this effect is reinforced at severe (with potentially more consequences) and moderate triage levels, while it is reversed at the consultation triage level (which implies longer waiting times but without significant clinical consequences).

C.2 MIMIC-IV

We performed the same analysis on MIMIC-IV. Table C5 shows the proportion of cases where the counterfactual was judged less severe ($cf > orig$) or more severe ($cf < orig$) than the original, stratified by sex and triage level.

Triage Score	$cf > orig$ (M)	$cf < orig$ (M)	$cf > orig$ (F)	$cf < orig$ (F)	Δ M-F $cf >$	Δ M-F $cf <$
1	25.9%	0.0%	26.1%	0.0%	-0.2	0.0
2	24.1%	5.3%	22.1%	6.6%	+2.0	-1.3
3	0.4%	25.2%	0.4%	29.2%	0.0	-4.0
4	0.0%	91.7%	0.0%	96.8%	0.0	-5.1

Table C5: Percentage of discordant predictions by triage level and sex (MIMIC-IV)

To quantify the strength of gender effects, we calculated odds ratios (OR) for the probability that the counterfactual was judged more severe than the original ($cf < orig$). Table C6 presents these results.

Triage Sc.	$cf < orig$ (M)	$cf \geq orig$ (M)	$cf < orig$ (F)	$cf \geq orig$ (F)	OR (M/F)	IC95%	p-val
1	0	2,562	0	1,895	—	—	—
2	536	9,632	515	7,228	0.80	[0.70–0.91]	0.001
3	2,159	6,417	2,106	5,099	0.81	[0.75–0.88]	<0.001
4	33	3	30	1	—	—	—

Table C6: Odds ratios for increased severity judgment ($cf < orig$) by triage level (MIMIC-IV)

The effect of gender varies according to triage level. At Triage 1 (most critical), no gender effect is observed: judgments are identical regardless of sex. At Triage 2, a slight tendency emerges where female counterfactuals (M→F) are judged less severe and male counterfactuals (F→M) more severe than their originals (+2.0 vs -1.3 percentage points). The most pronounced effect occurs at Triage 3, representing clinically ambiguous situations: changing from female to male (F→M) increases the probability of being perceived as more severe by approximately 25% (OR = 0.81, 95% CI [0.75–0.88], $p < 0.001$). This effect disappears again when severity is obvious (Triage 1) or trivial (Triage 4, though insufficient data preclude definitive conclusions).