Uncovering Judgment Biases in Emergency Triage: A Public Health Approach Based on Large Language Models

Ariel Guerra-Adames[1,2,3] ARIEL.GUERRA-ADAMES@U-BORDEAUX.FR
Marta Avalos-Fernandez[1,3] MARTA.AVALOS-FERNANDEZ@U-BORDEAUX.FR
Océane Doremus[1,2] OCEANE.DOREMUS@U-BORDEAUX.FR
Cédric Gil-Jardiné[1,2,4] CEDRIC.GIL-JARDINE@CHU-BORDEAUX.FR
Emmanuel Lagarde[1,2] EMMANUEL.LAGARDE@U-BORDEAUX.FR

[1] *University of Bordeaux, BPH Research Center, UMR U1219, INSERM, F-33000, Bordeaux, France*
[2] *AHeaD Team, BPH INSERM, F-33000, Bordeaux, France*
[3] *SISTM team, Inria centre at the University of Bordeaux, F-33405, Talence, France*
[4] *University Hospital of Bordeaux, Pole of Emergency Medicine, F-33000, Bordeaux, France*

Abstract

Judgment biases in emergency triage can adversely affect patient outcomes. This study examines sex/gender biases using four advanced language models fine-tuned on real-world emergency department data. We introduce a novel approach based on the testing method, commonly used in hiring bias detection, by automatically altering triage notes to change patient sex references. Results indicate a significant bias: female patients are assigned lower severity ratings than male patients with identical clinical conditions. This bias is more pronounced with female nurses or when patients report higher pain levels but diminishes with increased nurse experience. Identifying these biases can inform interventions such as enhanced training, protocol updates, and machine learning tools to support clinical decision-making.

Keywords: Public Health, Detecting and Measuring Human Bias, Health Inequity, Impact of Sex/Gender on Health, Responsible AI

Data and Code Availability Data for this study was collected from the Adult Emergency Department of the Bordeaux University Hospital. Because of patient confidentiality and our current agreement with the Bordeaux University Hospital, this data cannot be shared with others, and will therefore not be made available with this paper. Code for this paper will also not be made available, as some of the main scripts in the source repository also contain sensitive patient information, with the exception of prompts used which are included in the appendix.

Institutional Review Board (IRB) This study adhered to the MR-004 reference methodology as outlined by the French National Commission on Informatics and Liberty (CNIL), in association with the Phase I of the TARPON project[1] for data collection and processing.

1. Introduction

Emergency triage involves the rapid assessment and categorization of patients based on the severity of their conditions. Upon arrival at the emergency department (ED) of a hospital, a triage nurse evaluates patients by collecting information such as the reason for the visit, vital signs, and medical history. This data helps determine the urgency of the patient's condition, and a triage acuity score is assigned according to a specific scale. Various validated triage scales are used globally, such as the Manchester Triage System (MTS) developed in the United Kingdom (Azeredo et al., 2015), the Emergency Severity Index (ESI) developed in the United States (Tanabe et al., 2004), and the FRench Emergency Nurses Classification in-Hospital triage scale (FRENCH) (Taboulet et al., 2009) in France. The data from the Bordeaux University Hospital used in this study follows a 5-level triage scale.

Accurate triage is critical, as underestimating urgency (under-triage) can delay care and worsen patient outcomes, while overestimating urgency (over-triage) can lead to resource over-utilization and in-

[1] https://www.health-data-hub.fr/partenariats/tarpon

creased costs, particularly during peak periods in EDs (Fernandes et al., 2005). To evaluate the precision of existing triage scales, retrospective cohort studies have compared various scales, reporting accuracies ranging from 56 to 82%, along with differing levels of inter-rater agreement (Tam et al., 2018; Aubrion et al., 2022). Additionally, studies like the one conducted by Zachariasse et al. (2019) mention the role of biases as a factor affecting triage accuracy.

Judgment biases are cognitive shortcuts that individuals use to form opinions or make decisions based on incomplete or over-generalized information. In healthcare, judgment biases are particularly concerning as they can exacerbate inequalities, potentially delaying medical attention for some individuals while misallocating resources to others. Unlike fields such as human resources, where discrimination testing is often employed, ethical and practical constraints make similar testing in healthcare challenging (Croskerry, 2013).

These biases can have a profound impact on critical medical decision-making, such as emergency medical dispatching and emergency triage. For instance, triage nurses in EDs are often under pressure to make rapid decisions, which may be influenced by unconscious biases linked to factors like a patient's sex/gender, age, ethnicity, or insurance status (Peitzman et al., 2023; Essa et al., 2023), which may not necessarily have a clinical relevance. Complex interactions between these variables, such as those between the sex (biological characteristics related to male and female physiology) and gender (sociocultural roles and identities associated with masculinity and femininity) of the patient and that of the healthcare provider (Vigil et al., 2017), further complicate the landscape of potential biases.

Amid these challenges, the application of artificial intelligence (AI) in emergency medicine has gained traction, particularly for improving the efficiency of ED operations. Research has explored various aspects of emergency medicine, including prehospital settings (Rosemarin et al., 2019; Raff et al., 2024; Toy et al., 2024), emergency dispatch systems (Emami and Javanmardi, 2023), patient flow optimization (Arnaud et al., 2022; Liventsev et al., 2021), and emergency triage (Sánchez-Salmerón et al., 2022; Vântu et al., 2023; Defilippo et al., 2023; Mutegeki et al., 2023; Sax et al., 2023; Sears et al., 2024; ?).

Given that much clinical data is free-text, large language models (LLMs) have gained prominence for their efficiency and accuracy (Fraser et al., 2023; Stewart et al., 2023; Preiksaitis et al., 2024; Kaboudi et al., 2024; Masanneck et al., 2024; Huang et al., 2024; Colakca et al., 2024; Hoppe et al., 2024; Williams et al., 2024; Zaboli et al., 2024; Meral, 2024; Kim et al., 2024; Franc et al., 2024; Liu et al., 2024).

However, recent studies suggest that state-of-the-art LLMs like GPT-4 may inadvertently perpetuate racial and gender biases in medical tasks (Kotek et al., 2023; Zack et al., 2024), with some of these biases potentially present in the data indirectly (Chang et al., 2024). Researchers have raised concerns about the risks of premature AI adoption in clinical practice, emphasizing the need for a validated approach to ensure its integration genuinely improves patient care without reinforcing existing disparities (Friedman et al., 2024; Russon et al., 2024).

This study aims to detect and quantify sex/gender biases in ED data from the Bordeaux University Hospital using a workflow based on fine-tuning and few-shot learning with LLMs. We hypothesize that fine-tuned state-of-the-art LLMs can replicate the judgment biases of healthcare professionals when trained on data they have produced, thus serving as a proxy for discrimination testing. Conversely from traditional fairness approaches which aim to mitigate biases within AI systems, our approach uses LLMs to replicate human biases in a decision-making process. The literature presents strategies based on natural language processing (NLP) for assessing gender-related linguistic biases in recommendation letters Fu et al. (2023), gender disparities in scientific reviewing processes Verharen (2023), and job descriptions Frissen et al. (2022). To the best of our knowledge, our approach of using LLMs both to reproduce human behavior, including biases, and to quantify these biases is novel.

2. Related Works

Several studies have highlighted that sociodemographic factors, including ethnicity, sex, age, and health insurance status, contribute to triage errors (Portillo et al., 2023; Peitzman et al., 2023; Essa et al., 2023; Fekonja et al., 2023; Jafari et al., 2024; Puissant et al., 2024). Although notable discrepancies in prioritization based on age and ethnicity are well established, evidence regarding sex and gender remains less definitive (Arslanian-Engoren, 2000; Onal et al., 2022), likely due to nuances that can both favor and disadvantage women in different situations or through interactions with other factors. In Coisy et al. (2023), altering the visualization of simulated patients with chest pain affected prioritization decisions, with black-female patients being less likely to receive prompt care. Gender disparities in prioritization are further complicated by the fact that women often present with symptoms that are "atypical" to established standards for serious conditions, such as strokes, heart attacks, appendicitis, or acute poisonings (Preciado et al., 2021; Mnatzaganian et al., 2020; Lopez et al., 2021).

With the introduction of complex NLP models in the form of LLMs like GPT-4, it has become evident that these architectures are at a high risk of reproducing or amplifying the human biases contained in their training data. In Zack et al. (2024) for example, authors evaluated the potential of GPT-4 to reproduce racial and gender biases in four tasks in the clinical domain, concluding that the selected model does not accurately depict the demographic diversity of clinical scenarios, favoring existing racial and gender stereotypes. Another recent study Schmidgall et al. (2024) which compares four state-of-the-art LLMs affirms that GPT-4 appears to be more resilient to cognitive biases in a new proposed medical bias benchmark based on questions from the United States Medical Licensing Exam. Another benchmark for evaluating biases and equity risks in medical language models is proposed in Pfohl et al. (2024), this one being based on both manually-curated and synthetic clinical questions. The authors use the toolbox to evaluate their own model based on the PaLM model, concluding that a system to assess equity in LLMs such as the one proposed cannot yet determine if a given model can promote equitable health outcomes in practice.

In response to these possible inequities, and those found on other traditional models, the field of fairness in AI focuses on mitigating biases in AI-based systems in order to ensure that they do not replicate discriminatory behaviors across various demographic groups (Ferrara, 2023). However, authors like Buyl and de Bie (2024) argue that fairness in AI cannot be fully achieved through mathematical formalism alone, as AI systems often inherit biases from data and operate in difficult to quantify socio-technical contexts where human decision-making and environmental factors also influence outcomes. They conclude that while AI fairness tools have value, they must be applied with a nuanced understanding of these broader limitations. In clinical contexts, Liu et al. (2023) argue that as technical AI fairness methods focus on equality across demographic subgroups they may not align with clinical realities. For example, differences in outcomes based on age or gender might be justified by clinical factors rather than be indicative of biases. They suggest shifting the focus from "equality" to "equity" in healthcare AI, where fairness is defined as providing appropriate treatment based on individual needs rather than enforcing uniformity across groups. Both authors agree that a deeper understanding of the context surrounding human clinical reasoning is necessary to achieve properly fair AI systems. This is one of the main motivations of our study.

Diverging from work on AI fairness, our approach focuses on using AI systems–in our case LLMs–to study human biases in emergency medical decision-making, taking into account the individual needs of patients and the context surrounding patients and caregivers by including verbal information in the form of triage nurse notes. Our findings can then be used to help both healthcare professionals and their potential AI counterparts to make fairer decisions in emergency triage. AI fairness would, however, be an important foundation for a later stage of our project when decision-making assistant tools are developed.

3. Materials

The data used for this study comprises approximately 480,000 entries to the Adult ED of the Bordeaux University Hospital, containing reports of admissions having taken place between January 2013 and December 2021, each with its corresponding assigned triage score. These reports contain a number of parameters from each admission, such as the exact moment of the admission, the sex of the patient, the chief complaint, history of present illness, past medical history, an array of vital signs (heart rate, respiratory rate, blood pressure, among others), and the associated triage score. Additionally, information related to the triage nurses themselves was also collected and used to enrich the dataset, including sex of the triage nurse, number of years of experience at the date of the triage, and whether or not they have received specialized triage training at the date of the triage. Data were de-identified prior to analysis according to the General Data Protection Regulation (GDPR) and relevant European regulations, ensuring the protection of personal data and the privacy of participants (Dorémus et al., 2024). A small overview of the data and the distributions of some of its variables of interest can be found in the Appendix C.

The current 5-level triage scale and corresponding triage criteria used in the Bordeaux University Hospital were introduced in October 2014, with a full adoption by the 31st of December 2014. Because of this change, an initial filtering was performed to exclude all records from before January 2015, assuring a normalized triage scale. Additionally, a portion of samples were found to have missing values in its vital parameters fields due to triage nurses not taking them on admission as they may have considered them to not be relevant at a given moment. As vital parameters are important for the context of the triage process, these samples were removed, as well as samples which do not contain the sex of the patient or the sex of the triage nurse.

Because some specific admission motives or pre-existing conditions are sex-specific (for example, a prostate cancer can only affect biological males, menstrual pain can only affect biological women), all corresponding samples were also filtered out. This filtering was performed by verifying the past medical history, chief complaint, and history of present illness of each patient. The full list of words used for filtering and their variations can be found on the Appendix B.

Finally, according to information provided by triage nurses in a previous qualitative study (Avalos et al., 2024), samples that resulted in a triage score of 1 (resuscitation, immediate care needed) were also excluded, as these are unlikely to be subject to bias due to the life-threatening nature of the condition. Moreover, they represent a significant class imbalance problem. The succession of these filtering processes resulted in a sample size of 151,294. This sample was subsequently randomly divided into a train and test partition of equal sizes (75,647 samples each). We verified that, although randomly built, the patient's and nurse's sex, as well as the triage distribution, in both the train and test sets, were comparable to their distribution in the full dataset. A diagram describing the full pre-filtering process with the corresponding sample sizes is shown in the Appendix B.

Since we are working with LLMs, all the filtered tabular data was arranged into free-text strings with their respective variable names, which can then be appended to a prompt instructing to perform the triage. A synthetic example illustrating the resulting free-text format of each sample is provided as follows, along with its English translation.

Synthetic example in French:

Num. adm.: 1632467657 Jour de la semaine: Jeudi Heure: nuit Date heure: 2016-03-03 21:03:00 Sexe patient(e): M Age patient(e): 29 Sexe de l'infirm.: F Experience de l'infirm. (ans): 9 Nb. annnées autonomie de l'infirm.: 1 Motif de la visite: Otalgie Anamnese infirm.: Baisse de l'audition depuis 3,4 jours, travaille dans le batiment, aurait pris des poussières dans les oreilles. Pas de douleur. Saturation O2: 99 Detresse respiratoire?: non Freq. card.: 80 Tension systolique: 130 Tension diastolique: 80 Tension moyenne: 96.67 Douleur: 4 Temperature: 36,5

Translation into English:

Adm. Num.: 1632467657 Day of the week: Thursday Hour: night Date time: 2016-03-03 21:03:00 Sex of patient: M Age of patient: 29 Sex of triage nurse: F Exp. of triage nurse (yrs): 9 Num. of years of autonomy of nurse: 1 Chief complaint: Otalgia History of present illness: Hearing loss for 3-4 days, works in the building, might have gotten dust in the ears. No pain. O2 saturation: 99 Respiratory distress?: no Heart rate: 80 Systolic blood pressure: 130 Diastolic blood pressure: 80 Mean arterial pressure: 96.67 Pain: 4 Temperature: 36.5

4. Methods

Our proposed methodology aims at uncovering any possible judgment biases in emergency triage through two main tasks: accurately predicting triage scores by fine-tuning a LLM on human annotated triage notes, and automatically changing the variable of interest in triage notes with another LLM in a few-shot configuration.

4.1. Predicting Nurse's Triage with LLMs

The first task in our methodology consists of fine-tuning a medium-sized pre-trained LLM to predict the triage score of patients given all the information contained in each admission report, including patient-related information collected by the nurse, as well as data characterizing the nurse responsible for assigning the triage score. This last detail is of critical importance because we need to distinguish the behavior of each human evaluator (the nurses) in order to identify possible individual variables associated with specific judgment biases. The act of fine-tuning a model over the ensemble of all individual human evaluators without identifying them would result in a single evaluator which would better generalize the data, but which would fail to accomplish our goal of reproducing individual judgments. To illustrate this point, Figure 7 in the Appendix shows the mean triage score of the 50 triage nurses who triaged the most patients in the data before filtering (253,975 samples), with scores varying from 3.1 to almost 3.8. In a way, we could say that we aimed to 'overfit' our models on the existing data for the purpose of our study, rather than to make a generalist model which generalizes well over all the individual human experts.

For this study, we chose to compare four recently released LLMs: Mistral 7B [2], BioMistral 7B Labrak et al. (2024), Mixtral 8x7B [3] and Llama 3 8B [4]. These four models were selected mainly due to their size, which allows us to fine-tune and evaluate their quantized versions on our dataset within a reasonable timeframe. Additionally, they can be used locally, a necessary condition given the security requirements for handling health data. Furthermore, BioMistral 7B was selected as it is a version of Mistral 7B which was further pre-trained on medical question answering in languages including French. We chose to work with the 'Instruct' variant of all models, with a common prompt instructing the model to assign a triage score from 2 to 5 given the free-text clinical notes containing all the variables of interest. The model then returns a single token on top of the prompt, containing the predicted triage score. Details regarding the fine-tuning hyperparameters can be found on Appendix E. We did not consider comparing LLMs with other non-transformer-based models for the task of automated triage, as our previous work (Davis, 2022) showed that transformer-based architectures outperformed other NLP methods for this task.

Having fine-tuned the models and saved the adapters, the models were evaluated using the test partition of the data, calculating both classification and regression metrics, as the emergency triage score could be treated as either. For the regression metric, we calculated the quadratically weighted kappa (κ) score, as it is used to assess inter-observer reliability among triage nurses (Worster et al., 2004), as well as the mean absolute error to quantify the average magnitude of errors in the predictions without considering their direction. For classification metrics, we obtained precision, recall, specificity, and F1 scores (one-vs-all scores).

4.2. Transforming triage notes with LLMs

The second aspect of our work concerns the transformation of certain aspects of a patient's clinical notes using pre-trained LLMs in a few-shot learning modality. This transformation concerns only references to the sex of the patient in a triage note, but could be extended to automatically manipulate any aspect of said triage notes. To achieve this, three columns from the complete tabular version of the test data (sex of patient, history of present illness and past medical history) were employed to assemble free-text extracts containing references to the sex of patients. These can then be passed to a pre-trained LLM instructed to perform the changes in the text. An example of the desired output is shown as follows:

Original extract:
Sexe patient(e): M, Anamnèse patient(e) : Patient qui se plaint d'une d thoracique ECG fait et vu par medecin en box : GSC 14 , [...] ne se dit pas gêné pour respirer , ps de diarrhée

Transformed extract:
Sexe patient(e): F, Anamnèse patient(e) : Patiente qui se plaint d'une d thoracique ECG fait et vu par medecin en box : GSC 14 , [...] ne se dit pas gênée pour respirer , ps de diarrhée

In this example, shown only in French, words which are associated to the sex of a patient are highlighted in colors blue and red: *sexe patient(e): M/F* (meaning sex of the patient), *patient/patiente* (meaning the patient [male] versus the patient [female], or *gêné/gênée* (meaning [he/she] does not feel discomfort while breathing).

To accurately perform this automatic transformation on the entire test partition, we used the Mixtral 8x22B[5] SMoE model, quantized to 4-bit precision in a 7-shot learning configuration. This means that we included seven manually-annotated examples of the desired transformations on the prompt of the model, between the instruction and the notes to be changed. The model was instructed to respond in a comma-separated format, allowing us to store the

[2] mistralai/Mistral-7B-Instruct-v0.2
[3] mistralai/Mixtral-8x7B-Instruct-v0.1
[4] meta-llama/Meta-Llama-3-8B-Instruct
[5] mistralai/Mixtral-8x22B-v0.1

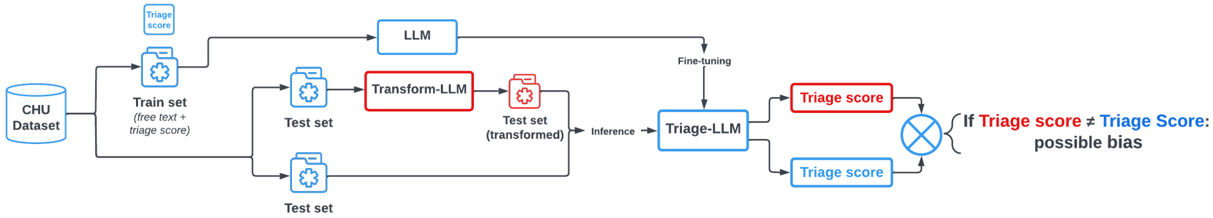

Figure 1: Downstream workflow to quantify biases in emergency triage.

transformed data following the same structure as the original data.

4.3. Finding Judgment Biases

Having established how to perform both emergency triage and triage note transformation with LLMs, we introduce the third main methodological contribution of this study: the downstream task of judgment bias detection. This task consists of predicting the triage scores for the entire test dataset, both before and after the transformations. If our original hypothesis, stating that LLMs can accurately imitate human patterns in emergency triage from text, is correct, then any existing biases related to the patient's sex in the data should cause significant differences between the two sets of predicted triage scores. The entire proposed workflow is illustrated in Figure 1.

Let us define x as an original triage note and x^* as a modified triage note, where the patient's sex and references to the patient's sex have been altered. Let $\mathcal{X}$ denote the set of all triage notes. Consider the function $g : \mathcal{X} \to \mathcal{X}$ as the transformation applied to x to obtain x^*, i.e., $x^* = g(x)$.

Let y and $\widehat{y} \in \{2, 3, 4, 5\}$ represent, respectively, the original triage score and the predicted triage score derived from the original triage notes x using a fine-tuned LLM f, such that $\widehat{y} = f(x)$. Similarly, let $\widehat{y^*} \in \{2, 3, 4, 5\}$ represent the predicted triage score based on the modified triage notes x^*, i.e., $\widehat{y^*} = f(x^*)$.

The objective of this downstream task is to compare the predicted triage scores $\widehat{y}$ and $\widehat{y^*}$ to identify significant differences that may indicate judgment biases. We aim to detect whether the transformation g in the triage notes results in systematic differences in the predicted triage scores due to the alteration in the triage notes:

$$H_0^* : \widehat{y} - \widehat{y^*} = 0 \quad \text{(no bias)}$$

$$H_1^* : \widehat{y} - \widehat{y^*} \neq 0 \quad \text{(possible presence of bias)}$$

If we reject H_0^* and confirm a significant difference between $\widehat{y}$ and $\widehat{y^*}$, it indicates potential biases related to the altered variable. However, this step alone does not fully validate the method nor confirm the existence of biases. To address this, we have designed a downstream task evaluation, detailed in the following subsection. This evaluation takes into account the conditions under which the transformation g operates, ensuring that any detected bias can be attributed to the applied transformation.

4.4. Downstream Task Evaluation

If we reject H_0^*, the next step involves attempting to replicate the same method in the reverse direction. This means that we will re-transform the already transformed triage notes from the test partition using the same method described in Section 4.2, and infer a triage score once more. We can then carry out the same type of comparison described in the previous subsection, which should allow us to observe the same type of bias, confirming an alternative hypothesis H_1^{**} that what was detected by the system was indeed bias related to the changed variable.

We can express this as $\widehat{y^{**}} = f(x^{**})$, where x^{**} represents the re-transformed triage notes, and $\widehat{y^{**}}$ is the predicted triage score based on these notes. We then test:

$$H_0^{**} : \widehat{y^{**}} - \widehat{y^*} = 0 \quad \text{(not confirmed)}$$

$$H_1^{**} : \widehat{y^{**}} - \widehat{y^*} \neq 0 \quad \text{(confirmation of bias)}$$

Rejecting H_0^{**} would strengthen the evidence that the detected bias is indeed related to the variable we manipulated (in this case, the patient's sex). Additionally, any discrepancies between the back-transformed triage notes x^{**} and the original triage notes x could indicate issues in the transformation process.

5. Results

Once fine-tuned, all models were evaluated on the test partition of the data. This meant they were instructed to predict a label $\widehat{y}$ which would be compared to its human counterpart y. The results, as presented in Table 1, demonstrate that the Mistral 7B and Llama 3 8B models achieved the highest precision, recall, F1-scores, all at 0.67, and quadratically weighted Cohen's kappa, κ at 0.71. They also tied for the lowest MAE at 0.34. The BioMistral 7B model, though slightly behind, performed closely with precision and recall at 0.66, F1-score at 0.65, κ at 0.69, and MAE at 0.35. The Mixtral 8x7B, despite its longer training time and larger parameter count, lagged with the lowest precision (0.64), recall (0.62), F1-score (0.61), κ (0.65), and highest MAE (0.39).

Model Name	Precision	Recall	Specificity	F1	κ	MAE
Mistral 7B	**0.67**	**0.67**	0.81	**0.67**	**0.71**	**0.34**
BioMistral 7B	0.66	0.66	0.81	0.65	0.69	0.35
Mixtral 8x7B	0.64	0.62	0.79	0.61	0.65	0.39
Llama 3 8B	**0.67**	**0.67**	**0.82**	0.66	**0.71**	**0.34**

Table 1: Classification and regression metrics for predicting triage scores from the test data.

The normalized confusion matrix of the classification results for the fine-tuned Mistral 7B model in Figure 2 indicates that, in most cases where the model misclassifies patients, the predicted triage level is typically either one level higher or lower than the ground truth. These results also hint at the possible consequences of having imbalanced classes in our train set, where categories which are better predicted are those with a bigger sample size (samples with triage score 3 and 5).

5.1. Transformation and evaluation of biases

The Mixtral 8x22B model was successfully used to transform tabular and free-text references to the sex of patients in the data. While we did explore ways to systematically evaluate the text transformation as it would require us to manually label a subset of triage notes, we empirically manually inspected 100 original and transformed triage notes pairs, and found that the model was able to change all 100 tabular references to a patient sex, and 98 of the free-text references. In the two cases where the model made a mistake, the notes were long and it was not clear whether the triage nurse was referring to the patient or another person who was with the patient. Having transformed all triage notes towards the opposite sex, and having established that the system works satisfactorily, we proceeded to evaluate the impact of the transformation.

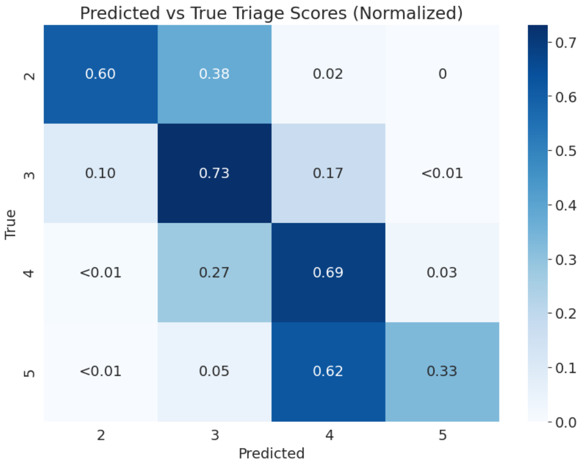

Figure 2: Normalized confusion matrix of classification results.

We performed a paired t-test between the triage scores predicted by the fine-tuned LLM before and after the sex transformation, concluding that there was a statistically significant difference between the two sets of predicted labels, and that the probability of observing such a difference by random chance is virtually nonexistent (p-value = 1.4524e-24).

To further explore this result, we proceeded to quantify the extent of this difference by calculating the change in the predicted triage scores based on the sex of the patient before and after transformation. As illustrated in Figure 3, several points need to be highlighted.

First, when transforming triage notes originally belonging to male patients towards the opposite sex (right side of the figure), we observed a higher percentage of samples triaged as less critical ($\approx$ 5%) compared to those which were triaged as more critical (1.81%) after the transformation. We interpret this difference in percentages as female patients being more likely to be under-triaged with respect to their male counterparts for the same clinical conditions. When looking at the percentages of over- and under-triage of notes originally belonging to female patients transformed towards the opposite sex (left side of the figure), the differences are less pronounced. We observe a higher percentage of cases where the transfor-

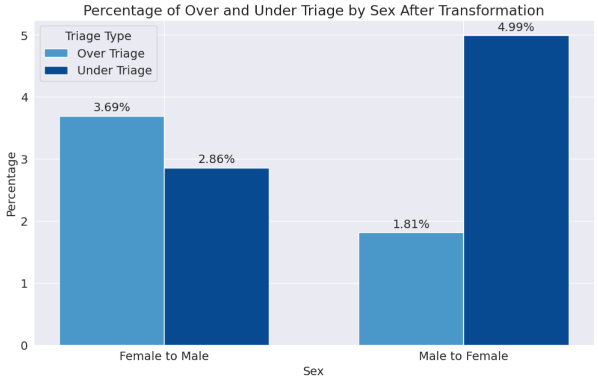

Figure 3: Percentage of over- and under-triaged samples after artificially transforming references to the sex of patients.

mation resulted in a more critical triage score (3.69%) versus those which were triaged less critical (2.86%). We interpret this difference as male patients being slightly more likely to be over-triaged with respect to their female counterparts for the same clinical conditions. Additional stratified analyses according to the sex of triage nurse, the expression of pain, and the experience of triage nurses can be found in the Appendix G, along with the mean triage scores per sex of patients before and after triage. Biases which tended towards under-triaging female patients were more pronounced when triage was performed by female nurses and when the patient expressed higher pain levels, and attenuated by triage nurses with more years of experience.

6. Conclusion

This study explores the presence of biases in emergency triage employing state-of-the-art LLMs to replicate and quantify these biases based on real-life data from a hospital. Our primary hypothesis, that LLMs can accurately mirror the judgment biases of triage nurses when fine-tuned with data from real admissions to an emergency department, was tested and substantiated through an innovative approach that, to the best of our knowledge, has never been implemented.

The evaluation of the selected models demonstrate that LLMs, particularly Mistral 7B and Llama 3 8B, are capable to imitate human emergency triage with a satisfactory accuracy. These models were able to capture patterns in the data that reflect human decision-making, including biases related to sociodemographic factors. Notably, the transformation experiments revealed that female patients are more likely to be under-triaged compared to their male counterparts presenting with the same clinical conditions. This trend was more pronounced when triage was performed by female nurses and when the patient reported higher pain levels, but it was attenuated when triage was conducted by nurses with more years of experience.

Future work should focus on refining the fine-tuning methodology through hyperparameter optimization, and the incorporation of non-verbal cues in the triage process through the introduction of multimodal data. Additionally, establishing a triage 'silver standard' through expert consensus could enhance the evaluation of the models and provide a more robust framework for assessing triage accuracy. Beyond model improvement, a deeper analysis of the identified biases will be crucial to determine whether these patterns are indeed biased decision-making or rooted in other contextual factors. This follow-up would explore interventions both at the machine level—by adjusting model architectures—and at the human level—through enhancing training and revising protocols for healthcare workers.

Another important step in a follow-up study would be to explore potential reasons for the asymmetric effect of transformation observed in Figure 3. We interpret this asymmetry as probably explained by a higher prevalence of severe cases for certain conditions in one sex, resulting in imbalanced data representation learned by the model during fine-tuning, or by residual, implicit gender indicators in the text, which may vary according to the gender of the patients. In connection with this point, it will be important to further analyze the data by stratifying according to diagnosis or the type of discharge from the emergency department. Additionally, a previous descriptive statistical analysis (Avalos et al., 2024) highlighted a potential nuance in gender biases: men tend to receive slightly more urgent scores, while women may receive somewhat more urgent scores for consultation visits. This analysis was adjusted for the sex of the nurses and the age of the patients but did not ensure comparability of clinical conditions. Nevertheless, this remains a lead to explore further.

By laying the groundwork for a deeper understanding of how LLMs can help detect and mitigate biases, this study contributes to the broader goal of improving fairness and patient outcomes in emergency medicine. Findings from this work can be applied to triage decision-making through targeted interventions, raising awareness about potential biases and promoting equitable clinical judgment.

7. Discussion

While the results of this study are encouraging and innovative, several areas still require improvement. The first major factor is the accuracy of the models. Emergency triage is an inherently challenging task, largely because not all the information available to the nurse during triage is captured in the recorded data we have access to. Many critical aspects of triage are non-verbal, such as patients' facial expressions or the tone of their voice, and these are rarely documented in the triage notes. One potential improvement could be to record and transcribe nurse-patient interactions during triage, incorporating this additional data into the triage notes in a prospective study. A similar approach was used in Pilleron et al. (2024), where theoretical triage scores were retrospectively assessed using audio recordings, observer notes, and nurses' handover reports.

The absence of standardized approaches for evaluating triage systems and the lack of clear guidelines on the frequency and methods for auditing triage practitioners make it challenging to establish a universally accepted 'gold standard' for triage (Zaboli, 2024). Each country—and even individual hospitals within France—follows its own protocols. Introducing a 'silver standard' for triage could enhance future studies. The variability in triage systems and the logistical burden of conducting comprehensive audits complicate the evaluation of ML-based models against multiple human evaluators, as the evaluators themselves may lack consistency. A future study could tackle this issue by assembling a panel of medical experts to assess a set of data points based on their expertise, thus helping to expand and refine the triage criteria for future applications.

An additional area for improvement is extending the fine-tuning duration and optimizing hyperparameters using a validation partition. Due to time constraints, the models were evaluated at the specified epochs. However, since then, we have extended the training periods, leading to observed improvements in classification accuracy. This indicates that given more time, the models can still achieve slightly better performance.

Bias mitigation in emergency triage To mitigate biases in triage decision-making, targeted interventions can be implemented through training programs, simulations, and serious games aimed at raising awareness and promoting equitable clinical judgment. Training should focus on sensitizing healthcare professionals to implicit gender biases by incorporating case-based learning and structured debriefings that highlight disparities in triage scoring between male and female patients for identical clinical conditions.

An initial triage training is mandatory in France for nurses that practice triage in ED. Training strategies have to include simulations with a some special focus on patient scenarios with varied gender presentations to help nurses recognize and adjust biased behaviors in real time. For continuous training, a before-and-after learning strategy can be constructed, combining initial assessments, targeted training, and post-training evaluations. The approach would begin with a baseline evaluation of triage nurses' decisions through simulated scenarios integrating various patient genders and reason for their visit. Following this, structured training sessions would focus on raising awareness of implicit gender and other biases (racial, socioeconomic, etc.). The objective of these sessions would be to provide tools to mitigate them. After the intervention, a second round of simulations would assess changes in performance, ensuring that nurses integrate unbiased practices into their routine triage.

In this context, serious games that simulate triage scenarios could be developed to allow nurses to reinforce unbiased clinical decision-making in a controlled setting. The advantage of this approach lies in its production of immediate feedback on their performance. These approaches, combined with continuous assessment and feedback, could contribute to reducing bias and improving fairness in emergency triage.

Acknowledgments

The authors would like to thank all nurses who participated in the study.

References

Emilien Arnaud, Mahmoud Elbattah, Christine Ammirati, Gilles Dequen, and Daniel Aiham Ghazali. Use of artificial intelligence to manage patient flow in emergency department during the covid-19 pandemic: a prospective, single-center study. *International Journal of Environmental Research and Public Health*, 19(15):9667, 2022.

Cynthia Arslanian-Engoren. Gender and age bias in triage decisions. *Journal of Emergency Nursing*, 26 (2):117–124, 2000. ISSN 0099-1767.

Antoine Aubrion, Romain Clanet, JP Jourdan, Christian Creveuil, E Roupie, and Richard Macrez. FRENCH versus ESI: comparison between two nurse triage emergency scales with referent scenarios. *BMC Emergency Medicine*, 22, 12 2022.

Marta Avalos, Dalia Cohen, Dylan Russon, Melissa Davids, Océane Dorémus, Gabrielle Chenais, Eric Tellier, Cédric Gil-Jardiné, and Emmanuel Lagarde. Detecting human bias in emergency triage using llms: Literature review, preliminary study, and experimental plan. In *The International FLAIRS Conference Proceedings*, volume 37, 2024.

Thereza Raquel Machado Azeredo, Helisamara Mota Guedes, Ricardo Alexandre Rebelo de Almeida, Tânia Couto Machado Chianca, and José Carlos Amado Martins. Efficacy of the manchester triage system: a systematic review. *International emergency nursing*, 23(2):47–52, 2015.

Maarten Buyl and Tijl de Bie. Inherent limitations of ai fairness. *Communications of the ACM*, 67(2): 48–55, 2024.

Trenton Chang, Mark Nuppnau, Ying He, Keith E Kocher, Thomas S Valley, Michael W Sjoding, and Jenna Wiens. Racial differences in laboratory testing as a potential mechanism for bias in AI: A matched cohort analysis in emergency department visits. *PLOS Global Public Health*, 4(10):e0003555, 2024.

Fabien Coisy, Guillaume Olivier, François-Xavier Ageron, Hugo Guillermou, Mélanie Roussel, Frédéric Balen, Laura Grau-Mercier, and Xavier Bobbia. Do emergency medicine health care workers rate triage level of chest pain differently based upon appearance in simulated patients? *European Journal of Emergency Medicine*, pages 10–1097, 2023.

Cansu Colakca, Mehmet Ergın, Habibe Selmin Ozensoy, Alp Sener, Selahattin Guru, and Ayhan Ozhasenekler. Emergency department triaging using ChatGPT based on emergency severity index principles: a cross-sectional study. *Scientific reports*, 14(1):22106, 2024.

Pat Croskerry. From mindless to mindful practice—cognitive bias and clinical decision making. *N Engl J Med*, 368(26):2445–2448, 2013.

Melissa Davis. Development and validation of an automated triage system in the emergency department of Bordeaux University Hospital. Master's thesis, Université de Bordeaux, France, 2022. Available at https://dumas.ccsd.cnrs.fr/dumas-03856402v1.

Annamaria Defilippo, Giuseppe Bertucci, Cosimo Zurzolo, Pierangelo Veltri, and Pietro Guzzi. On the computational approaches for supporting triage systems. *Interdisciplinary Medicine*, 1(3): e20230015, 2023.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

Océane Dorémus, Dylan Russon, Benjamin Contrand, Ariel Guerra-Adames, Marta Avalos, Cédric Gil-Jardiné, and Emmanuel Lagarde. Harnessing moderate-sized language models for reliable patient data de-identification in emergency department records: An evaluation of strategies and performance (preprint). 2024. doi: 10.2196/preprints.57828.

Payam Emami and Karim Javanmardi. Enhancing emergency response through artificial intelligence in emergency medical services dispatching; a letter to editor. *Archives of Academic Emergency Medicine*, 11(1), 2023.

Changaiz Dil Essa, Gideon Victor, Sadia Farhan Khan, Hafisa Ally, and Abdus Salam Khan. Cognitive biases regarding utilization of emergency severity index among emergency nurses. *The American Journal of Emergency Medicine*, 73:63–68, 2023.

Zvonka Fekonja, Sergej Kmetec, Urška Fekonja, Nataša Mlinar Reljić, Majda Pajnkihar, and Matej Strnad. Factors contributing to patient safety during triage process in the emergency department: A systematic review. *Journal of Clinical Nursing*, 32, 01 2023.

Christopher MB Fernandes, Paula Tanabe, Nicki Gilboy, Loren A Johnson, Rebecca S McNair, Alexander M Rosenau, Peter Sawchuk, David A Thompson, Debbie A Travers, Nancy Bonalumi, et al. Five-level triage: a report from the acep/ena five-level triage task force. *Journal of Emergency Nursing*, 31(1):39–50, 2005.

Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3, 2023.

Jeffrey Michael Franc, Lenard Cheng, Alexander Hart, Ryan Hata, and Atilla Hertelendy. Repeatability, reproducibility, and diagnostic accuracy of a commercial large language model (ChatGPT) to perform emergency department triage using the Canadian triage and acuity scale. *Canadian Journal of Emergency Medicine*, 26(1):40–46, 2024.

Hamish Fraser, Daven Crossland, Ian Bacher, Megan Ranney, Tracy Madsen, and Ross Hilliard. Comparison of diagnostic and triage accuracy of ada health and webmd symptom checkers, chatgpt, and physicians for patients in an emergency department: Clinical data analysis study. *JMIR Mhealth Uhealth*, 11:e49995, Oct 2023.

Ari B Friedman, M Kit Delgado, and Gary E Weissman. Artificial intelligence for emergency care triage—much promise, but still much to learn. *JAMA Network Open*, 7(5):e248857–e248857, 2024.

Richard Frissen, Kolawole Adebayo, and Rohan Nanda. A machine learning approach to recognize bias and discrimination in job advertisements. *AI & SOCIETY*, 38:1–14, 10 2022.

Sunyang Fu, Darren Q Calley, Veronica A Rasmussen, Marissa D Hamilton, Christopher K Lee, Austin Kalla, and Hongfang Liu. Gender-based language differences in letters of recommendation. *AMIA Summits on Translational Science Proceedings*, 2023:196, 2023.

John Michael Hoppe, Matthias K Auer, Anna Strüven, Steffen Massberg, and Christopher Stremmel. ChatGPT with GPT-4 outperforms emergency department physicians in diagnostic accuracy: Retrospective analysis. *J Med Internet Res*, 26:e56110, 2024.

Ting-Yun Huang, Chee-Fah Chong, Heng-Yu Lin, Tzu-Ying Chen, Yung-Chun Chang, and Ming-Chin Lin. A pre-trained language model for emergency department intervention prediction using routine physiological data and clinical narratives. *International Journal of Medical Informatics*, 191: 105564, 2024. ISSN 1386-5056.

Kaileen Jafari, Brian Burns, Dwight Barry, Cassandra Koid, Tina Tan, and Emily Hartford. Triage discordance in an academic pediatric emergency department and disparities by race, ethnicity, and language for care. *Pediatric Emergency Care*, 40 (10):681–687, 2024.

Navid Kaboudi, Saeedeh Firouzbakht, Mohammad Shahir Eftekhar, Fatemeh Fayazbakhsh, Niloufar Joharivarnoosfaderani, Salar Ghaderi, Mohammadreza Dehdashti, Yasmin Mohtasham Kia, Maryam Afshari, Maryam Vasaghi-Gharamaleki, et al. Diagnostic accuracy of chatgpt for patients' triage; a systematic review and meta-analysis. *Archives of Academic Emergency Medicine*, 12(1): e60, 2024.

Jae Hyuk Kim, Sun Kyung Kim, Jongmyung Choi, and Youngho Lee. Reliability of chatgpt for performing triage task in the emergency department using the korean triage and acuity scale. *Digital Health*, 10:20552076241227132, 2024.

Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701139.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024.

Mingxuan Liu, Yilin Ning, Salinelat Teixayavong, Mayli Mertens, Jie Xu, Daniel Shu Wei Ting, Lionel Tim-Ee Cheng, Jasmine Chiat Ling Ong, Zhen Ling Teo, Ting Fang Tan, et al. A translational perspective towards clinical ai fairness. *NPJ Digital Medicine*, 6(1):172, 2023.

Xiaoni Liu, Rui Lai, Chaoling Wu, Changjian Yan, Zhe Gan, Yaru Yang, Xiangtai Zeng, Jin Liu, Liangliang Liao, Yuansheng Lin, Hongmei Jing, and Weilong Zhang. Assessing the utility of artificial intelligence throughout the triage outpatients: a prospective randomized controlled clinical study. *Frontiers in Public Health*, 12, 2024.

Vadim Liventsev, Aki Härmä, and Milan Petković. Towards effective patient simulators. *Front Artif Intell Appl.*, 4, 2021. ISSN 2624-8212.

Ryan Lopez, Megan Snair, Susana Arrigain, Jesse D Schold, Frederic Hustey, and Laura E Walker. Sex-based differences in timely emergency department evaluations for patients with drug poisoning. *Public Health*, 199:57–64, 2021.

Lars Masanneck, Linea Schmidt, Antonia Seifert, Tristan Kölsche, Niklas Huntemann, Robin Jansen, Mohammed Mehsin, Michael Bernhard, Sven G Meuth, Lennert Böhm, and Marc Pawlitzki. Triage performance across large language models, chatgpt, and untrained doctors in emergency medicine: Comparative study. *J Med Internet Res*, 26:e53297, Jun 2024.

Ateş S. Günay S. Öztürk A. Kuşdoğan M. Meral, G. Comparative analysis of chatgpt, gemini and emergency medicine specialist in esi triage assessment. *The American Journal of Emergency Medicine*, 81: 146–150, 2024.

George Mnatzaganian, Janet E Hiller, George Braitberg, Michael Kingsley, Mark Putland, Melanie Bish, Kathleen Tori, and Rachel Huxley. Sex disparities in the assessment and outcomes of chest pain presentations in emergency departments. *Heart*, 106(2):111–118, 2020.

Henry Mutegeki, Alvin Nahabwe, Joyce Nakatumba-Nabende, and Ggaliwango Marvin. Interpretable machine learning-based triage for decision support in emergency care. In *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 983–990, 2023.

Ege G Onal, Kit Knier, Alexander W Hunt, John M Knudsen, David M Nestler, Ronna L Campbell, Kristine M Thompson, Kharmene L Sunga, Laura E Walker, Bo E Madsen, 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. *Journal of the American College of Emergency Physicians Open*, 3(5):e12792, 2022.

Cassandra Peitzman, Jossie A Carreras Tartak, Margaret Samuels-Kalow, Ali Raja, and Wendy L Macias-Konstantopoulos. Racial differences in triage for emergency department patients with subjective chief complaints. *Western journal of emergency medicine*, 24(5):888, 2023.

Stephen R Pfohl, Heather Cole-Lewis, Rory Sayres, Darlene Neal, Mercy Asiedu, Awa Dieng, Nenad Tomasev, Qazi Mamunur Rashid, Shekoofeh Azizi, Negar Rostamzadeh, et al. A toolbox for surfacing health equity harms and biases in large language models. *arXiv preprint arXiv:2403.12025*, 2024.

Benjamin Pilleron, Delphine Douillet, Yoakim Furon, Carole Haubertin, Elsa Parot-Schinkel, Bruno Vielle, Pierre-Marie Roy, and Laurent Poiroux. Nurses' moral judgements during emergency department triage - a prospective mixed multicenter study. *International Emergency Nursing*, 75: 101479, 2024.

Elyse N Portillo, Chris A Rees, Emily A Hartford, Zachary C Foughty, Michelle L Pickett, Colleen K Gutman, Bashar S Shihabuddin, Eric W Fleegler, Corrie E Chumpitazi, Tiffani J Johnson, et al. Research priorities for pediatric emergency care to address disparities by race, ethnicity, and language. *JAMA Network Open*, 6(11):e2343791–e2343791, 2023.

Salena M Preciado, Adam L Sharp, Benjamin C Sun, Aileen Baecker, Yi-Lin Wu, Ming-Sum Lee, Ernest Shen, Maros Ferencik, Shaw Natsui, Aniket A Kawatkar, et al. Evaluating sex disparities in the emergency department management of patients with suspected acute coronary syndrome. *Annals of emergency medicine*, 77(4):416–424, 2021.

Carl Preiksaitis, Nicholas Ashenburg, Gabrielle Bunney, Andrew Chu, Rana Kabeer, Fran Riley, Ryan Ribeira, and Christian Rose. The role of large language models in transforming emergency medicine: Scoping review. *JMIR Med Inform*, 12:e53787, May 2024.

Madeleine M Puissant, Isha Agarwal, Elizabeth Scharnetzki, Anya Cutler, Hadley Gunnell, and Tania D Strout. Racial differences in triage assessment at rural vs urban maine emergency departments. *Internal and emergency medicine*, pages 1–11, 2024.

Daniel Raff, Kurtis Stewart, Michelle Christie Yang, Jessie Shang, Sonya Cressman, Roger Tam, Jessica Wong, Martin C Tammemägi, Kendall Ho, et al. Improving triage accuracy in prehospital emergency telemedicine: Scoping review of machine learning–enhanced approaches. *Interactive Journal of Medical Research*, 13(1):e56729, 2024.

Hanan Rosemarin, Ariel Rosenfeld, and Sarit Kraus. Emergency department online patient-caregiver scheduling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 695–701, 2019.

Dylan Russon, Marta Avalos, Ariel Guerra-Adames, Cédric Gil-Jardiné, and Emmanuel Lagarde. Ai-driven emergency patient flow optimization is both an unmissable opportunity and a risk of systematizing health disparities: Human bias in ed: Implications for ai-driven patient flow optimization. *The International FLAIRS Conference Proceedings*, 37 (1), 2024.

Rocío Sánchez-Salmerón, José L Gómez-Urquiza, Luis Albendín-García, María Correa-Rodríguez, María Begoña Martos-Cabrera, Almudena Velando-Soriano, and Nora Suleiman-Martos. Machine learning methods applied to triage in emergency services: A systematic review. *International Emergency Nursing*, 60:101109, 2022.

Dana R Sax, E Margaret Warton, Oleg Sofrygin, Dustin G Mark, Dustin W Ballard, Mamata V Kene, David R Vinson, and Mary E Reed. Automated analysis of unstructured clinical assessments improves emergency department triage performance: A retrospective deep learning analysis. *Journal of the American College of Emergency Physicians Open*, 4(4):e13003, 2023.

Samuel Schmidgall, Carl Harris, Ime Essien, Daniel Olshvang, Tawsifur Rahman, Ji Woong Kim, Rojin Ziaei, Jason Eshraghian, Peter Abadir, and Rama Chellappa. Addressing cognitive bias in medical language models. *arXiv preprint arXiv:2402.08113*, 2024.

Kim Sears, Sam Belbin, Elyas Rashno, Drishti Sharma, Kevin Woo, Farhana Zulkernine, Ciprian Daniel Neagu, Bita Amani, and Furkan Alaca. Implementing triage-bot: Supporting the current practice for triage nurses. *Surgical technology international*, 44:sti44–1804, 2024.

Jonathon Stewart, Juan Lu, Adrian Goudie, Glenn Arendts, Shiv Akarsh Meka, Sam Freeman, Katie Walker, Peter Sprivulis, Frank Sanfilippo, Mohammed Bennamoun, et al. Applications of natural language processing at emergency department triage: A narrative review. *Plos one*, 18(12): e0279953, 2023.

Pierre Taboulet, Véronique Moreira, Laurent Haas, Rapheal Porcher, Adelia Braganca, Jean-Paul Fontaine, and Marie-Cecile Poncet. Triage with the french emergency nurses classification in hospital scale: reliability and validity. *European Journal of Emergency Medicine*, 16(2):61–67, 2009.

Hon Lon Tam, Siu Fung Chung, and Chi Kin Lou. A review of triage accuracy and future direction. *BMC Emergency Medicine*, 18:1–7, 2018.

Paula Tanabe, Rick Gimbel, Paul R Yarnold, Demetrios N Kyriacou, and James G Adams. Reliability and validity of scores on the emergency severity index version 3. *Academic emergency medicine*, 11(1):59–65, 2004.

Jake Toy, Jonathan Warren, Kelsey Wilhelm, Brant Putnam, Denise Whitfield, Marianne Gausche-Hill, Nichole Bosson, Ross Donaldson, Shira Schlesinger, Tabitha Cheng, et al. Use of artificial intelligence to support prehospital traumatic injury care: A scoping review. *Journal of the American College of Emergency Physicians Open*, 5(5):e13251, 2024.

Andreea Vântu, Anca Vasilescu, and Alexandra Băicoianu. Medical emergency department triage data processing using a machine-learning solution. *Heliyon*, 9(8), 2023.

Jeroen PH Verharen. ChatGPT identifies gender disparities in scientific peer review. *Elife*, 12:RP90230, 2023.

Jacob Miguel Vigil, Patrick Coulombe, Joe Alcock, Sarah See Stith, Eric Kruger, and Sara Cichowski. How nurse gender influences patient priority assignments in us emergency departments. *Pain*, 158 (3):377, 2017.

Christopher YK Williams, Travis Zack, Brenda Y Miao, Madhumita Sushil, Michelle Wang, Aaron E

Kornblith, and Atul J Butte. Use of a large language model to assess clinical acuity of adults in the emergency department. *JAMA Network Open*, 7(5):e248895–e248895, 2024.

Andrew Worster, Nicki Gilboy, Christopher M Fernandes, David Eitel, Kevin Eva, Rose Geisler, and Paula Tanabe. Assessment of inter-observer reliability of two five-level triage and acuity scales: a randomized controlled trial. *Canadian Journal of Emergency Medicine*, 6(4):240–245, 2004.

Arian Zaboli. Establishing a common ground: the future of triage systems. *BMC Emergency Medicine*, 24(1):148, 2024.

Arian Zaboli, Francesco Brigo, Serena Sibilio, Michael Mian, and Gianni Turcato. Human intelligence versus chat-gpt: who performs better in correctly classifying patients in triage? *The American Journal of Emergency Medicine*, 79:44–47, 2024.

Joany M Zachariasse, Vera van der Hagen, Nienke Seiger, Kevin Mackway-Jones, Mirjam van Veen, and Henriette A Moll. Performance of triage systems in emergency care: a systematic review and meta-analysis. *BMJ open*, 9(5):e026471, 2019.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, et al. 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, 2024.

Appendix A. Triage scale used for the study

Level	Situation	Risk	Delay
1	Life-threatening distress	In the following minutes	Medical att. <1 min
2	Impairment of a vital organ or severe traumatic injury	In the following hours	Medical att. <20 min
3	Unstable or complex functional or lesional impairment	In the following 24 hours	Medical att. <90 min
4	Stable functional or lesional impairment	None	Medical att. <120 min
5	No obvious functional or lesional impairment	None	Medical att. <240 min

Table 2: Triage scale used in the Bordeaux University Hospital.

Appendix B. Pre-filtering

As previously mentioned, all patients admitted to the ED in the Bordeaux University Hospital are assigned a class of chief complaint from a list of 189 possible motives belonging to the ICD-10. Some of these, are clearly indicative of the sex of the patient, as they involve pathologies or sex-specific organs. As a first round of pre-filtering references to a patient's sex which cannot be easily transformed to the opposite sex, we removed samples which had been classified as any of the following classes:

- Recent abdominal pain
- Recent pelvic or genital pain
- Female urogenital problem without pain
- Abdominal mass or distension, ascites
- Female genital bleeding
- Foreign body in the genitourinary tract

This pre-filtering was not enough, as both the past medical history and history of present illness columns sometimes also contain references to a patient's sex. We then performed a regular expression search in the strings of both columns, for rows which contained any of the following words or word segments:

```
uter|utér|ovair|ovarien|vagin|vulv|
mamma|fallope|ivg|prostat|menopause|
ménorrhée|grossesse|fausse couche|
cancer du sein|Gyné|Obsté|gyneco|obste|
règles|hystérectomie|hysterectomie|
endométriose|endometriose|K sein|
ligature trompes|cesarienne|
enceinte|curetage|cezarienne|
testicul|peni|scrot|
pénien|séminale|andropause|
```

The remaining pre-filtering steps as described in Section 3 are shown in Figure 4 along with the corresponding sample sizes at each step.

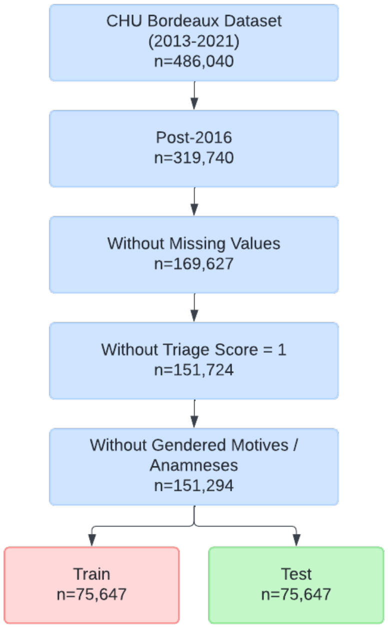

Figure 4: Data pre-processing workflow.

Appendix C. Dataset overview

To illustrate some of the relationships between variables in the data as well as potential sources of biases related to such variables, a brief exploratory data analysis was performed. Many factors which may not necessarily have a clinical relevance have been suspected to impact the assignment of triage scores in EDs. This overview is centered around two types of factors which have been found to influence emergency triage: socio-demographic factors and external factors. It excludes samples from before 2016 for reasons to be explained in the following subsection, as well as a small number of pediatric admissions which were performed in the concerned ED. These visualizations are obtained before normalizing the dataset by eliminating rows with missing values.

The first of these factors, illustrated as the distribution of triage scores attributed to patients by sex, shows a significant class imbalance between triage classes, as illustrated in Table 3. This class imbalance is most evident for the extreme triage scores (1 and 5), as these are much more rarely attributed to patients who arrive at an emergency department.

			Triage Score				
		Total	1	2	3	4	5
Sex	**F**	147 931	590	27 016	61 483	50 110	8 732
	M	166 683	778	31 947	61 504	59 804	12 650

Table 3: Counts of triage scores per sex of patient

This distribution also illustrates another effect mentioned by the literature: substantial discrepancies in triage scores depending on the sex of the patient. While these could be partially explained by the number of men and women which present themselves to an ED, the effect is less clinically explainable when observing the proportion of triage scores attributed by male and female triage nurses, as illustrated in Table 4 and Figure 5.

			Triage Score				
		Total (%)	1	2	3	4	5
Nurse' sex	**F**	247 069	1 077 (0.4)	46 165 (18.7)	97 977 (39.7)	84 852 (34.3)	16 998 (6.9)
	M	65 991	289 (0.4)	12 469 (18.9)	24 409 (37.0)	24 511 (37.1)	4 313 (6.5)

Table 4: Counts of triage score per sex of triage nurse.

These observations may also indicate biases stemming from both the patient's sex and the triage nurse's sex. More specifically, female nurses may assign slightly higher urgency levels, particularly for male patients, compared to their male counterparts, who tend to assign slightly lower urgency levels, regardless of the patient's sex.

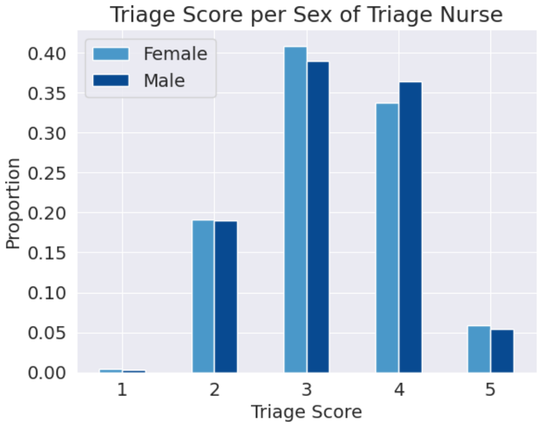

Figure 5: Proportion of triage score per sex of triage nurse.

Another potential source of bias described in the literature is the age of patients. While the age of a patient can sometimes be clinically relevant during triage (e.g., older patients are at higher risk of developing potentially life-threatening complications), this alone does not fully explain the observed distribution shown in Figure 6. In particular, the trend of assigning lower triage scores to younger patients compared to those over 50, while initially logical, may also suggest the presence of biases influencing these scores.

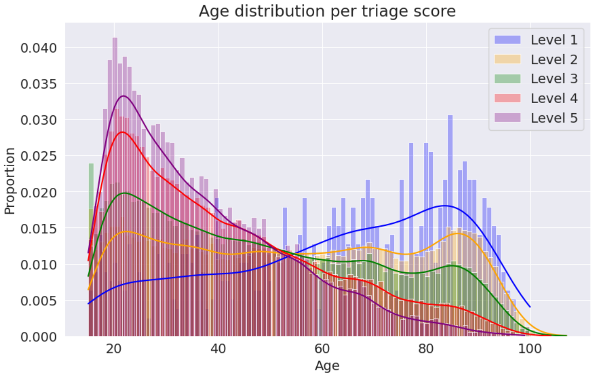

Figure 6: Distribution of ages per triage score.

External factors identified in the literature may influence triage scores. We examined the day of the week as a potential factor, given the observed fluctuations. As shown in Table 5, there is an increased number of admissions on weekends. However, when analyzing the detailed count of admissions by triage score for each day of the week, it becomes evident that on days with higher admission rates, patients may have a higher likelihood of receiving a lower triage score, possibly as a strategy to optimize hospital resources.

		Triage Score					
		1	2	3	4	5	Total
Day of the Week	Monday	195	8 445	18 060	15 723	2 938	45 361
	Tuesday	194	8 523	16 991	14 042	2 790	42 540
	Wednesday	175	8 511	16 975	13 812	2 671	42 144
	Thursday	205	8 330	17 026	14 389	2 704	42 654
	Friday	185	8 343	17 467	14 948	2 703	43 646
	Saturday	193	8 546	18 590	19 172	3 806	50 307
	Sunday	221	8 265	17 878	17 828	3 771	47 963

Table 5: Admissions per day of the week and triage score.

Appendix D. Average triage scores per triage nurse

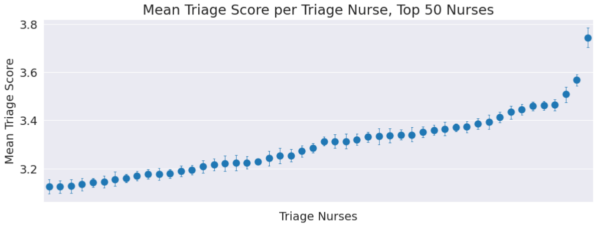

Figure 7: Mean average score of top 50 triage nurses.

Appendix E. Fine-tuning hyperparameters

All fine-tuning was performed using the latest implementation of QLoRA (Dettmers et al., 2024) for PyTorch in the Python programming language. The weights of all models were loaded in 4-bit precision, and all computations were performed on 16-bit floating point precision. The LoRA hyperparameters varied for each model: for the two Mistral 7B variants and LLama 3 we used a LoRA α of 64, LoRA dropout of 0.1, and LoRA r of 8, using LoRA on all layers. For the Mixtral model we used the same parameters, with a LoRA r of 16. Learning rate was fixed as a constant for all models at 5e−5, with a maximum gradient normalization of 0.3, weight decay of 0.001, a warm-up ratio of 0.03, and the AdamW optimizer. Batch sizes were fixed at 12 for the 7 and 8-billion parameter models, and 4 for the Mixtral model, with a maximum sequence length of 1024 tokens per sample for all models. Training was performed until stabilization of the loss, which was empirically found to be 7 epochs for the 7 and 8-billion parameter models, and 5 epochs for the Mixtral model. The models fine-tuned with QLoRa were saved as adapter models, which requires them to be loaded with the original model weights at inference time.

All fine-tuning of models mentioned in this paper were executed on either an Ubuntu 22.04 LTS server with 4 NVIDIA A100 80GB GPUs or a workstation with Ubuntu Desktop and an NVIDIA RTX 3090 24GB GPU. Inference and data filtering were performed on the same workstation with an NVIDIA RTX 3090 GPU, an Intel Xeon processor, and Ubuntu Desktop 22.04 LTS as operating system.

Training times varied in function of model size and number of epochs, as shown in table 6.

Model Name	Parameter No.	Approx. Epochs	Total Time
Mistral 7B	7 Billion	7	250h
BioMistral 7B	7 Billion	5	160h
Mixtral 8x7B	47 Billion	5	330h
Llama 3 8B	8 Billion	7	280h

Table 6: Training times and parameter numbers for the four models chosen.

Unsurprisingly, because of its size, Mixtral 8x7B took considerably longer to fine-tune versus its 7-8 billion parameter counterparts. Training time was a determining factor in selecting the model chosen to perform emergency triage.

Appendix F. Prompts for Triage and Transformation

Two prompts were used: one for the triage task and one for the transformation task. We received guidance from experts in the ED. The triage prompt varied depending on whether it was used for fine-tuning or inference, as it included or excluded the triage scores from the training/test partition. On the other hand, the transformation prompt remained the same due to the few-shot nature of the task. The samples provided to the model for the few-shot task were intentionally diverse: some included references to sex, orthographic errors, or misplaced capitalization, while others did not. The final version of the prompt used for the few-shot task was the result of empirical adjustments until satisfactory performance in the transformation task was achieved.

Original triage prompt for fine-tuning in French:

`<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 #####`
{sample[“text”]}
`##### Fin de l'anamnèse #####`
Score de triage: `[/INST]` {sample[“triage_score”]}`</s>`

Original triage prompt for inference in French:

`<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 #####`
{sample[“text”]}
`##### Fin de l'anamnèse #####`
Score de triage: `[/INST]</s>`

English translation:

`<s>[INST]` You are a system that helps triage patients in the emergency department of a hospital. Given the medical history (which contains various kinds of information about 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.
`##### Start of medical history #####`
{sample[“text”]}
`##### End of medical history #####`
Triage score: `[/INST] </s>`

Original transformation prompt in French:

`<s> [INST]` Je vais vous donner l'anamnèse d'un(e) patient(e) en français. Le ou la patient(e) est soit un homme, soit une femme. Votre tâche consiste à remplacer toutes les références au sexe du ou de la patient(e), telles que les mots sexués comme "patient/patiente", "orienté/orientée" ou "fatigué/fatiguée", par le sexe opposé. Vous ne devez pas renvoyer que l'anamnèse modifiée. N'utilisez pas l'écriture neutre comme "traité(e)". S'il n'y a pas des références au sexe a changer, ne changez rien. Ne corrigez pas les fautes d'orthographe ou les majuscules mal placées tels que (EPileptique). Les anamnèses changées doivent être dans un format csv (sexe, anamnèse, antécédents).

[Examples 1-7 omitted for brevity]

`##### Anamnèse à changer #####`
Anamnèse originale :
Sexe patient : {data_point["sex"]}, Anamnèse : {data_point["anam_ioa"]},
Antécédents : {data_point["antecedents"]}

Anamnèse changée : `[/INST] </s>`

English translation:

`<s> [INST]` I will provide you with the medical history of a patient in French. The patient is either a man or a woman. Your task is to replace all references to the patient's sex, such as gendered words like "patient/patiente," "orienté/orientée," or "fatigué/fatiguée," with the opposite sex. You should only return the modified medical history. Do not use neutral forms like "traité(e)." If there are no gendered references to change, make no changes. Do not correct spelling or capitalization errors, such as (EPileptique). The modified medical histories should be returned in CSV format (sex, medical history, medical background).

[Examples 1-7 omitted for brevity]

`##### Medical history to change #####`
Original medical history:
Patient gender: {data_point["sex"]}, Medical history: {data_point["anam_ioa"]},
Medical background: {data_point["antecedents"]}

Changed medical history: `[/INST] </s>`

Appendix G. Additional results

When expanding to show the difference per triage score, as it is shown in Figure 9 we observe the same differences being aggravated by certain triage scores, as is the case for triage scores 2 and 4. Differences in triage scores closer to 1 are arguably more serious, as this score is reserved for patients who are at an increased risk of dying. Given the large sample size, confidence intervals in this and all subsequent figures are approximated using the normal distribution.

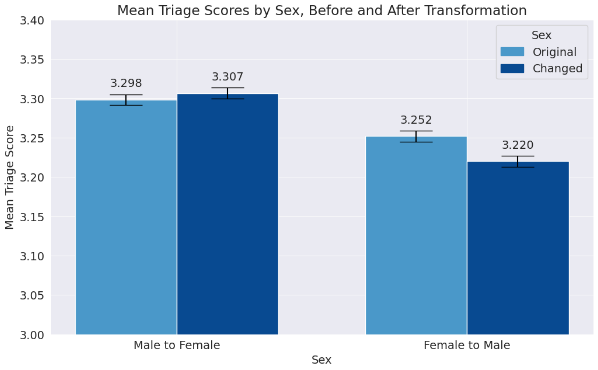

Figure 8: Mean triage scores before and after artificially transforming references to the sex of patients. 95%CI are based on normal approximation.

G.1. Stratification of results

Authors in the literature describe that some factors–not necessarily clinical–have been found to aggravate over and under triage in patients under the same conditions. To this end, we stratified our results according to some of the factors described in the literature as possibly aggravating triage score differences between male and female patients. The first of these, the interaction between the sex of the triage nurse and the sex of a patient is illustrated in Figure 10 which shows the mean difference of triage score by sex of triage nurse and sex of patient, the previously-described effect of under-triage observed in female patients is aggravated when these are admitted by a female triage nurse. Male patients, on the contrary, appear to be triaged equally regardless of the sex of the triage nurse.

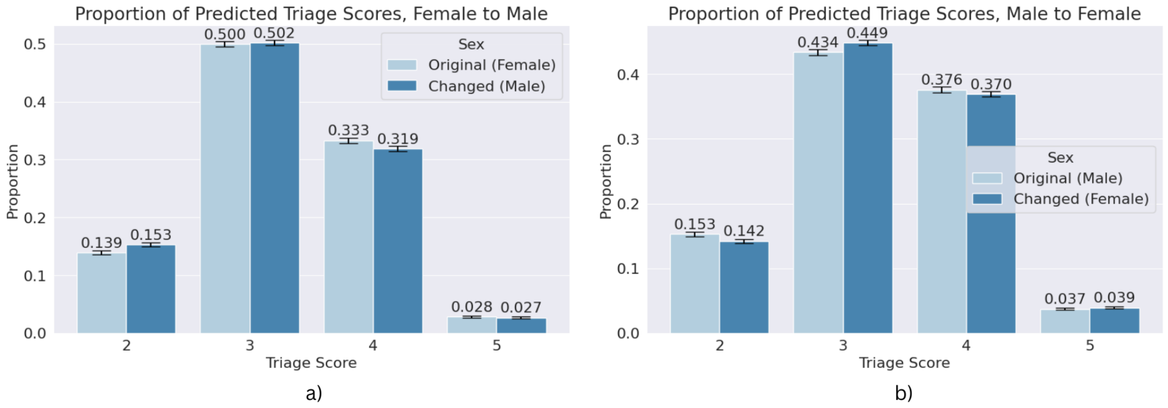

Figure 9: Triage score proportions per triage score per sex for a) Female to Male, b) Male to Female. 95%CI are based on normal approximation.

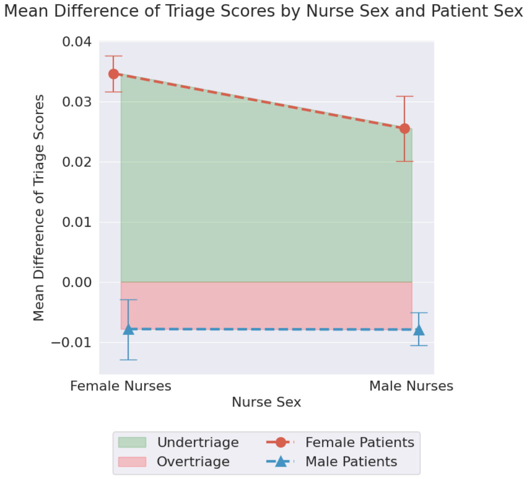

Figure 10: Mean difference of triage score after transformation, according to the sex of the triage nurse. 95%CI are based on normal approximation.

We also explored the possible relation between the expression of pain in patients and their triage score[6]. As illustrated on Figure 11, we observed an aggravating effect of under-triage for Female patients which correlated to the expressed pain score.

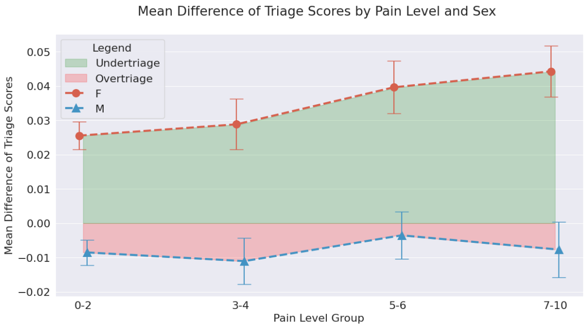

Figure 11: Mean difference of triage score after transformation, according to expressed level of pain. 95%CI are based on normal approximation.

When observing the mean differences with respect to the age of patients as shown on Figure 12, we observed two main effects. First, it appears that male patients tend to get over-triaged with an advanced age, while female patients appear do be less undertriaged under the same conditions. This is consistent with remarks made by triage nurses in French hospitals as described in Avalos et al. (2024), where they affirmed to sometimes ‘prioritize’ patients with an advanced age even if their current clinical condition does not necessarily need it. It is worth considering however that a patient’s age, while not always an explicit triage criteria, can carry clinical relevance in certain cases. Hence, it is difficult to completely distinguish cognitive biases from relevant clinical decisions when dealing with a patient’s age.

[6] Pain scores are self-reported by patients, and go between 0 and 10

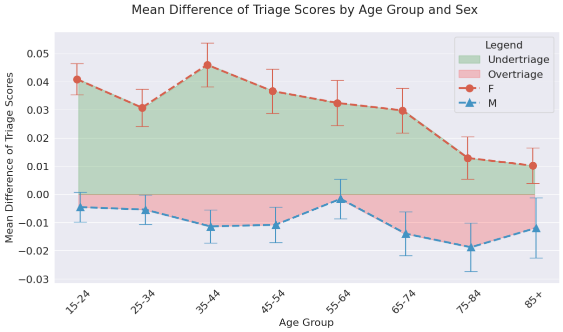

Figure 12: Mean difference of triage score after transformation, according to age. 95%CI are based on normal approximation.

The experience of triage nurses was also identified as a key factor influencing triage in the qualitative study described in Avalos et al. (2024). We explored the relationship between nurses' experience in triage (measured in years of experience as a triage nurse at the time of admission) and the impact of sex transformation on triage (Figure 13), with a focus on the most urgent triage score, originally a 2. While more experienced nurses exhibited similar variability in assigning triage scores after the sex transformation (but maintaining the same other clinical and external conditions), this variability was independent of the patient's sex. In contrast, less experienced nurses demonstrated a greater influence of the patient's sex when under-triaging after transformation: female patients were more frequently under-triaged.

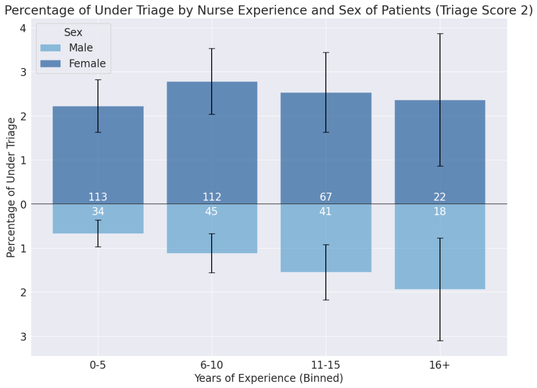

Figure 13: Percentage of under-triage per nurse experience for patients that should have been triaged as 2. 95%CI are based on normal approximation.

Finally, to test our hypothesis, we replicated the same workflow described earlier with the already transformed data, re-transforming it to the original sex of the patients. If these biases are present in the triage patterns of human nurses and are accurately replicated by the fine-tuned model, we should observe the same effect when analyzing mean triage scores per patient sex. The results of this analysis, shown in Figure 14, confirm our suspicions, although they reveal a small difference between the originally predicted triage scores and their re-transformed counterparts. This may indicate the influence of other factors we have not considered, such as potential biases inherent in the language model itself or clinical reasons that should be analyzed on a case-by-case basis.

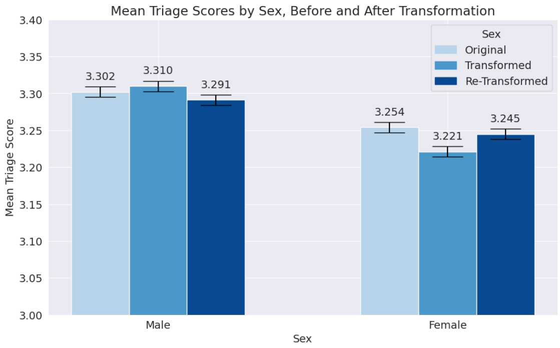

Figure 14: Mean difference of triage score per triage nurse experience. 95%CI are based on normal approximation