

Enhancing Patient Safety Through Machine Learning: Detecting and Prioritizing Risks in Clinical Narratives

by

Hongbo Chen

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Mechanical and Industrial Engineering
University of Toronto

© Copyright by Hongbo Chen 2026

Enhancing Patient Safety Through Machine Learning: Detecting and Prioritizing Risks in Clinical Narratives

Hongbo Chen

Doctor of Philosophy

Graduate Department of Mechanical and Industrial Engineering
University of Toronto

2026

Abstract

Patient safety remains a central priority in healthcare systems, as adverse events continue to contribute to preventable harm, mortality, and substantial resource burdens. Incident reporting systems capture large volumes of patient safety events (PSEs) containing critical signals of risk; however, manual review processes for accurate classification and identification of high-severity cases are increasingly impractical. Emerging evidence further indicates that patient safety risks also arise from clinical documentation practices, particularly stigmatizing language that may adversely affect patient psychological well-being. Given the scale and complexity of healthcare narrative data, scalable analytical approaches are essential for strengthening institutional safety surveillance. This dissertation develops and evaluates machine learning and natural language processing approaches to detect, prioritize, and analyze patient safety risks embedded within healthcare narratives.

First, accurate classification of PSEs by event type is essential to ensure that incident reports are directed to the appropriate patient safety specialists for timely review. To support this objective, we investigated natural language processing methods for automated PSE classification.

By leveraging transformer-based contextual representations, the models demonstrated significantly improved performance relative to conventional static representations.

Second, recognizing that healthcare institutions must prioritize the investigation of high-severity PSEs under resource constraints, we developed ranking-based machine learning models to prioritize events by severity. Using over 100,000 incident reports from a large academic health system, transformer-based ranking models consistently outperformed conventional classification methods and reporter-assigned severity baselines.

Third, we examined data-efficient detection of stigmatizing language in electronic health records, an emerging source of non-physical patient harm that may influence clinician judgment and patient well-being. We compared in-context learning with fully supervised models for stigma detection under data-scarce conditions. The proposed framework achieved performance comparable to supervised fine-tuning while requiring substantially fewer annotated examples.

Collectively, this dissertation advances scalable and data-efficient machine learning approaches for analyzing healthcare narratives in support of institutional patient safety monitoring. This work demonstrates how machine learning methods can enhance patient safety by augmenting expert review processes and strengthening safety surveillance in complex healthcare systems.

Acknowledgments

I would like to express my earnest gratitude to my supervisor, Professor Eldan Cohen, for his guidance, mentorship, and unwavering support throughout my doctoral studies. His intellectual rigor, thoughtful feedback, and commitment to high standards have profoundly shaped my development as a researcher. I am sincerely grateful for the trust he placed in me and for the many opportunities he provided to grow both academically and professionally.

I would also like to thank my committee members, Professor Sinisa Colic, Professor Michael W. Carter for their time, insight, and constructive feedback. Their thoughtful questions and critical perspectives not only strengthened this work but also helped guide my research in the right direction at pivotal stages of my doctoral training.

I am especially grateful to my former committee member, Professor Mark Chignell, whose encouragement and support during my doctoral journey meant a great deal to me. His guidance and kindness left a lasting impression, and I am deeply thankful for the time he dedicated to mentoring and supporting me. His memory remains an inspiration.

To my labmates at the Optimization and Machine Learning Lab and colleagues at University Health Network, thank you for the stimulating discussions and collaborations that made this journey both intellectually rewarding and personally meaningful. The shared challenges and successes along the way are memories I will always value.

I would also like to express my sincere appreciation to my friends, Songhao Xu whose friendship and steady encouragement have been a constant source of strength throughout this journey. During the more challenging periods of my doctoral studies, their support and belief in me provided reassurance and perspective that I deeply value.

Finally, I owe my heartfelt gratitude to my parents. Their strong belief in me, constant encouragement, and sacrifices have made this achievement possible. Their support has been the foundation upon which this journey was built.

This dissertation would not have been possible without the guidance, encouragement, and support of all those acknowledged above.

Table of Contents

Acknowledgments.....	iv
Table of Contents	v
List of Tables	xi
List of Figures	xiii
Chapter 1 Introduction	1
1 Introduction	1
1.1 Thesis Statement	3
1.2 Problem Statement	3
1.3 Research Objectives	4
1.4 Summary of Contributions.....	5
1.4.1 Advancing Reliable and Explainable AI for Patient Safety Event Classification (Chapter 3)	5
1.4.2 Enhancing Institutional Severity Triage Through Ranking-Based Machine Learning (Chapter 4).....	5
1.4.3 Data-Efficient Detection of Stigmatizing Language via In-Context Learning (Chapter 5)	6
1.5 Organization of the Dissertation	7
Chapter 2 Background	8
2 Background	8
2.1 Patient Safety Monitoring	8
2.1.1 Patient Safety Events	8
2.1.2 Current Triage and Review Workflows	10
2.2 Machine Learning and Natural Language Processing	11
2.2.1 Text Representation Methods	11
2.2.2 Natural Language Processing Learning Tasks.....	13
2.3 Machine Learning for Patient Safety Event Classification	16

2.4 Machine Learning for Triageing High-severity Patient Safety Events	17
2.5 Machine Learning for Detecting Stigmatizing Language in Clinical Documentation	18
Chapter 3 Advancing Reliable and Explainable AI for Patient Safety Event Classification.....	20
3 A Machine Learning Approach with Human-AI Collaboration for Automated Classification of Patient Safety Event Reports: Algorithm Development and Validation Study	20
3.1 Abstract.....	20
3.2 Introduction.....	22
3.2.1 Study Objectives	24
3.3 Methods.....	24
3.3.1 Data Collection	24
3.3.2 Data Preprocessing.....	25
3.3.3 Classifier Training	25
3.3.4 Classifier Evaluation	26
3.3.5 Development and Assessment of Explainability	26
3.3.6 Interface Development.....	27
3.3.7 Ethical Considerations	28
3.4 Results.....	28
3.4.1 Performance Comparison.....	28
3.4.2 Performance on Classifying Individual Event Types	29
3.4.3 LIME-Based Explainability Analysis	31
3.4.4 PSE Reporting System Interface.....	33
3.5 Discussion.....	35
3.5.1 Overview.....	35
3.5.2 Principal Findings	35
3.5.3 Comparison with Previous Work.....	39
3.5.4 Limitations	40

3.6 Conclusions	41
Chapter 4 Enhancing Institutional Severity Triage Through Ranking-Based Machine Learning	42
4 Machine Learning to Prioritize High-Severity Patient Safety Events for Institutional Investigation: Algorithm Development and Validation Study.....	42
4.1 Abstract	42
4.2 Introduction	44
4.2.1 Study Objectives	46
4.3 Methodology	46
4.3.1 Study Design	46
4.3.2 Dataset Description	47
4.3.3 Model Development and Validation	49
4.3.4 Model Evaluation and Comparison	53
4.3.5 Interface Design	54
4.3.6 Ethical Considerations	55
4.4 Results	55
4.4.1 Performance Comparison of Classification and Ranking Models	55
4.4.2 Comparison with the Reporter-Assigned Severity Baseline.....	58
4.5 Discussion	60
4.5.1 Overview	60
4.5.2 Principal Findings	61
4.5.3 Limitations	62
4.6 Conclusion	63
Chapter 5 Data-Efficient Detection of Stigmatizing Language via In-Context Learning	64
5 Efficient Detection of Stigmatizing Language in Electronic Health Records via In-Context Learning: Comparative Analysis and Validation Study.....	64
5.1 Abstract	64

5.2	Introduction	66
5.2.1	Study Objectives	71
5.3	Methods	71
5.3.1	Dataset Description	71
5.3.2	Comparative Analysis for Stigmatizing Language Detection	73
5.3.3	Textual Entailment Approach	74
5.3.4	SetFit Approach	75
5.3.5	Fully Supervised Fine-tuning Approach	76
5.3.6	In-Context Learning Approach	77
5.3.7	Prompting Strategies for the ICL Approach	78
5.3.8	Design of Stigma Detection Guided Prompt	83
5.3.9	Performance Evaluation and Comparison	84
5.3.10	Fairness Evaluation	85
5.3.11	Ethical Considerations	87
5.4	Results	88
5.4.1	Overview	88
5.4.2	Performance Evaluation for Zero-shot Approaches	88
5.4.3	Performance Evaluation for Few-shot Approaches	91
5.4.4	Performance Comparison of the ICL with the Fully Supervised Fine-tuning Approach	96
5.4.5	Statistical Comparison of Model Performance	97
5.4.6	Fairness Evaluation and Comparison	98
5.5	Discussion	104
5.5.1	Overview	104
5.5.2	Principal Findings	105
5.5.3	Clinical Adoption and Privacy Considerations	108

5.5.4	Comparison with Prior Work	111
5.5.5	Limitations	113
5.6	Conclusions	115
Chapter 6	Conclusions and Future Directions	116
6	Conclusions and Future Directions	116
6.1	Overview and Synthesis of the Thesis	116
6.2	Summary of Contributions.....	117
6.2.1	Advancing Reliable and Explainable AI for Patient Safety Event Classification.....	117
6.2.2	Enhancing Institutional Severity Triage Through Ranking-Based Machine Learning	118
6.2.3	Data-Efficient Detection of Stigmatizing Language via In-Context Learning	119
6.3	Limitations and Future Research	120
6.3.1	Data Quality, Subjectivity, and Generalizability	120
6.3.2	Expanding the Detection of Non-Physical Patient Harm	121
6.3.3	Prospective Evaluation, Deployment, and Continuous Adaptation	122
6.4	Concluding Remarks.....	124
References or Bibliography		126
Appendices.....		149
Appendix 3.1	Data Preprocessing Procedures.	149
Appendix 3.2	Evaluation Metrics for Examining Machine Learning Classifiers.	151
Appendix 3.3	The Performance of ML Classifiers in Classifying PSE Report Event Type. .	153
Appendix 4.1	Feature Extraction Procedures for Fixed-Representation-Based Models.	156
Appendix 4.2	Hyperparameter Settings, Tuning Ranges, and Final Selected Configurations for Feature-Based and Transformer-Based Models.	158
Appendix 4.3	Mathematical Definitions of Evaluation Metrics.	172
Appendix 4.4	Demonstration of the Designed Severity Ranking Interface.	175

Appendix 4.5 Evaluation Results of All Developed Ranking and Classification Models.	177
Appendix 5.1 Dataset Description and Stigmatizing Language Examples.	191
Appendix 5.2 Detailed Code Implementation for the Prompting Strategies.	193
Appendix 5.3 Paired t-test Comparison of Top-performing Models in Zero-shot, Few-shot, and Supervised Fine-tuning Approaches.	195
Appendix 5.4 Performance Metrics of the Few-shot In-Context Learning Models.	197
Appendix 5.5 Performance and Performance Disparities of Models across Sex, Age, and Race Demographic Attributes.	203

»

List of Tables

Table 3-1 Prevalence of patient safety event reports by event type.	25
Table 3-2 Performance of top-performing ML classifiers trained with static and contextual text representations.	29
Table 3-3 Performance of SVM+RoBERTa-base on the individual event type.....	29
Table 3-4 The five most prevalent important words for each event type that were derived from the SVM classifier trained with RoBERTa-base.	32
Table 3-5 Performance of the top-performing ML classifier on reports that have relevant words highlighted and reports with irrelevant words highlighted.	33
Table 4-1 Distribution of PSE reports by severity labels.	48
Table 4-2 Summary of models used in this study, including their abbreviations and corresponding full model names. Models with the suffix “RA” were developed for ranking tasks, while those with the suffix “CL” were developed for classification tasks.	50
Table 4-3 . Comparative performance of the best-performing classification model (BIOBERT-CL) and ranking model (BERT-RA) across two evaluation settings.	55
Table 4-4 Comparative performance of the best-performing ranking model (BERT-RA) and the Reporter-Assigned Severity Baseline across two evaluation settings.	58
Table 5-1 Demographic information of the stigmatizing dataset used in this study.	72
Table 5-2 Pretrained language models used in the textual entailment approach.	74
Table 5-3 Pretrained sentence transformers used in the SetFit approach.	75
Table 5-4 Pretrained language models used in the fully supervised fine-tuning approach.	77
Table 5-5 General-purpose LLMs used in the ICL approach.	78
Table 5-6 Prompting strategies and templates used in the ICL approach.....	82

Table 5-7 Performance of pretrained language models utilizing the textual entailment approach.	88
Table 5-8 Performance of general-purpose LLMs utilizing the zero-shot ICL approach.	89
Table 5-9 Performance of pretrained sentence transformers utilizing the SetFit approach.....	91
Table 5-10 Performance of general-purpose LLMs utilizing the few-shot ICL approach.	93
Table 5-11 Comparison of top-performing models in SetFit and few-shot ICL approaches.	95
Table 5-12 Performance of pretrained language models for the fully supervised fine-tuning approach.	96
Table 5-13 Performance of top-performing models in zero-shot, few-shot, and fully supervised fine-tuning approaches.	97
Table 5-14 Paired t-test results comparing the top-performing models in the zero-shot and few- shot settings, and between the best-performing models in ICL and supervised fine-tuning approaches.....	97
Table 5-15 The largest absolute performance disparity across demographic attributes for different approaches in detecting stigmatizing language.	99
Table 5-16 The highest observed performance disparities of ICL and supervised fine-tuning models across sex, age, and race subgroups.	104

List of Figures

Figure 2-1. Example incident report illustrating the documentation of a patient safety event, including structured reporting fields and a free-text narrative description of the event.	9
Figure 2-2. Overview of the patient safety event management workflow, including assignment by event type, specialist triage, and multidisciplinary review of serious events.	10
Figure 3-1 Confusion matrix for the testing set evaluation with the SVM classifier trained with RoBERTa-base.	31
Figure 3-2 LIME explanations of the SVM classifier trained with RoBERTa-base.	32
Figure 3-3 Interface visualization of a patient safety event report classifier coupled with the local interpretable model-agnostic explanations technique.	34
Figure 4-1 Overview of the study workflow, including data processing, model development, and interface design.	47
Figure 4-2 Model development workflow for fixed-representation-based and transformer-based models. Numbers indicate the proportion of the dataset allocated to each stage of model training, validation, and testing.	52
Figure 5-1 Workflow for developing a supervised machine learning model in stigmatizing language detection.	70
Figure 5-2 Examples of four distinct prompting strategies utilized in this study.	81

Chapter 1

Introduction

1 Introduction

Patient safety remains a central priority in modern healthcare systems. Despite decades of reform following the publication of *To Err Is Human*, preventable adverse events continue to impose substantial clinical and economic burdens [1]. Adverse events are preventable incidents that cause or have the potential to cause harm to patients during healthcare delivery [2,3]. In the United States, adverse events have been estimated to contribute to more than 250,000 deaths annually, ranking among the leading causes of mortality, behind heart disease and cancer [4]. In Canada, recent national data indicate that adverse events occur in approximately 6% of hospitalizations, corresponding to about one in every 17 hospital stays, and that multiple harmful events occur in roughly one-quarter of these cases [5]. Beyond their clinical consequences, adverse events generate profound economic costs. Patient harm has been estimated to reduce global economic growth by approximately 0.7% annually, with indirect costs of unsafe care reaching approximately USD 118 trillion worldwide between 2000 and 2020 [6].

Patient safety event (PSE) reporting systems, also referred to as incident reporting systems, have been widely implemented across healthcare institutions worldwide as a foundational mechanism for reducing adverse events and strengthening patient safety oversight [7,8]. These systems allow frontline healthcare professionals to document adverse events, and unsafe conditions using structured fields and free-text narrative descriptions [9]. Several jurisdictions have established legislation or national frameworks governing aspects of patient safety incident reporting [112–114].

The widespread adoption of these systems has led to the accumulation of large repositories of safety-related data. For example, approximately 2.3 million patient safety events were reported to the National Reporting and Learning System in England between April 2021 and March 2022 alone [10]. These reports are intended to facilitate root cause analysis, inform quality improvement initiatives, and guide preventive strategies [7,11]. However, the sheer volume and complexity of submitted reports present substantial operational challenges. Because comprehensive investigation is resource-intensive, and only a limited proportion of reported events undergo in-depth review.

As a result, critical safety signals may remain embedded within unstructured narrative documentation, constraining the full realization of institutional learning and harm prevention objectives [12,13].

While patient safety has traditionally emphasized the prevention of physical injury arising from technical, systemic, or operational failures, the scope of harm in healthcare extends beyond bodily damage [14,15]. There is growing recognition that disrespectful or stigmatizing conduct toward patients can give rise to dignitary harms, a form of preventable, non-physical harm that undermines patients' sense of moral worth and may adversely affect their psychological well-being [16–18]. When patients are treated as unreliable, noncompliant, or morally inferior, the resulting harm may not manifest as immediate physical injury but can erode therapeutic relationships and compromise care engagement [19–21].

Stigmatizing language embedded within clinical documentation constitutes a particularly consequential manifestation of dignitary harm. In healthcare contexts, stigmatizing language refers to terminology or descriptions that carry negative connotations, portray patients in a pejorative manner, assign blame, or convey judgment based on social identity, medical condition, or personal circumstances [22]. Unlike transient verbal interactions, written documentation persists within electronic health records and can influence how subsequent clinicians interpret patients' credibility, behavior, and symptoms [23,24]. As such, stigmatizing language within clinical notes may shape clinical reasoning, affect treatment decisions, and contribute to inequities in care delivery [23,24,36]. Importantly, such language is often subtle and highly context-dependent rather than overtly derogatory [25,26]. The implicit nature of these expressions limits the effectiveness of simple keyword-based detection methods and necessitates nuanced analytical approaches, thereby posing substantial challenges for large-scale monitoring of documentation quality within electronic health records [25,26].

Healthcare narratives have become central to institutional patient safety efforts, functioning both as repositories of critical safety signals and as potential mechanisms through which harm may be perpetuated. Their scale, contextual nuance, and unstructured nature challenge the sustainability of manual review and highlight the need for systematic analytical approaches.

1.1 Thesis Statement

The central thesis of this dissertation is that advanced natural language processing and machine learning methods can enable scalable and systematic analysis of healthcare narratives to strengthen institutional patient safety monitoring. Specifically, this work demonstrates that contextual text representations, ranking-based learning frameworks, and in-context learning approaches can enable accurate classification of patient safety events, effective prioritization of high-severity incidents, and data-efficient detection of stigmatizing language in clinical documentation.

1.2 Problem Statement

Healthcare institutions increasingly rely on narrative data, such as patient safety event (PSE) reports and clinical documentation, to monitor risks and improve patient safety [2,26–28]. However, the volume and complexity of these narratives have grown substantially, making comprehensive manual review processes difficult to sustain at scale [10,29]. Although expert review remains essential for classifying PSEs by event type, prioritizing high-severity cases, and monitoring documentation quality, these processes are constrained by time, staffing, and variability in interpretation.

Recent advances in machine learning and natural language processing provide new opportunities to analyze healthcare narratives more efficiently. Yet important gaps remain. First, much of the existing literature on PSE classification relies on static text representations that may inadequately capture contextual nuances in clinical narratives [2,13,30,31]. Although transformer-based language models enable contextual text representations [32,33], their application in PSE classification remains limited, indicating an opportunity to examine whether such models can improve classification performance in patient safety reporting data. Second, current approaches to severity modeling largely emphasize categorical classification [30,34], whereas institutional safety workflows require effective mechanisms for prioritizing high-severity events under constrained resources [34,35]. This misalignment suggests the need to explore alternative modeling formulations that directly account for relative event severity and prioritization. Third, existing methods for detecting stigmatizing language predominantly rely on supervised learning approaches that depend on large annotated datasets [25,36,37]. Developing such datasets is costly and time-consuming, as accurate annotation requires both domain expertise to interpret clinical context and linguistic proficiency to identify subtle expressions of stigma [19,36]. These

constraints highlight the need to investigate data-efficient approaches capable of detecting stigmatizing language under limited annotation availability.

Taken together, these limitations highlight the need for machine learning approaches capable of more effectively analyzing healthcare narratives to support accurate event classification, prioritization of high-severity incidents, and efficient detection of stigmatizing language. Addressing these gaps is essential for enabling scalable analysis of healthcare narratives and strengthening patient safety monitoring in modern healthcare systems.

1.3 Research Objectives

The overarching objective of this dissertation is to develop machine learning approaches for scalable analysis of healthcare narratives to strengthen institutional patient safety monitoring.

To achieve this aim, the dissertation pursues the following research objectives:

- Objective 1: Examine transformer-based contextual text representations for patient safety event classification using incident report narratives. This objective compares the efficacy of contextual text representations with traditional static text representations for patient safety event classification and examines the use of explainability techniques to enhance transparency and support human–AI collaboration within institutional patient safety review workflows.
- Objective 2: Assess ranking-based machine learning frameworks for prioritizing high-severity patient safety events using large-scale incident report data. This objective models severity as a relative ordering problem rather than a categorical classification task. It evaluates whether ranking formulations better align with institutional triage priorities compared with conventional classification models and reporter-assigned severity ratings.
- Objective 3: Develop and evaluate data-efficient approaches for detecting stigmatizing language in electronic health records through in-context learning. This objective investigates whether large language models, guided by domain-informed prompting strategies, can achieve performance comparable to fully supervised fine-tuning under zero-shot and few-shot conditions, thereby reducing reliance on extensive annotated datasets.

1.4 Summary of Contributions

This dissertation contributes to institutional patient safety monitoring by developing machine learning approaches that enable scalable analysis of healthcare narratives. The primary contributions are as follows:

1.4.1 Advancing Reliable and Explainable AI for Patient Safety Event Classification (Chapter 3)

This chapter advances the methodological and translational foundations of AI-assisted patient safety event (PSE) classification by demonstrating how contextualized language modeling and explainable AI can be integrated into institutional safety analysis workflows. Through a rigorous benchmarking study, it provides empirical evidence that transformer-derived contextual representations substantially outperform traditional static representations in PSE classification, yielding meaningful improvements in both accuracy and F1-score. Beyond predictive performance, this work extends conventional model evaluation by explicitly examining the relationship between explanation quality and classification reliability. By applying LIME to analyze model reasoning and linking explanation relevance to downstream accuracy, the chapter introduces an explanation-informed perspective in safety-critical applications. The observed disparity in performance between cases with relevant versus irrelevant highlighted terms underscores the importance of transparent reasoning and structured human oversight when deploying AI systems in clinical environments. Finally, this chapter translates methodological advances into a workflow-aligned human–AI collaboration framework. It presents a prototype reporting interface that integrates probabilistic recommendations and real-time explanation visualization, enabling reporters and patient safety specialists to retain decision authority while benefiting from algorithmic support. By operationalizing level-2 automation within incident reporting systems, the work bridges the gap between algorithm development and practical implementation. Collectively, this chapter establishes a foundation for trustworthy, workflow-integrated AI systems capable of strengthening PSE classification.

1.4.2 Enhancing Institutional Severity Triage Through Ranking-Based Machine Learning (Chapter 4)

This chapter enhances institutional patient safety monitoring by reframing high-severity event triage as a ranking problem rather than a conventional classification task. While prior research has

focused on predicting categorical severity labels, real-world triage requires ranking events by severity so that the most serious cases can be systematically prioritized for comprehensive investigation. By introducing and validating learning-to-rank models for prioritizing serious safety events, this work aligns machine learning objectives directly with operational decision-making needs. Using over 100,000 patient safety event reports from a large academic health system, the study systematically compares ranking-based and classification-based approaches across expert-adjudicated and deployment-representative datasets. The ranking models consistently outperformed classification models across multiple evaluation metrics, with performance gains widening at broader review cutoffs. Beyond methodological innovation, this chapter benchmarks model performance against the Reporter-Assigned Severity Baseline, reflecting current institutional practice. The substantial improvement in retrieving high-severity events exposes the limitations of relying solely on subjective reporter-assigned severity. As a whole, this work establishes ranking-based machine learning as a reliable, workflow-aligned approach for strengthening institutional severity triage.

1.4.3 Data-Efficient Detection of Stigmatizing Language via In-Context Learning (Chapter 5)

This chapter establishes in-context learning (ICL) as a viable and data-efficient framework for detecting stigmatizing language in electronic health records. In contrast to approaches that rely on large annotated datasets for supervised fine-tuning, this work demonstrates that strategically prompted general-purpose large language models can achieve competitive performance in zero-shot and few-shot settings. Although supervised fine-tuning achieved the highest overall performance, ICL attained closely comparable F1-scores with substantially fewer labeled examples, highlighting its scalability under annotation constraints. Beyond benchmarking against textual entailment, SetFit, and fine-tuned transformer models, this chapter introduces a stigma detection-guided prompting strategy that embeds domain-specific linguistic characteristics into the model's reasoning process. The findings demonstrate that prompt configuration meaningfully shapes detection accuracy and that integrating clinically grounded linguistic guidance improves sensitivity to subtle and context-dependent expressions of stigmatization. Collectively, this work advances a scalable and annotation-efficient paradigm for automated stigma monitoring in clinical documentation systems.

1.5 Organization of the Dissertation

This dissertation is organized into six chapters.

- Chapter 1 introduces the background, thesis statement, problem statement, objectives, and key contributions.
- Chapter 2 provides background by describing existing institutional patient safety workflows, outlining relevant machine learning and natural language processing methodologies, and reviewing prior research on machine learning applications for patient safety event analysis and the detection of stigmatizing language.
- Chapter 3 presents machine learning models for patient safety event classification using contextual text representations. Model performance is systematically benchmarked against approaches based on static text representations, and a prototype interface incorporating an explainable AI technique is developed to illustrate how such models can support human–AI collaboration within institutional incident reporting workflows.
- Chapter 4 evaluates ranking-based machine learning models for prioritizing high-severity patient safety events using large-scale incident report data. The chapter analyzes the extent to which ranking formulations better reflect operational triage priorities compared with categorical classification approaches. Performance is systematically benchmarked against classification-based models and reporter-assigned severity baselines using established ranking metrics.
- Chapter 5 examines automated detection of stigmatizing language in electronic health records, comparing in-context learning and supervised fine-tuning approaches under data-scarce conditions. The chapter further investigates the use of domain-informed prompting strategies to guide large language models, evaluating whether such strategies improve stigma detection performance relative to standard prompting approaches.
- Chapter 6 synthesizes the dissertation findings, discusses methodological, practical, and ethical implications for institutional patient safety monitoring, and outlines limitations and directions for future research.

Chapter 2 Background

2 Background

This chapter provides the background for the research presented in this dissertation. It first introduces patient safety monitoring, including patient safety events, how they are generated and documented, and current workflows for classification, triage, and review. It also discusses emerging concerns related to stigmatizing language in clinical documentation. The chapter then reviews relevant machine learning and natural language processing methods for analyzing textual data, including text representation approaches, transformer-based language models, classification frameworks, and learning-to-rank methods. Finally, prior research on machine learning for patient safety event analysis and stigmatizing language detection is summarized to contextualize the methods developed in the subsequent chapters.

2.1 Patient Safety Monitoring

2.1.1 Patient Safety Events

Patient safety events (PSEs) refer to incidents occurring during the provision of healthcare that have the potential to result in patient harm. These events may include adverse events that cause injury, near misses that are detected before harm occurs, and unsafe conditions that increase the risk of future harm [3]. Although definitions vary slightly across health systems, PSEs are generally understood as deviations from expected clinical processes that threaten patient safety. Common types of PSEs include medication errors, surgical complications, healthcare-associated infections, and diagnostic errors [38,39].

PSEs are typically documented through incident reporting systems, in which clinicians voluntarily report safety-related incidents following their occurrence [9]. These reports typically include structured fields (i.e., event type, location, and severity level) as well as free-text narratives describing the circumstances of the event and the resulting patient outcomes (Figure 2-1) [13]. The narrative component often captures important information about contributing factors, sequence of events, and clinical decision-making processes that may not be reflected in structured fields alone.

PATIENT INCIDENT REPORT FORM	
Use this form to report accidents, injuries, medical situations, criminal activities, traffic incidents, or student behavior incidents. If possible, a report should be completed within 24 hours of the event.	
Date of Report: _____, 20____	
PERSON INVOLVED	
Full Name: _____	Address: _____
Identification: ☐ Driver's License No. _____ ☐ Passport No. _____	
☐ Other: _____	
Phone: (____) ____-____	E-Mail: _____
THE INCIDENT	
Date of Incident: _____, 20____	Time: ____:____ ☐ AM ☐ PM
Location: _____	
Describe the Incident: _____	

INJURIES	
Was anyone injured? ☐ Yes ☐ No	
If yes, describe the injuries: _____	

WITNESSES	
Were there witnesses to the incident? ☐ Yes ☐ No	
If yes, enter the witnesses' names and contact info: _____	

Figure 2-1. Example incident report illustrating the documentation of a patient safety event, including structured reporting fields and a free-text narrative description of the event.

Traditionally, patient safety efforts have focused primarily on preventing physical injury caused by medical errors. However, contemporary patient safety frameworks increasingly recognize that harms may also arise from communication practices [16,18,40]. Emotional or dignitary harms, such as those resulting from disrespectful communication or stigmatizing language, can undermine trust in clinicians, influence clinical decision-making, delay patient engagement in care, and adversely affect patient experience and psychological well-being [18,23,24,41]. For instance, a study found that clinicians exposed to the stigmatizing term “substance abuser” as opposed to the more neutral phrase “person with substance use disorder” were more likely to view the patient as personally culpable and deserving of punitive action rather than treatment [42].

The potential impact of stigmatizing language is further amplified by its widespread presence in clinical documentation. A large survey study reported that stigmatizing language appears in a substantial proportion of clinical notes, with prevalence estimates ranging from 42.8% to 59.1% across history and physical notes, consultation notes, and discharge summaries [43].

2.1.2 Current Triage and Review Workflows

Following the submission of a patient safety event (PSE) report, the event typically enters a multi-stage institutional workflow involving assignment, triage, and review (Figure 2-2) [11,34,44,45]. In the first phase, the report is assigned to the appropriate patient safety specialist or operational team based on the event type [27,46]. For example, reports describing events such as incorrect medication dosing may be directed to patient safety specialists responsible for medication safety.

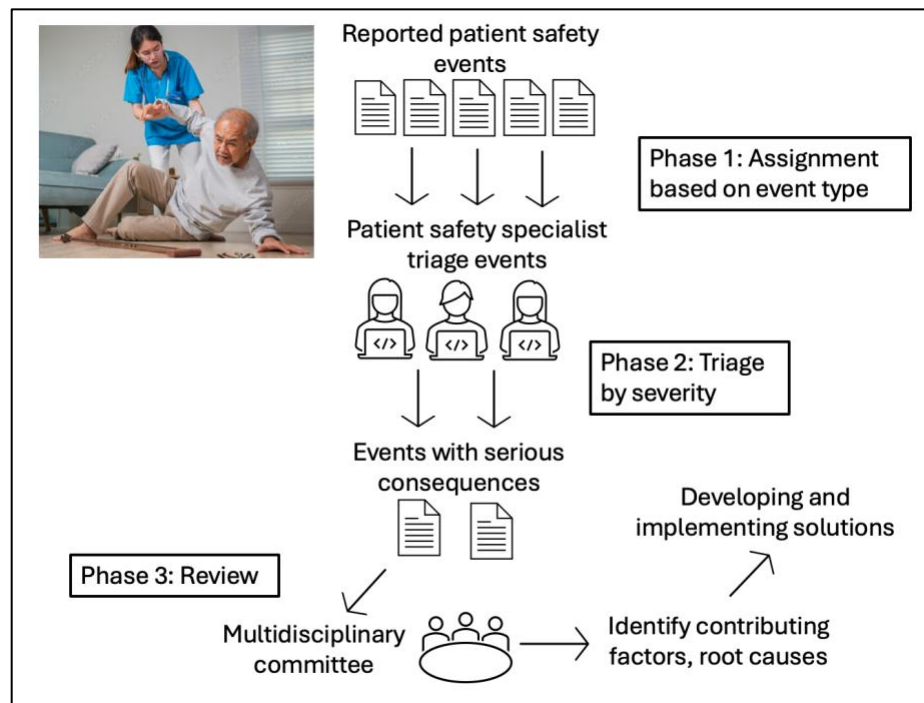


Figure 2-2. Overview of the patient safety event management workflow, including assignment by event type, specialist triage, and multidisciplinary review of serious events.

After this initial assignment step, the designated patient safety specialist reviews the report and performs triage based on the severity and urgency of the event [47,48]. At this stage, the specialist evaluates whether the reported incident may have serious consequences and whether escalation is warranted for more comprehensive institutional investigation. Because healthcare organizations

receive a large volume of incident reports [10], and this triage process is essential for prioritizing limited investigative resources toward the most consequential cases.

Higher-severity events are escalated to a comprehensive review process. In this phase, these serious events are examined by a multidisciplinary committee that may include point-of-care clinicians, patient safety specialists, clinical leaders, and other domain experts. The purpose of this review process is to systematically identify contributing factors, examine underlying system vulnerabilities, and support the development and implementation of corrective actions aimed at preventing recurrence [49]. Through multidisciplinary discussion, reviewers analyze the sequence of events, clinical context, and organizational conditions that may have contributed to the incident. This process often involves examining human factors, communication breakdowns, workflow design, and other latent system issues that may increase the likelihood of similar events occurring in the future [50,51]. In this way, the review process enables healthcare organizations to conduct structured investigations of serious safety events and develop targeted interventions to mitigate future patient harm.

2.2 Machine Learning and Natural Language Processing

2.2.1 Text Representation Methods

This dissertation uses machine learning methods to analyze healthcare narratives, including clinical documentation and patient safety event reports. Because machine learning algorithms operate on numerical inputs, raw text data must first be converted into a numerical representation before being used as input to machine learning models. This process is commonly referred to as text representation or feature extraction, which transforms unstructured text into structured vectors that capture meaningful linguistic information [52]. Broadly, text representation methods can be categorized into static representations and contextual representations, which differ in how they encode semantic relationships within text.

Static text representation methods generate fixed representations of words or documents without considering the surrounding context in which a word appears. One of the simplest approaches is term frequency, which represents a document by counting the number of occurrences of each word within the text. In this representation, each document is converted into a vector where each dimension corresponds to a word in the vocabulary, and the value represents how frequently the

word appears in the document [53]. A commonly used extension of this approach is term frequency-inverse document frequency (TF-IDF). TF-IDF improves upon simple frequency counts by weighting words according to their importance within the entire dataset. Specifically, TF measures how frequently a word appears in a document, while inverse document frequency (IDF) reduces the weight of words that appear frequently across many documents but carry limited discriminative information [54]. As a result, TF-IDF assigns higher weights to words that are more informative for distinguishing between documents. Another widely used static representation method is Global Vectors for Word Representation (GloVe), which learns dense vector representations for words by leveraging global word co-occurrence statistics from large text corpora [55]. GloVe learns a high-dimensional embedding space in which words appearing in similar contexts are located closer to each other in the vector space, enabling the representation to capture semantic relationships between words. To obtain document-level representations using GloVe, the embeddings of individual words within a document are typically aggregated, for example, by averaging the word vectors.

In contrast to static methods, contextual text representation methods generate word embeddings that depend on the context in which the word appears [56]. In these approaches, the representation of a word varies depending on the surrounding words in the sentence, allowing the same word to take on different meanings in different contexts. Contextual representations are commonly produced by transformer-based models, such as BERT and its variants [32,33], which have become the dominant architecture for many natural language processing tasks [57–59]. Transformer models rely on a self-attention mechanism that enables the model to learn relationships between words across an entire sequence [60], rather than processing text sequentially as in traditional recurrent neural networks. This mechanism allows the model to capture long-range dependencies and complex semantic relationships within text. These models are typically pretrained on large-scale corpora using self-supervised objectives, such as masked language modeling and next sentence prediction, enabling them to learn general linguistic structures and semantic patterns from raw text. The contextual representations produced by transformer-based models are able to capture more nuanced and informative language features compared with traditional static text representations.

The key distinction between static and contextual representations lies in their treatment of word meaning. Static representations assign a single fixed vector to each word regardless of context,

whereas contextual representations generate different embeddings for the same word depending on its usage in a particular sentence. This ability to incorporate contextual information allows transformer-based models to capture subtle linguistic patterns, which is particularly important for analyzing complex healthcare narratives in which meaning is often context-dependent.

2.2.2 Natural Language Processing Learning Tasks

Natural language processing (NLP) encompasses a range of learning tasks aimed at extracting meaningful information from textual data. These tasks vary depending on the analytical objective but commonly include text classification, information extraction, and learning-to-rank approaches [58,61–64]. In healthcare settings, NLP methods are increasingly used to analyze clinical narratives, electronic health records, and incident report descriptions to support decision-making and quality improvement initiatives [2,25,26,36,45,65].

Among these tasks, text classification and learning-to-rank approaches are particularly relevant for analyzing healthcare narratives. These two paradigms address different operational needs. Classification frameworks are typically used when the goal is to assign discrete labels to clinical text, such as identifying clinical conditions from narrative notes, classifying patient safety event reports by event type, or detecting stigmatizing language in clinical documentation [65–67]. In contrast, learning-to-rank methods are designed to rank items based on their relative importance and have been applied to prioritize high-risk patients or rank clinical alerts by urgency [68–70]. The sections below describe common text classification and learning-to-rank methods used in this dissertation.

2.2.2.1 Text Classification Frameworks

Text classification in natural language processing has evolved through several methodological paradigms that differ in how textual information is represented and how classification models are trained [58,61]. Early approaches primarily relied on static representations, such as term frequency or term frequency–inverse document frequency. These methods convert textual data into numerical vectors that can serve as inputs to conventional machine learning algorithms, including logistic regression, decision trees, and support vector machines [71–73]. However, the reliance on static feature representations limits their ability to capture contextual relationships and subtle semantic nuances present in complex clinical narratives.

Recent advances in deep learning have substantially improved the performance of text classification systems through the development of pretrained language models based on transformer architectures [32,60]. Models such as Bidirectional Encoder Representations from Transformers and related variants generate contextualized representations of text by modeling the relationships among words within their surrounding context [32,33,74]. These pretrained models can be adapted to downstream tasks through supervised fine-tuning on labeled datasets, allowing them to learn task-specific decision boundaries while retaining the linguistic knowledge acquired during pretraining. Supervised fine-tuning of transformer-based models is currently considered the state-of-the-art approach for many text classification tasks.

In addition to fully supervised approaches, alternative learning settings have been explored for scenarios where labeled data are limited. One such setting is zero-shot classification, in which models perform classification tasks without task-specific labeled examples [75,76]. A commonly used approach for zero-shot classification is the textual entailment framework, which formulates classification as a natural language inference problem [77]. In this framework, the input text is treated as a premise, and the candidate label is expressed as a hypothesis. The model then evaluates whether the premise entails the hypothesis, enabling classification decisions to be made by leveraging the linguistic knowledge acquired during pretraining. Another learning setting is few-shot learning, where models are trained using only a small number of labeled examples [61]. Methods such as the SetFit framework leverage pretrained sentence transformers to generate semantically meaningful sentence embeddings and subsequently train lightweight classification models using these representations [78]. By exploiting high-quality pretrained embeddings, such approaches can achieve competitive performance even when only a limited number of annotated examples are available.

More recently, in-context learning (ICL) has emerged as a data-efficient approach for performing text classification with general-purpose large language models [76,79,80]. Rather than updating model parameters through conventional training procedures, ICL enables models to perform classification by incorporating task instructions and a small number of illustrative examples directly within the input prompt. By leveraging the knowledge acquired during large-scale pretraining, this approach allows models to adapt to new tasks with minimal labeled data. Depending on the number of examples provided, ICL can operate in zero-shot or few-shot settings, making it particularly useful in scenarios where annotated datasets are limited [26,65].

The performance of text classification models is commonly assessed using several evaluation metrics that capture different aspects of predictive performance. Accuracy measures the proportion of correctly classified instances among all predictions and provides a general indication of model performance. However, accuracy alone may be insufficient when class distributions are imbalanced. To address this limitation, precision and recall are frequently used. Precision measures the proportion of predicted positive instances that are truly positive, reflecting the reliability of positive predictions, while recall quantifies the proportion of actual positive instances that are correctly identified by the model, reflecting the model's ability to detect relevant cases. To balance these two perspectives, the F1-score is often reported as the harmonic mean of precision and recall, providing a single measure that reflects both correctness and completeness of the predictions. For multiclass classification problems, these metrics are computed separately for each class and then aggregated across classes. Macro averaging assigns equal weight to each class, while weighted averaging accounts for class imbalance by weighting each class according to its number of instances.

2.2.2.2 Learning-to-Rank Methods

Learning-to-rank (LTR) methods are machine learning approaches designed to order items according to their relative relevance or importance rather than assigning discrete labels. Unlike classification models, which predict categorical outcomes for individual instances, ranking models learn to produce an ordered list in which the most relevant items appear near the top. This paradigm was originally developed for information retrieval systems, where the goal is to rank documents according to their relevance to a user query [64].

LTR methods are commonly categorized into three frameworks, including pointwise, pairwise, and listwise approaches. Pointwise methods treat ranking as a regression or classification problem by predicting a relevance score for each instance independently [81]. When implemented as a classification task, models estimate the probability that an instance belongs to a target class, and items are subsequently ranked according to their predicted probabilities or scores. Pairwise methods learn relative ordering by comparing pairs of items and predicting which item should be ranked higher [62]. By optimizing pairwise comparisons, the model learns relative preferences between instances, which can be aggregated to produce a final ranking. Listwise approaches extend this concept by considering the entire ranked list during training and directly optimizing ranking-

specific evaluation metrics, such as normalized discounted cumulative gain (NDCG) [82]. In doing so, the model learns to produce an ordering of instances that maximizes the overall quality of the ranked list.

Ranking models are typically evaluated using metrics that assess the quality of the ranked list rather than the accuracy of individual predictions. One commonly used metric is Average Precision, which summarizes how well relevant items are ranked near the top of the list by averaging precision values at the positions where relevant items are retrieved. Another widely used metric is NDCG, which evaluates ranking quality by considering both the relevance of retrieved items and their positions in the ranked list, assigning greater importance to items appearing near the top. In addition, Precision@k measures the proportion of relevant items among the top k positions of the ranked list, reflecting the model's ability to correctly identify high-priority items within a limited review set. Complementarily, Recall@k quantifies the proportion of all relevant items that appear within the top k positions, indicating how effectively the ranking retrieves important cases. Together, these metrics provide a comprehensive assessment of ranking performance by evaluating both the accuracy and the ordering of relevant items within the ranked list.

2.3 Machine Learning for Patient Safety Event Classification

Accurately classifying patient safety event (PSE) reports by event type is essential for directing reported incidents to the appropriate patient safety specialists for follow-up investigation and management. With the growing volume of PSE reports generated by incident reporting systems [10,29], researchers have explored the use of machine learning methods to improve the efficiency of this classification task [2,13,30,31,47,67,83,84]. In this formulation, the narrative description within the PSE report serves as the input, and the objective is to predict the corresponding event type.

Early studies primarily relied on traditional machine learning algorithms combined with static text representation methods, most commonly term frequency or term frequency–inverse document frequency (TF–IDF) features. For example, Wang et al. applied logistic regression and support vector machine (SVM) classifiers with term frequency and TF–IDF features to classify ten types of patient safety event reports, achieving an F1-score of up to 0.783 using a dataset of 2,860 reports [2]. Similarly, Fong et al. investigated the use of SVM, random forest, and logistic regression

models with TF-IDF representations to classify miscellaneous PSE reports, reporting an accuracy of 92.0% using a dataset of 70,051 reports [31]. Evans et al. investigated the use of SVM and naïve Bayes classifiers with TF-IDF representations to automatically classify event type in PSE reports, reporting an area under the receiver operating characteristic curve of 0.891 using a dataset of 31,333 reports [67].

Other studies have examined more narrowly defined classification tasks. For instance, Ong et al. evaluated the feasibility of automatically classifying two types of PSE reports using term-frequency representations with SVM and naïve Bayes classifiers [13]. Their study reported classification accuracy of up to 98% using a dataset of 372 reports. However, this task was formulated as a binary classification problem. Binary classification problems are generally less complex than the multiclass settings typically encountered in operational patient safety monitoring.

Overall, existing studies demonstrate the potential of machine learning for supporting PSE report classification. However, most prior approaches rely on traditional representations of text, such as term frequency or TF-IDF, which may have limited ability to capture contextual information within clinical narratives. These limitations highlight the need for approaches that leverage contextual language representations capable of modeling semantic relationships within incident narratives. Utilizing contextual text representations generated from transformer-based large language models may enable more accurate classification of patient safety events.

2.4 Machine Learning for Triage High-severity Patient Safety Events

Timely identification of high-severity patient safety events (PSEs) is critical for enabling rapid institutional response and preventing the recurrence of serious patient harm. Because the investigation of safety events requires substantial clinical expertise and organizational resources, patient safety specialists must triage incoming reports to prioritize those associated with the greatest potential harm [34,35]. As the volume of incident reports continues to increase, manual triage has become increasingly challenging, highlighting the need for scalable methods to support PSE triage.

Prior studies have explored the application of machine learning and natural language processing techniques to support the automated severity assessment of PSE reports. For example, Ong et al. applied naïve Bayes and support vector machine (SVM) models to identify extreme-risk incidents using textual features derived from narrative descriptions [34]. Their SVM model achieved an F-measure of 0.86 and an AUC of 0.92 in identifying extreme-risk incidents in a multi-type dataset, and detected 96.4% of extreme-risk cases in a specialized dataset of patient misidentification incidents. Evans et al. investigated supervised machine learning approaches for automatically predicting the severity of patient safety incidents from free-text reports [67]. More recently, Wang et al. examined the use of deep learning models, including convolutional neural networks, to classify both the type and severity of safety events using incident report narratives [30]. Using data collected from two independent incident report systems, their model achieved F1-scores exceeding 85% for common incident categories and improved severity classification performance by 11.9%–45.1% compared with SVM-based baselines.

Although these studies demonstrate the feasibility of using machine learning methods to support severity assessment, most prior work has formulated the task as a classification problem, in which each report is assigned to a predefined severity category. While classification models can achieve strong predictive performance, this formulation may not fully reflect the practical needs of patient safety specialists. In operational settings, safety teams are primarily concerned with prioritizing the most severe events for investigation, rather than assigning categorical severity labels to every report. As a result, methods that directly support prioritization may be better suited for real-world safety triage workflows. Ranking-based machine learning approaches provide a promising alternative by directly modeling the prioritization of high-severity events.

2.5 Machine Learning for Detecting Stigmatizing Language in Clinical Documentation

Language in clinical documentation can influence how patients are perceived and subsequently treated within healthcare systems. In particular, stigmatizing language can result in dignitary harms and may negatively affect patients' psychological well-being, trust in clinicians, and engagement with healthcare services [16,18,23,24].

To detect stigmatizing language in clinical text, early studies explored keyword-based approaches. Specifically, these approaches rely on predefined lists of words commonly associated with

stigmatizing language, such as “abuse”, “fail”, or “combative” [85–87]. However, the meaning of a word is often dependent on its surrounding context. For example, the word “fail” may appear in stigmatizing phrases such as “the patient failed treatment,” but it may also appear in clinically appropriate contexts such as “treatment failure.” As a result, keyword-based methods have inherent limitations in accurately identifying such language in clinical documentation.

Recognizing these limitations, recent studies have applied machine learning approaches to detect stigmatizing language in clinical text using manually annotated clinical notes [25,26,36]. For example, Harrigian et al. developed supervised models to classify stigmatizing language in electronic health records, reporting macro F1-scores of up to 0.93 using BERT-based models [25]. Similarly, Sun et al. developed a logistic regression classifier with TF–IDF features to identify stigmatizing language in history and physical notes, achieving an F1-score of 0.94 [36]. In these approaches, candidate sentences containing potentially stigmatizing expressions were first identified, after which supervised models classified whether the language reflected stigmatizing descriptions of patients.

However, supervised learning approaches typically rely on large quantities of manually annotated clinical text corpora, often requiring thousands of labeled examples. Producing such annotations is both time-consuming and resource-intensive. This challenge is compounded by the fact that stigmatizing language is often subtle and context-dependent, requiring both clinical domain expertise and linguistic judgment to determine whether a phrase reflects stigmatizing intent [19,25,36]. These challenges highlight the need for more data-efficient machine learning approaches that can accurately detect stigmatizing language while reducing reliance on extensive manual annotation. One emerging approach is in-context learning, in which large language models perform classification by conditioning on a small number of labeled examples provided within the prompt [65,76,88]. This paradigm enables effective detection while substantially reducing the need for large annotated datasets.

Chapter 3

Advancing Reliable and Explainable AI for Patient Safety Event Classification

3 A Machine Learning Approach with Human-AI Collaboration for Automated Classification of Patient Safety Event Reports: Algorithm Development and Validation Study

3.1 Abstract

Background: Adverse events refer to incidents with potential or actual harm to patients in hospitals. These events are typically documented through patient safety event (PSE) reports, which consist of detailed narratives providing contextual information on the occurrences. Accurate classification of PSE reports is crucial for patient safety monitoring. However, this process faces challenges due to inconsistencies in classifications and the sheer volume of reports. Recent advancements in text representation, particularly contextual text representation derived from transformer-based language models, offer a promising solution for more precise PSE report classification. Integrating the machine learning (ML) classifier necessitates a balance between human expertise and artificial intelligence (AI). Central to this integration is the concept of explainability, which is crucial for building trust and ensuring effective human-AI collaboration.

Objective: This study aims to investigate the efficacy of ML classifiers trained using contextual text representation in automatically classifying PSE reports. Furthermore, the study presents an interface that integrates the ML classifier with the explainability technique to facilitate human-AI collaboration for PSE report classification.

Methods: This study used a data set of 861 PSE reports from a large academic hospital's maternity units in the Southeastern United States. Various ML classifiers were trained with both static and contextual text representations of PSE reports. The trained ML classifiers were evaluated with multiclass classification metrics and the confusion matrix. The local interpretable model-agnostic explanations (LIME) technique was used to provide the rationale for the ML classifier's predictions. An interface that integrates the ML classifier with the LIME technique was designed for incident reporting systems.

Results: The top-performing classifier using contextual representation was able to obtain an accuracy of 75.4% (95/126) compared to an accuracy of 66.7% (84/126) by the top-performing classifier trained using static text representation. A PSE reporting interface has been designed to facilitate human-AI collaboration in PSE report classification. In this design, the ML classifier recommends the top 2 most probable event types, along with the explanations for the prediction, enabling PSE reporters and patient safety analysts to choose the most suitable one. The LIME technique showed that the classifier occasionally relies on arbitrary words for classification, emphasizing the necessity of human oversight.

Conclusions: This study demonstrates that training ML classifiers with contextual text representations can significantly enhance the accuracy of PSE report classification. The interface designed in this study lays the foundation for human-AI collaboration in the classification of PSE reports. The insights gained from this research enhance the decision-making process in PSE report classification, enabling hospitals to more efficiently identify potential risks and hazards and enabling patient safety analysts to take timely actions to prevent patient harm.

3.2 Introduction

Since the publication of the seminal report on patient safety – To Err Is Human [1], the importance of preventing adverse events in healthcare has been widely recognized. Adverse events refer to unintended or unexpected incidents that occur during hospital care that cause harm to a patient [14]. Common adverse events include complications, falls, and medication errors. These events can lead to prolonged hospital stays, permanent harm to patients, life-saving interventions, or even contributing to patient deaths [14,15]. Unfortunately, adverse events remain one of the top ten leading causes of death and disability worldwide, resulting in 251,454 deaths annually in the US alone [4]. In Organisation for Economic Co-operation and Development (OECD) countries, 15% of total hospital activity is the direct result of adverse events [6]. The global cost of adverse events has been estimated at 42 billion USD annually [89].

Patient Safety Event (PSE) reporting systems, also called incident reporting systems, have been widely adopted in hospitals across the world as part of their efforts to mitigate adverse events and improve patient safety [7,8]. Several jurisdictions have established legislation or national frameworks governing aspects of patient safety incident reporting [112–114]. The primary purpose of the PSE reporting system is to provide healthcare organizations with a centralized system for tracking and analyzing PSEs, thereby facilitating continuous learning and maintaining a record of PSEs for risk assessment and prevention [7,11]. PSE reporting systems are tools that allow frontline healthcare personnel to voluntarily report adverse events, near-misses, and unsafe conditions [9]. Each PSE report includes structured data, such as event types, patient harm level, date and location of the event, as well as unstructured data, including a free text section that contains the factual description of the event and the patient's outcome [13]. Following submission, PSE reports are reviewed by relevant hospital staff such as risk managers, patient safety analysts, nurse managers, physicians, and biomedical engineers to identify areas for patient safety and quality improvement within the hospital [44].

Accurately classifying PSE reports into their appropriate event type is crucial to ensure that these reports are directed to the relevant patient safety analyst, support organizational learning, identify patterns and trends in adverse events, and ultimately, prioritize measures to reduce adverse events [27,46]. An event type refers to a specific class of events that share common characteristics [3]. Examples of event types include fall, medication related issue, and diagnosis error [31,90]. PSE

reporting systems may have upwards of 20 categories of events. These classification taxonomies organize PSE reports according to shared incident characteristics [3]. The descriptions of event types are not always readily accessible to PSE reporters and patient safety analysts [46]. Previous studies have found that the classification of PSE reports is inconsistent depending on the reporter's profession, interpretation of the adverse event, and understanding of the PSE classification taxonomy [46,91]. Furthermore, 25% of PSE reports are labeled with vague or non-specific categories such as "miscellaneous" and "other" and require time-consuming retrospective analysis for reclassification [92]. These problems are further exacerbated by the growing volume of PSEs reported [90,93]. For instance, hospitals in the state of New South Wales in Australia reported close to 195,000 PSEs in 2020 [29], while there were approximately 2.3 million PSEs reported to the National Reporting & Learning System in England from April 2021 to March 2022 [10].

In light of these challenges, it is imperative to find an efficient solution to ensure the reliable classification of PSE reports. Recent studies have utilized static text representations and supervised machine learning (ML) techniques to automate the PSE report classification [2,31,67]. However, static text representations ignore the ordering of the words and do not account for the differences in word meaning across different contexts. These limitations may result in suboptimal classification performance. With the emergence of deep learning, contextual text representation produced from transformer-based deep learning (DL) models has achieved state-of-the-art performance on a wide range of natural language processing tasks including text classification [94]. The contextual representation of each word is based on its surrounding context within the text, allowing for a more accurate understanding of its usage across different contexts and facilitating knowledge transfer across languages [56]. Therefore, employing contextual text representation in training ML classifiers presents a promising opportunity for achieving a more precise classification of PSE reports.

The integration of ML models into PSE reporting systems has important implications for human-AI collaboration, given the role of the incident reporter (front-end) and patient safety analyst (back-end). Various approaches for employing ML classifiers can be developed, including at different levels of automation, however, unifying the strength of both human expertise and artificial intelligence (AI) offers the most promising route for effective implementation [95–97]. A crucial determinant for successfully implementing the human-AI collaboration approach is the decision transparency [98,99], which is often referred to as explainability. Explainability is the concept that

an ML model's prediction can be explained in a way that human operators can comprehend and reconstruct the model's reasoning [99]. Incorporating explainability techniques in human-AI collaboration is paramount as it facilitates a deeper understanding of the factors influencing the predictions, thereby fostering trust and understanding between human experts and AI systems. Therefore, embedding explainability into the human-AI collaboration holds significant potential for enhancing PSE report classification.

3.2.1 Study Objectives

The main aim of this study is to examine the efficacy of contextual text representation in improving the accuracy of PSE report classification. To accomplish this, we trained, evaluated, and compared various ML classifiers with both static and contextual text representations. Additionally, we developed an interface to illustrate the integration of the ML classifier in an event reporting system to support human-AI collaboration for PSE report classification. Moreover, we enhanced the explainability of the ML classifiers by employing an explainable AI technique. Furthermore, we have investigated the ML classifier's performance under two conditions differentiated by whether the explanation is valid for the predicted event type. Based on this analysis, we offer recommendations for optimizing human-AI collaboration in the context of PSE report classification.

3.3 Methods

3.3.1 Data Collection

The data set for this study was obtained from a large academic hospital located in the Southeastern United States. A total of 861 PSE reports from the labor and delivery and mother-baby units were extracted from the PSE reporting system from January 1, 2019, to December 31, 2020. Each PSE report was assigned to a single event type from a set of 25 classes, such as complication of the surgery, fall, medication-related, and supply issues. Event type labels were reviewed by patient safety experts, providing a high-quality reference standard for model training and evaluation. The ML classifiers were trained exclusively on PSE reports from the 7 most frequently occurring event types. This selection was intended to create a more balanced training data set to reduce sampling bias and the risk of overfitting. The selected PSE reports used for training ML classifiers constitute approximately 72.8% (627/861) of the extracted reports (Table 3-1).

Table 3-1 Prevalence of patient safety event reports by event type.

Event Type	Frequency of occurrence	Percentage of total extracted reports
Care coordination/communication	186	21.6
Laboratory test	122	14.2
Medication related	89	10.3
Omission/errors in assessment, diagnosis, monitoring	67	7.8
Maternal	58	6.7
Equipment/devices	56	6.5
Supplies	49	5.7
Total	627	72.8

3.3.2 Data Preprocessing

The free-text section of PSE reports was preprocessed before feeding into ML classifiers as input features. The preprocessing procedures include text normalization, feature extraction, data splitting, and data augmentation is presented in Appendix 3.1 Data Preprocessing Procedures.

3.3.3 Classifier Training

A range of ML classifiers, including multinomial logistic regression (MLR), support vector machine (SVM), extreme gradient boosting, light gradient boosting, random forest (RF), k-nearest neighbor (KNN), and multilayer perceptron, were used for the classification of PSE reports. While SVM is a binary classifier, it is also capable of performing multiclass classification using the one-versus-one strategy. This involves treating the multiclass classification problem as a series of binary classification problems, creating $n \times (n - 1) / 2$ binary classifiers for each pair of classes, where n represents the total number of classes, and the final classification is based on the majority vote of all binary classifiers. Extreme gradient boosting, light gradient boosting, and RF are tree-based ensemble algorithms that are commonly used in text classification tasks [31,100]. The KNN classifier predicts the class of a data point based on the majority class among its nearest neighbors

in the training data set. Multilayer perceptron is a feedforward neural network consisting of multiple layers of interconnected neurons and trained using backpropagation. For model development, the incident description field was used as the sole input feature, while the corresponding event type served as the prediction target, thereby minimizing the risk of data leakage from structured reporting fields.

To optimize the performance of ML classifiers, we used the 5-fold cross-validation grid search technique to identify the best combination of hyperparameters. During this process, a range of values of important hyperparameters (i.e., regularization strength) is assessed with 5-fold cross-validation. For each combination of hyperparameters, the training set is randomly split into 5 distinct folds, and then the ML classifier is trained and evaluated 5 times, picking a different fold for evaluation every time and training on the remaining 4 folds. The optimized combination of hyperparameters is determined based on the average performance of the classifier on the $F1$ -score across the 5-fold cross-validation runs.

3.3.4 Classifier Evaluation

We evaluated the performance of the trained classifiers on the testing set with standard classification metrics, including accuracy, precision, recall, $F1$ -score, and area under the receiver operating characteristic curve. We also evaluated classifiers on top-2 accuracy, which measures the proportion of predictions where the correct event type is among the top 2 highest probability event types predicted by the classifier. The definitions and mathematical formulas of the evaluation metrics are shown in Appendix 3.2 Evaluation Metrics for Examining Machine Learning Classifiers. Each of these metrics provides a distinct perspective on the performance of the classifier, and collectively, they offer a comprehensive understanding of how well the classifier is functioning. Since we framed PSE report classification as a multiclass text classification problem, the precision, recall, $F1$ -score, and area under the receiver operating characteristic curve are computed for each class and combined using a weighted average where the weights correspond to the number of data points in each class.

3.3.5 Development and Assessment of Explainability

As the contextual text representation is generated from transformer-based neural network, which has a black box nature, we used the local interpretable model-agnostic explanations (LIME)

technique to analyze the top-performing ML classifier trained with the contextual text representation. LIME is a post hoc, local perturbation technique that provides the explanation for a single prediction. LIME generates perturbed data by randomly removing words from a text document and trains a locally explainable model with perturbed data to simulate the original classifier's prediction [101]. By measuring how the classifier's prediction changes under these perturbations, LIME reflects the contributions of each word to the prediction. The importance of each word can then be assessed for a single prediction, revealing whether the ML classifier has learned to use relevant words for classifying PSE reports. We used LIME to generate explanations for the top-performing classifier's prediction, specifically by highlighting the words that the classifier deems influential for the prediction. We presented 3 distinct cases: one where the classifier effectively leveraged relevant words for accurate prediction, another where it failed to do so, and a final case that illustrated the explanation for a misclassification. In addition, we analyzed the top 5 most prevalent words identified by LIME for each event type.

A total of 2 human factors graduate students were recruited to assess the quality of the LIME explanations. For each PSE report in the test data set, the reviewers were asked to determine independently if any of the highlighted words were relevant to the predicted event type. Based on these evaluations, the reports were then categorized into 2 distinct groups: those in which the highlighted terms were deemed relevant to the predicted event types and those where they were deemed irrelevant. Discrepancies were resolved through discussions. The interrater reliability index (Cohen κ) was calculated to quantify the level of agreement between the reviewers. The ML classifier's accuracy and F1-score were evaluated for these 2 groups of PSE reports. A subsequent comparison will explore the influence of explanation quality on prediction reliability.

3.3.6 Interface Development

In the typical workflow of PSE report classification, reporters need to provide a narrative description of the event as well as key attributes such as the event type, level of harm, date, and location of the event. Subsequent to this initial classification, the patient safety analyst will review the submitted report and decide if it needs to be recategorized to better reflect the nature of the event [31,102]. To support efficient and reliable categorization, the classifier will need to provide reporters with real-time support during the reporting process. We developed a PSE reporting interface to illustrate the integration of the ML classifier and the LIME explainability technique.

In the design, the ML classifier provides multiple high-probability event types along with explanations for its prediction and allows the user to select the most appropriate event type. The interface was developed in Figma [103] and designed using guidance from previous research on incident reporting systems, including question type, mandatory and optional questions, and taxonomy for event type and harm level [104,105].

3.3.7 Ethical Considerations

The study was approved by the Medical University of South Carolina Hospital’s institutional review board (Pro00105892). Following data extraction, PSE reports were anonymized in accordance with privacy regulation guidelines. Informed consent was not required, as the study qualified as a quality improvement project involving secondary use of anonymized data. No compensation was applicable for the use of these data. Figures and supplementary materials in this manuscript do not contain any identifiable information regarding individual participants or users.

3.4 Results

3.4.1 Performance Comparison

We evaluated the trained ML classifier’s classification performance on both static and contextual text representations (Appendix 3.3 The Performance of ML Classifiers in Classifying PSE Report Event Type.). The performance of the top-performing ML classifier trained with static and contextual text representations is shown in Table 3-2. Our results showed that for static text representation, the MLR classifier trained with term frequency–inverse document frequency (TF-IDF) achieved the best performance, with an *F1*-score of 0.631 and an accuracy of 66.7% (84/126). On the other hand, for contextual text representation, the SVM classifier trained with RoBERTa-base outperformed others, with an *F1*-score of 0.753 and an accuracy of 75.4% (95/126). The SVM classifier trained with RoBERTa-base showed a 19.3% relative improvement in *F1*-score and a 13% (11/85) relative improvement in accuracy compared to the MLR classifier trained with TF-IDF for contextual text representation. In addition, we compared the accuracy (95/126, 75.4%) and top 2 accuracy (107/126, 84.9%) of the SVM classifier trained with RoBERTa-base and observed that 9.5% (12/126) of PSE reports’ true event type was predicted as the second highest probability event type by the classifier, which represents 39% (12/31) of misclassified PSE reports.

Table 3-2 Performance of top-performing ML classifiers trained with static and contextual text representations.

Metrics	Top-performing ML model trained with the static text representation			Top-performing ML model trained with the contextual text representation		
	Performance	ML classifier	Text Representation	Performance	ML classifier	Text Representation
Accuracy (%)	66.67	MLR	TF-IDF	75.40	SVM	RoBERTa-base
Top 2 Accuracy (%)	85.71	MLR	TF-IDF	88.10	MLP	xlm-RoBERTa-base
Precision	0.707	KNN	TF-IDF	0.757	SVM	RoBERTa-base
Recall	0.667	MLR	TF-IDF	0.754	SVM	RoBERTa-base
F1	0.631	MLR	TF-IDF	0.753	SVM	RoBERTa-base

3.4.2 Performance on Classifying Individual Event Types

We analyzed the performance of the SVM classifier trained with RoBERTa-base on individual event types (Table 3-3). The F1-score measure for different event types ranged from 0.958 (laboratory test) to 0.400 (omission or errors in assessment, diagnosis, and monitoring).

Table 3-3 Performance of SVM+RoBERTa-base on the individual event type.

Event Type	Precision	Recall	F1 score
Care coordination/communication	0.721	0.838	0.775
Laboratory test	1.000	0.920	0.958
Medication related	0.765	0.722	0.743

Omission/errors in assessment, diagnosis, monitoring	0.417	0.385	0.400
Maternal	0.750	0.750	0.750
Equipment/devices	0.700	0.636	0.667
Supplies	0.778	0.700	0.737

Figure 3-1 shows the confusion matrix for the SVM classifier trained with RoBERTa-base evaluated on the test set. A confusion matrix is a table that visualizes the performance of a classifier. The main diagonal value is the number of PSE reports that have been classified as true event types, whereas off-diagonal values are the number of PSE reports that have been wrongly classified. To facilitate interpretation, diagonal percentages in the confusion matrix represent the percentage of correctly classified instances within each ground-truth category (i.e., class-wise recall). For example, the diagonal percentage in the first row (83.8%) indicates that 83.8% of Care Coordination events were correctly classified as Care Coordination. While the classifier was able to classify the majority of event types of PSE reports correctly, there is a consistent misclassification of the omission or errors in assessment, diagnosis, or monitoring PSE report as the care coordination or communication (coordination) event type.

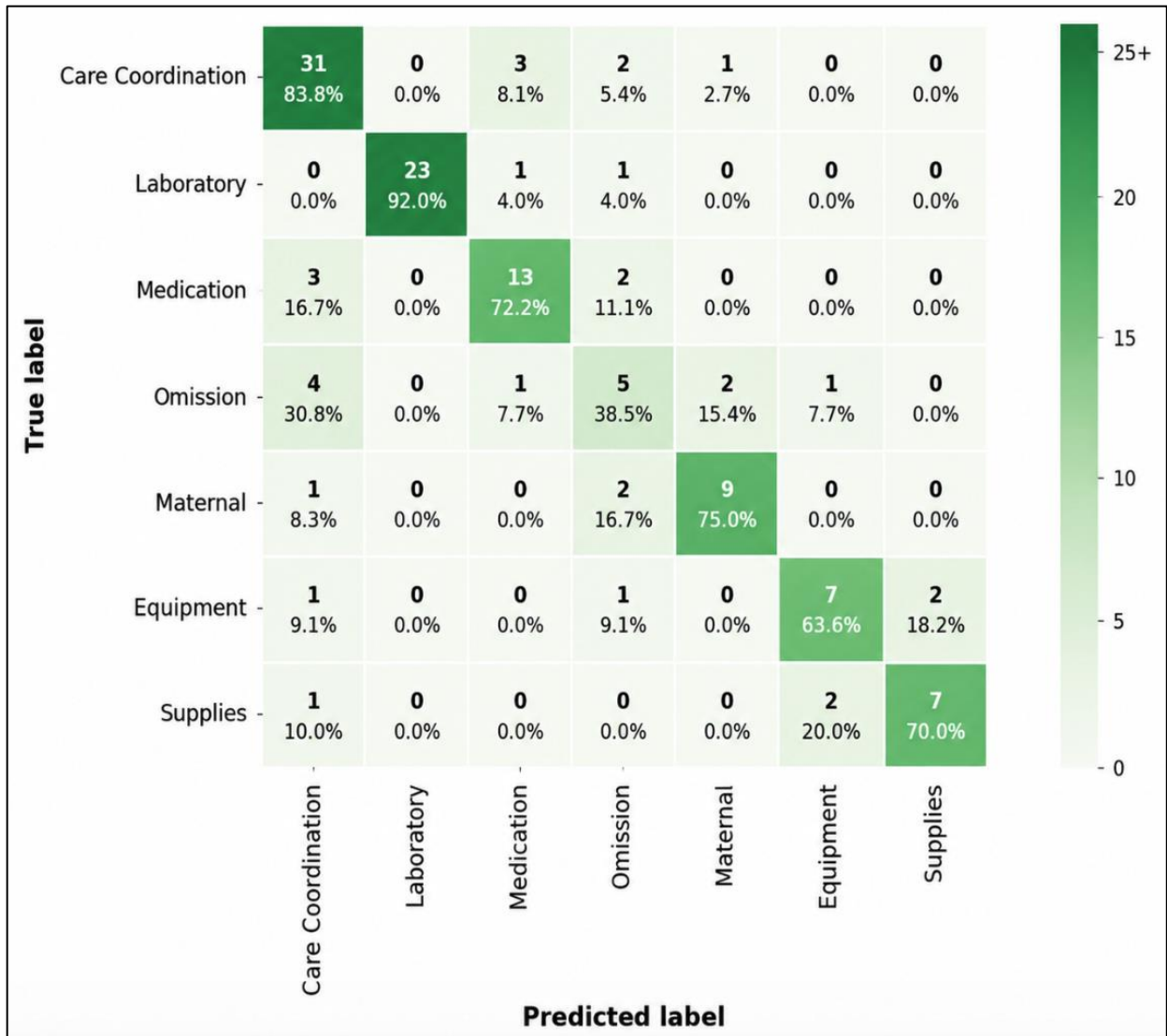


Figure 3-1 Confusion matrix for the testing set evaluation with the SVM classifier trained with RoBERTa-base.

3.4.3 LIME-Based Explainability Analysis

We used LIME to evaluate whether the SVM classifier trained with RoBERTa-base has leveraged informative words for classification. Figure 3-2 presents 3 examples of explanations for the classifier's predictions. At the top of Figure 3-2, LIME identified “ketorolac,” “ibuprofen,” and “doses” from the PSE report as important words for classifying the report into the medication-related event type, which is reasonable given the report's association with incorrect medication doses. Conversely, in the middle of Figure 3-2, LIME highlighted “our,” “handle,” and “or” from the text as important words for classifying the report into the equipment or device event type.

Although the predicted event type was correct, the classifier relied on irrelevant words for the classification. At the bottom of Figure 3-2, a case of misclassification is shown. LIME highlighted “pitocin,” “pump,” “available,” and “use” as influential words for classifying the PSE report into medication-related event type when it belongs to the equipment class. In addition, for each event type, we extract the 5 most prevalent words that were deemed important for the classifier’s prediction across the whole data set (Table 3-4). This inclusion of stop words (i.e., “was,” “not,” and “till”) among influential terms, as shown in Table 3-4, demonstrated that the classifier does not always rely on relevant words for making classifications.

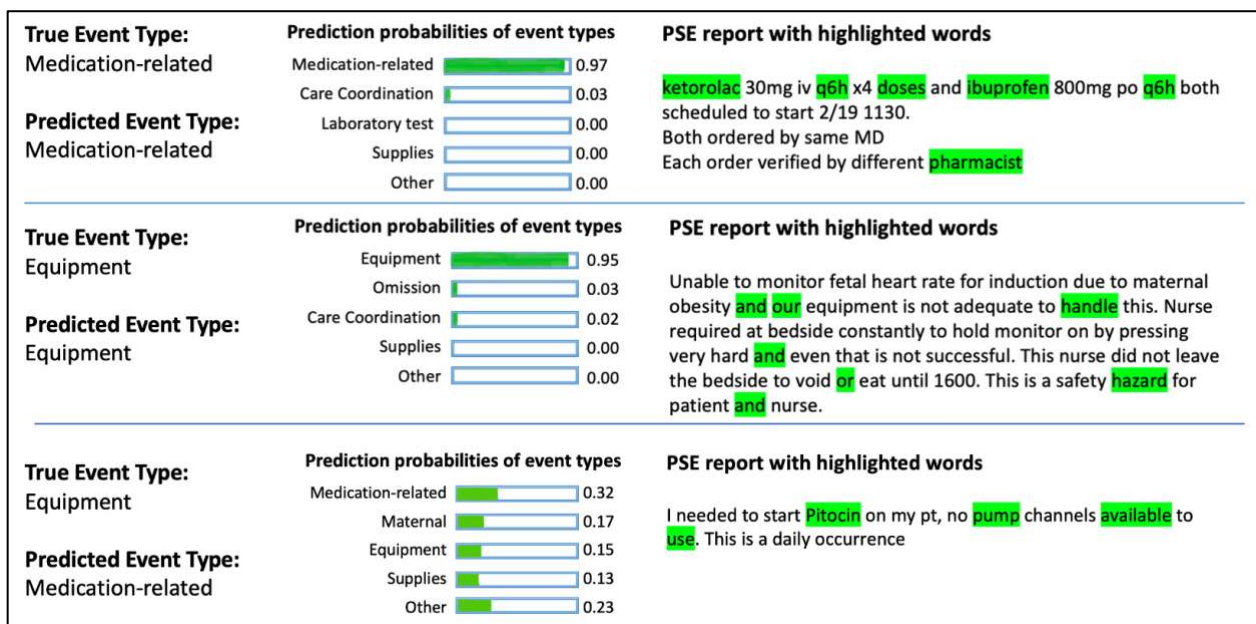


Figure 3-2 LIME explanations of the SVM classifier trained with RoBERTa-base.

Table 3-4 The five most prevalent important words for each event type that were derived from the SVM classifier trained with RoBERTa-base.

Event Type	Prevalent Influential Words highlighted by LIME
Care coordination/communication	requested, delayed, patient, not, follow
Laboratory test	specimen, lab, labels, collection, results
Medication related	patches, doses, orders, medication, pitocin
Omission/errors in assessment, diagnosis, monitoring	warning, patient, was, till, late

Maternal	baby, hysterectomy, stable, pumping, hemorrhage
Equipment/devices	instruments, trays, notified, malfunctioning, faulty
Supplies	vendor, sterile, available, needed, OR

After reviewing the LIME explanations for each PSE report in the test data set, 73.8% (93/126) of the reports were categorized into a subset where at least 1 highlighted word was deemed relevant to the predicted event type. The remaining reports comprised a second subset where no highlighted words were relevant. The interrater reliability index measured by Cohen κ between the 2 reviewers was 0.83, indicating substantial agreement. Table 3-5 presents the performance of the top-performing ML classifier for both subsets. For the first subset, the classifier achieved an accuracy of 84% (78/93) and an *F1*-score of 0.825. In contrast, the second subset showed a classifier accuracy of 52% (17/33) and an *F1*-score of 0.549.

Table 3-5 Performance of the top-performing ML classifier on reports that have relevant words highlighted and reports with irrelevant words highlighted.

	PSE reports with relevant words highlighted	PSE reports with irrelevant words highlighted
Number of PSE reports	93	33
Percentage of test dataset (%)	73.81	26.19
Accuracy (%)	83.87	51.51
F1	0.825	0.549

3.4.4 PSE Reporting System Interface

We designed an event reporting interface that integrates both the ML classifier and the LIME explainability technique. Figure 3-3 shows the event classification screen, where reporters enter a

narrative description of the event after providing the details of the event, including date, time, unit, and information about the patient and reporter. Before describing the event in narrative form, reporters also choose among factors that contributed to the incident and the level of harm experienced by the patient. Once the reporter enters their narrative and selects the “classify” button, the system activates the ML classifier. Subsequently, the interface displays the top 2 most probable event types, along with their associated probability distributions, in the lower left section. Simultaneously, the LIME technique will identify influential words that significantly contributed to the predicted event type, highlighting these words in green in the upper section of the dashboard. Based on the predicted event types and words highlighted for their influence on the prediction, the reporter may select the most suitable event type from a drop-down menu located in the lower-right section of the dashboard. Following this selection, reporters are queried on whether they agree with the classifier’s prediction, and the collected data can be used to guide subsequent refinement of the ML classifier.

Patient safety event description:
Please specify the details of what happened and how it may have happened.

ketorolac 30mg iv q6h x4 doses and ibuprofen 800mg po q6h both scheduled to start 2/19 1130.
Both ordered by same MD
Each order verified by different pharmacist

Classify
* Words highlighted are influential for recommended event types.

Machine learning classifier’s prediction

Event Type	Probability
Medicated-related	0.97
Care Coordination	0.03

Select the event type :

Select an option

Do you agree or disagree with classifier’s prediction:

Agree

Disagree

Figure 3-3 Interface visualization of a patient safety event report classifier coupled with the local interpretable model-agnostic explanations technique.

3.5 Discussion

3.5.1 Overview

PSE event reporting systems are commonly used in health systems and hospitals across the world [106]. Data collected in PSE reporting systems drive quality improvement and patient safety efforts and supports regulatory reporting requirements for hospitals. The erroneous classification of PSE reports can impede the learning capabilities of the PSE reporting system, leading to suboptimal performance in detecting and preventing potential patient safety hazards [91]. It can also result in a substantial time cost for reclassifying PSE reports and compromise the integrity of a PSE database when analysts are investigating trends in events to develop effective solutions [31]. Previous studies have trained ML classifiers with static text representations for automatic PSE classification [2,13,31,67]. This study aimed to investigate whether using contextual text representations can further improve the accuracy of classifying PSE reports. We trained and evaluated a range of ML classifiers using both static and contextual text representations. To the best of our knowledge, this is the first time that contextual text representation has been used for training ML classifiers for PSE report classification. We analyzed the confusion matrix of the top-performing classifier to identify prevalent misclassified event types. Furthermore, aiming for more accurate and reliable PSE report classification, we incorporated an explainability technique to support human-AI collaboration and designed an interface to illustrate the possible integration of the ML classifier in PSE reporting systems.

3.5.2 Principal Findings

In this study, we extensively investigated the potential of using contextual representation for improving PSE report classification. The leading classifier trained with the static text representation (MLR trained with TF-IDF) was able to achieve an accuracy of 66.7% (84/126). This accuracy considerably exceeds the baseline accuracy of 29.4% (37/126), which involves classifying all PSE reports into the majority event type. However, using contextual text representation proved more efficacious. The SVM trained with contextual text representation (RoBERTa-base) was able to achieve an accuracy of 75.4% (95/126), reflecting a 13% (11/84) relative improvement in accuracy compared to the best-performing classifier trained with static text representation. While the achieved accuracy of 75.4% may not appear outstanding in isolation, it represents a significant advance compared with static text representations and exceeds the

baseline, given the limited size of the data set. The improvement in classifier performance can be attributed to the use of contextual text representations, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which relies primarily on word frequency and co-occurrence to represent text. Moreover, contextual text representation does not require explicit text normalization while also avoiding issues associated with high-dimensionality and sparsity commonly found in static text representations. Hence, when training ML classifiers for PSE reporting systems, contextual text representation should be prioritized over static text representation to ensure the highest level of accuracy in classifying PSE reports.

As part of our investigation, we evaluated the performance of the top-performing classifier trained with contextual text representation on individual event types. While the classifier demonstrated impressive performance in accurately classifying laboratory test PSE reports (F1-score=0.958), it struggled with classifying omissions or errors in assessment, diagnosis, and monitoring PSE reports, resulting in an unsatisfactory F1-score of 0.400. To investigate this discrepancy, we analyzed the confusion matrix for the classifier and discovered that omissions or errors in assessment, diagnosis, and monitoring PSE reports were frequently misclassified as the coordination event type. This misclassification can be attributed to the multiclass nature of PSE reports. For example, a failure to document the removal of a patient's epidural catheter (omission or errors in assessment, diagnosis, and monitoring) could lead to a medication ordered by a physician (such as Lovenox) being withheld by the pharmacy due to a complication risk (coordination). On the other hand, the laboratory test is a more distinct event type in comparison to the other event types, and the classifier was able to correctly classify the majority of these reports. The observation obtained from the confusion matrix implies that PSE reports can potentially have more than 1 event type. This finding is consistent with previous studies [2,67]. This reflects an important distinction between the multiclass structure imposed by the reporting system and the multidimensional nature of the underlying events. Although the model assigns each report to one of several mutually exclusive categories, a PSE may involve characteristics associated with multiple event types. In practice, incident reporting systems typically require reporters or patient safety personnel to assign a single primary event type to facilitate routing and investigation of the report. However, when a PSE spans multiple contributing domains, forcing a

single-category assignment may oversimplify the underlying event and contribute to misclassification or reclassification during downstream review, which may affect the timeliness and efficiency of directing reports to the most appropriate personnel for follow-up and system-level learning. The multi-event type nature of PSE reports also underscores the need for further refinement in the development of the PSE taxonomy to create more distinctive event types. Another potential solution for addressing the multiclass nature of PSEs is to enable multiple event-type assignments [107]. It is important to highlight that the confusion matrix should be interpreted in relation to the number of test-set reports available for each event category. Some categories contained relatively few observations; therefore, their class-specific results may be unstable and disproportionately affected by a small number of correct or incorrect predictions. Consequently, the apparent performance differences across underrepresented categories should be interpreted cautiously and should not be considered conclusive evidence of the model's ability to classify these event types. Evaluation using larger and more representative samples is needed to obtain more reliable class-specific performance estimates. Alternatively, the ML classifier can provide several probable event types, allowing the user to select the most appropriate one. We evaluated the top 2 accuracy of the top-performing ML classifier trained with contextual text representation and observed that 39% (12/31) of misclassified PSE reports' true event type was predicted as the second-highest probability event type by the classifier. The finding suggests that there is a greater chance for the ML classifier to provide the correct event type when considering multiple options. As event reporting systems usually encompass over 20 event types, which can be difficult to memorize or access [31], narrowing down the PSE report's potential event types to a smaller range also reduces the cognitive workload for PSE reporters during the classification process [108] and enhances the efficiency of reclassifying PSE reports for patient safety analysts.

We used LIME to showcase 3 predictions' explanations and demonstrated cases where the ML classifier used informative words for classifying the PSE report and where it used irrelevant words for classification. These results highlight the importance of not solely relying on the ML classifier's prediction and underscore the need for explainability and transparency in using the ML classifier for PSE report classification. Additionally, we showed the top 5 most prevalent words the ML classifier deemed important in the PSE reports for each event type. These words are indicative of the prevalent themes and issues within specific event types. Understanding the context and relationships between these prevalent informative words and specific event types can

potentially provide valuable insights into the factors contributing to different types of PSEs. Furthermore, we have evaluated the top-performing ML classifier's performance on 2 subsets of PSE reports, differentiated by whether the highlighted word by LIME is relevant to the predicted event type. Our findings reveal that the majority of PSE reports (93/126) have at least 1 relevant word highlighted, with the classifier achieving an accuracy of 84% (78/93) on these reports. Conversely, accuracy drops to 52% (17/33) when irrelevant words are highlighted. Such a disparity in performance emphasizes the necessity for additional scrutiny from reporters and patient safety analysts, particularly when dealing with PSE reports that have irrelevant words highlighted.

While previous research has focused on the development of ML classifiers, none of these previous works have investigated the potential integration of the classifier within the PSE reporting system in a manner that aligns with the workflow of the front-end reporter. We designed an interface to demonstrate the feasibility of a collaborative human-AI approach for event categorization. The interface provides the PSE reporter with multiple probable event types and associated explanations for the ML classifier's prediction. This approach aligns with the principles of level 2 automation, where ML classifiers aids human decision-making rather than fully automating it [108]. This collaboration optimally combines human expertise with ML capabilities, potentially reducing cognitive workload and memorization of the taxonomy while also reducing the risks associated with overreliance on automation. Numerous studies have shown that the human-AI collaboration approach can improve the decision-making process [109–111], indicating its potential for enhancing PSE report classification. Furthermore, the interface also integrates the LIME explainability technique, which offers real-time insights into the rationale for the probable event types. Given the role of reporters and patient safety analysts in the incident reporting process, the use of explainability techniques can also increase trust in the recommendation provided by the ML classifier as it provides transparent and interpretable reasoning for the classification decisions [47,48]. Using LIME to highlight top informative words in real time for a PSE report can assist PSE reporters by emphasizing keywords in their narratives that are linked to the proposed classification. Highlighting informative words can also facilitate patient safety analysts working at the back end by providing insights into why a specific event type was chosen for classification. Such transparency not only clarifies current recommendations but also guides analysts in identifying influential terms for future report classifications. Previous research has illustrated the value of automation transparency in supporting appropriate levels of trust in the system, including

decision support systems [29]. Additionally, regularly checking the explanations of the ML classifier's prediction enables continuous monitoring of the classifier's performance, identification of issues, and refinement [49]. As we have only designed the interface, additional research is needed to test the effectiveness of this approach in PSE report classification. Assessing the interface's impact on cognitive workload and decision-making accuracy is essential for ensuring its usability and adoption in the event reporting system. We plan to undertake a usability testing study with health care professionals in a subsequent study.

3.5.3 Comparison with Previous Work

Research into the use of ML classifiers for the automation of PSE report classification has been relatively scarce. Wang et al [2] used logistic regression and SVM with the binary count, term frequency, and TF-IDF text representation to classify ten types of PSE reports, reaching an F1-score as high as 0.783. However, they used a considerably larger data set ($n=2860$). Fong et al [31] achieved an accuracy rate of 92.0% (284/309) when they examined the usage of an ML classifier for classifying miscellaneous PSE reports using SVM, RF, and logistic regression with TF-IDF [16]. They also used a much larger data set ($n=70,051$). Ong et al [13] investigated the feasibility of using an ML classifier to automatically classify 2 types of PSE reports, including inadequate clinical handover and incorrect patient identification. They used Bag of Words model for text representation and trained both SVM and naive Bayes on classifying PSE reports, reaching accuracy as high as 98% (364/372). However, they framed the problem as a binary classification problem, which inherently has a higher baseline accuracy compared to our investigation. In this study, we've performed an in-depth comparative analysis with the available PSE data set and compared the established methods of classifying PSE reports and our novel method of using contextual text representations for classification. Our findings reveal that our proposed method outperforms the traditional models in terms of accuracy (i.e., 84/126, 66.7% vs 95/126, 75.4%) and F1-score (i.e., 0.631 vs 0.753). This underlines the significance of our approach and its potential to advance the field of using ML classifiers for PSE report classification.

Direct comparison of absolute model performance with prior studies should be interpreted cautiously, as external datasets were not publicly available for replication or benchmarking. Instead, this study focused on comparing the relative performance of machine learning approaches using static and contextual text representations within our own dataset. Differences in performance

across studies may also reflect variation in dataset characteristics, including sample size, as larger datasets may provide models with greater opportunities to learn representative patterns.

3.5.4 Limitations

There are several limitations to this study. First, the PSE reports used to train and evaluate the machine learning classifiers were obtained from selected maternal care units at a single hospital in the United States. The findings therefore primarily reflect the reporting practices, terminology, clinical workflows, and event taxonomy represented in this setting. Performance may differ in other clinical departments, where the types of safety events and language used to describe them may vary, or in other institutions that use different reporting systems and category definitions. Second, because of the limited dataset size, the analysis included only the seven most prevalent event categories. Less frequent event types were excluded and may be more difficult to classify because fewer examples would be available for model training and because their narratives may overlap with those of more common categories. Accordingly, the reported performance should not be assumed to extend to infrequent or previously unseen event types. External validation using reports from diverse clinical departments and institutions, followed by local retraining or calibration where necessary, is required before broader implementation. [13] Third, the quality of the LIME explanations was assessed by 2 graduate students; thus, further investigation is needed for a more robust validation of explanation quality. Furthermore, we have not yet empirically tested the interface for potential decision-making biases it may introduce. Future research should also extend the explainability component of the PSE classification framework. The present study used LIME to highlight words that influenced individual predictions; however, alternative explainability methods, such as SHAP and integrated gradients, could be compared in terms of faithfulness, stability, consistency, and interpretability. The clinical relevance of the resulting explanations should also be evaluated by patient safety specialists across a broader range of event types. Prospective studies could further examine whether providing explanations improves classification accuracy, decision time, user trust, and the appropriate identification of uncertain or erroneous predictions compared with presenting model recommendations alone. In addition, aggregating explanations across reports may help identify recurring linguistic patterns, potential spurious associations, and institutional reporting biases, thereby supporting continuous model monitoring and refinement.

Future research should evaluate classifiers trained with contextual text representations using larger and more diverse datasets that include less frequent event types, a broader range of clinical departments, and institutions using different reporting taxonomies. Where reporting categories differ across institutions, category mapping and local validation should be performed before deployment. Additionally, while we plan to refine the interface and test whether it supports event classification, future research can continue to investigate the appropriate way of incorporating the ML classifier into the reporting and reviewing workflow of PSE report classification and examine various human-AI collaboration approaches. Future studies should explore the potential biases (i.e., automation bias) that the interface may introduce into the analysts' decision-making process. Furthermore, future research should prospectively evaluate the classification system within the PSE review workflow. In addition to classification accuracy, evaluation should examine the time required to classify and review reports, agreement between patient safety specialists and model recommendations, changes in reviewer workload and backlog, and the usability of the interface. Comparisons between model-assisted and current review processes would help determine whether the system improves workflow efficiency and decision-making without introducing excessive reliance on model recommendations.

3.6 Conclusions

Improving the precision of PSE report classification is a multifaceted task, involving both the refinement of event type taxonomies and adequate training of hospital staff on reporting systems. Despite these challenges, machine learning classifiers offer substantial potential to enable scalable and systematic classification of large volumes of incident reports within institutional safety monitoring workflows.

The findings of this study demonstrate that contextual text representations substantially outperform static representations, leading to more accurate and reliable classification of PSE reports. These results highlight the importance of leveraging pretrained language models to capture nuanced clinical context in healthcare narratives. In addition, the integration of explainability techniques supports transparency in model predictions, fostering trust and facilitating informed decision-making during the review process. Furthermore, the proposed integration of machine learning classifiers within an event reporting interface with collaborative decision-making capabilities provides a practical pathway toward operational deployment. By improving the accuracy and

efficiency of PSE classification, these approaches contribute to more effective identification of safety risks and support timely interventions to mitigate patient harm.

Chapter 4

Enhancing Institutional Severity Triage Through Ranking-Based Machine Learning

4 Machine Learning to Prioritize High-Severity Patient Safety Events for Institutional Investigation: Algorithm Development and Validation Study

4.1 Abstract

Background: Patient safety events (PSEs) are preventable incidents that cause, or have the potential to cause harm to patients during their medical journey. Although incident reporting systems capture large volumes of such events, only a small proportion undergo comprehensive investigation due to the resource-intensive nature of the review process. Patient safety specialists are tasked with triaging PSEs to prioritize high-severity cases that warrant timely institutional investigation; however, the rapidly increasing volume of events has made manual triage increasingly impractical.

Objective: This study aimed to develop and validate ranking-based machine learning models for prioritizing high-severity PSEs and to compare their performance with conventional classification-based models and the Reporter-Assigned Severity Baseline.

Methods: A total of 101,239 PSE reports were retrospectively extracted from a large Canadian academic health system between January 2017 and March 2025. Each report included a free-text incident description and a corresponding severity label denoting the level of severity, assigned through an institutional review process. Both fixed-representation-based and transformer-based models were developed using ranking and classification frameworks. Model performance was

evaluated using ranking metrics, including Average Precision (AP) and Normalized Discounted Cumulative Gain (NDCG) at multiple cutoff thresholds ($k = 10, 25, 50, 100, 200$), as well as Precision@ k and Recall@ k . The best-performing model was further benchmarked against the Reporter-Assigned Severity Baseline.

Results: The ranking model, BERT-RA, achieved the highest mean AP of 0.91 (± 0.03), representing a 23.0% relative improvement over the best classification model (BIOBERT-CL; 0.74 ± 0.06 ; $P < 0.0008$). BERT-RA demonstrated superior performance across ten of sixteen evaluation metrics, with larger relative gains at broader evaluation cutoffs (e.g., +14.8% in NDCG@200, +25.8% in Precision@200, +27.0% in Recall@200). Compared with the Reporter-Assigned Severity Baseline, BERT-RA achieved significantly higher performance across sixteen metrics, retrieving 80.0% of high-severity PSEs within the top 200 ranked events, compared with 41.0% retrieved under the Reporter-Assigned Severity Baseline.

Conclusions: The BERT-RA demonstrated superior capability in prioritizing high-severity PSEs compared with both classification approaches and the Reporter-Assigned Severity Baseline. Integrating this model into triage workflows may offer a scalable and efficient solution for identifying and prioritizing high-severity safety events.

4.2 Introduction

Systematic reporting and analysis of adverse events are essential for improving patient safety in modern healthcare systems. According to the World Health Organization, adverse events are incidents associated with healthcare that result in, or have the potential to result in, harm to a patient, many of which are potentially preventable [2,3]. Common examples include medication errors, unsafe surgical procedures, and diagnostic errors [38,39]. In Canada, recent national data indicate that unintended harm occurs in approximately 6% of hospitalizations, corresponding to about one in every 17 hospital stays, and that multiple harmful events occur in roughly one-quarter of these cases [5]. Beyond clinical consequences, adverse events also impose substantial economic burdens. The indirect costs of unsafe care were estimated at USD 118 trillion globally between 2000 and 2020 [6].

The significant impact of adverse events has led some jurisdictions to establish legislation or formal frameworks addressing patient safety incident reporting [112–114]. As of 2021, nine out of thirteen jurisdictions in Canada require healthcare providers to report adverse events [114]. Systematic reporting enables institutions to detect unsafe practices and implement preventive strategies. Adverse events are typically documented as patient safety event (PSE) reports through institutional incident reporting systems (IRSs) [9,115]. PSE reports typically comprise both unstructured fields—such as narrative descriptions of the incident—and structured fields, including the date, location, event type, and severity of harm [13,45]. The submitted reports are investigated by designated hospital personnel, such as risk managers, patient safety specialists, nurse managers, and physicians, to identify contributing factors and root causes [44,116,117]. Although these reports offer valuable learning opportunities, their investigation is resource-intensive, demanding substantial time, expert knowledge, and interdepartmental coordination [12]. As a result, only a limited proportion of PSE reports undergo comprehensive investigations [34].

Hospitals prioritize the investigation of high-severity PSEs to optimize the effectiveness of limited resources [34,35]. The prioritization of high-severity PSEs is also driven by legislated investigation requirements, as well as ethical obligations to address patient harm, support affected individuals and families, and prevent recurrence through system-level mitigation [44,118]. Event reporters typically assign a severity rating within the IRS to reflect the perceived seriousness of an event. These ratings inform prioritization decisions for investigation efforts. However, research has shown that severity ratings vary substantially among healthcare professionals, due to differences in clinical roles and familiarity with the severity taxonomy [91,119]. To address this inconsistency, patient safety specialists frequently conduct retrospective manual reviews of submitted reports to identify high-severity cases that warrant timely investigation [34,47,48]. Nevertheless, the rapidly growing volume of reports has made manual triage increasingly impractical, underscoring the urgent need for scalable and efficient triage solutions. For example, over 4.4 million PSE reports were submitted to the National Reporting and Learning System in England between 2020 and 2022 [120]. In addition, reliance on manual triage consumes substantial clinical and safety personnel resources that could otherwise be directed toward quality improvement initiatives, and may introduce delays in response that increase organizational risk.

Recent studies have applied natural language processing (NLP) techniques to automatically classify the severity of PSE reports [2,30,34,67]. Ong et al. employed Naïve Bayes and Support Vector Machine models for severity classification [34], while Wang et al. investigated convolutional neural networks for predicting severity levels [30]. However, most prior studies have relied on classification-based approaches that focus on maximizing categorical accuracy, which may not fully align with the practical needs of patient safety specialists. In institutional settings, these specialists are responsible for prioritizing high-severity PSEs for investigation under significant time constraints and limited resources. In this context, accurately ranking the most critical events at the top of the review list can more effectively support their workflow and decision-making processes.

Ranking-based approaches explicitly learn the relative ordering of instances according to their importance or relevance, making them particularly suitable for this prioritization task. Instead of assigning each report to a discrete severity level, these methods estimate a continuous severity score and order events from most to least severe. By emphasizing relative priority rather than categorical accuracy, ranking-based models align more closely with the operational needs of

patient safety specialists. To our knowledge, no prior studies have systematically evaluated ranking approaches for prioritizing PSEs by severity.

4.2.1 Study Objectives

This study aimed to evaluate the effectiveness of machine learning models for prioritizing high-severity PSEs using retrospectively collected incident reports. We developed both ranking-based and classification-based models and compared their performance in prioritizing PSEs by severity. The top-performing model was further evaluated against the Reporter-Assigned Severity Baseline to assess its potential to improve severity triage practices. Finally, we designed an interactive interface to demonstrate how such models can be integrated into clinical workflows to support real-time triage.

4.3 Methodology

4.3.1 Study Design

This study aimed to develop and validate machine learning models for prioritizing high-severity PSEs. Figure 4-1 presents an overview of the study workflow, comprising three phases: 1) data extraction and partitioning, 2) model development and validation, and 3) interface design and conceptual demonstration.

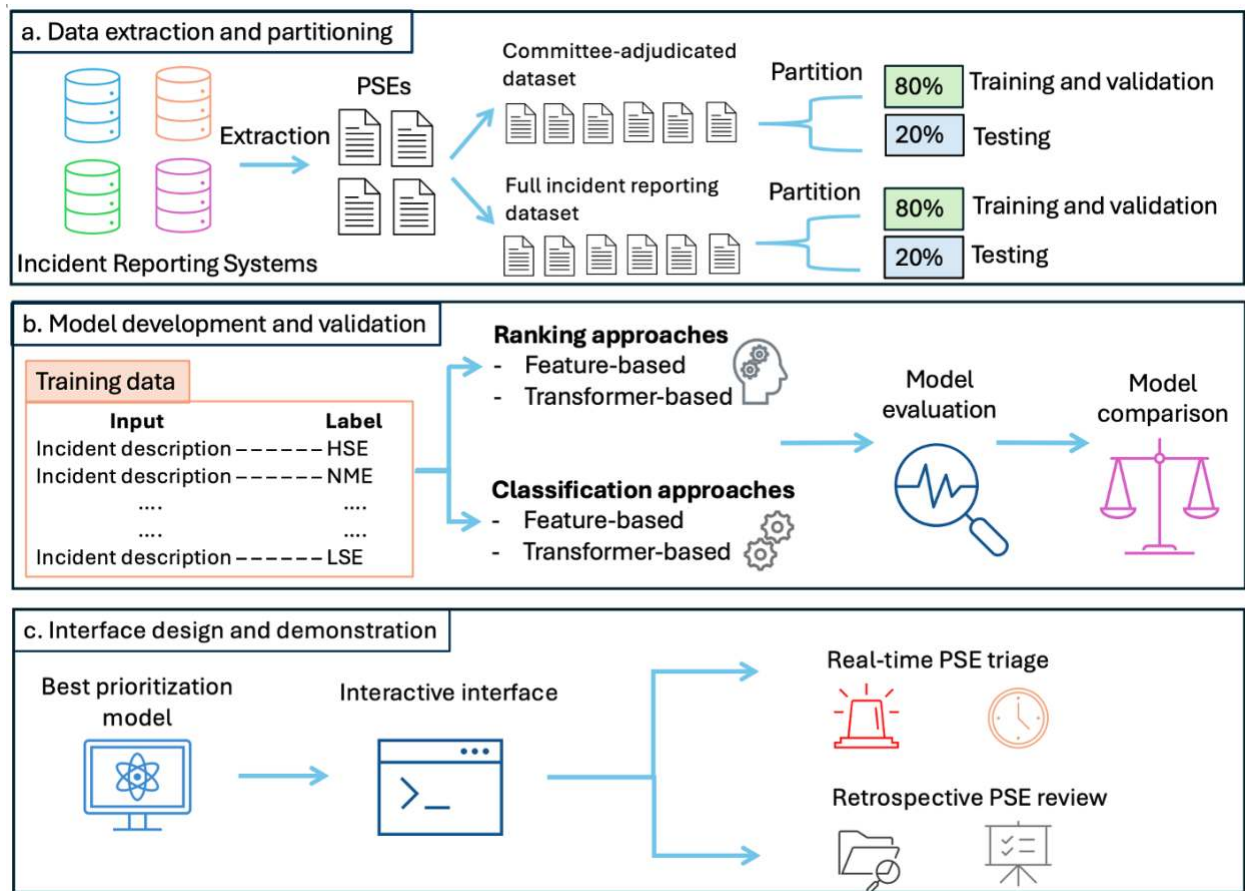


Figure 4-1 Overview of the study workflow, including data processing, model development, and interface design.

4.3.2 Dataset Description

This study utilized de-identified PSE reports collected from a large academic health system in Eastern Canada. A total of 101,246 reports were extracted from institutional IRSs between January 2017 and March 2025. Each PSE was assigned a severity label to indicate the level of harm associated with the event. The classification includes four categories: (1) High-Severity Event (HSE), (2) Low-Severity Event (LSE), (3) Near Miss Event (NME), and (4) No Safety-Concern Event (NSE). High-severity events refer to incidents associated with substantial patient harm, including life-threatening outcomes or death. Low-severity events include incidents with limited or no clinically significant harm. Near miss events are incidents that had the potential to cause harm but were prevented before reaching the patient. Non-safety events refer to reports that do not represent a patient safety issue or do not involve a deviation from standard care processes.

Severity labeling followed a two-stage institutional review process. First, patient safety specialists screened PSE reports submitted through the IRS to identify those with potential for high severity. Reports identified through this initial screening were subsequently reviewed by a multidisciplinary committee of point-of-care clinicians, patient safety specialists, operational leaders, and other domain experts. Based on this review, severity labels were assigned according to established institutional classification practices.

We composed two analytic datasets to support complementary but distinct evaluation objectives. Table 4-1 presents the distribution of PSE reports by severity category across the two datasets. The first dataset comprised PSE reports that underwent multidisciplinary committee review and received severity labels. These expert-adjudicated labels provide a high-quality reference standard and were used to support robust assessment and comparison of model performance. This dataset included 2,675 reports, of which 1,037 (38.8%) were classified as HSE (Table 4-1, Panel A). Throughout this study, the term patient safety event (PSE) refers to any reported safety-related incident that could have resulted, or did result, in unnecessary harm to a patient, whereas HSE, LSE, NME, and NSE denote the severity categories used in the institutional classification system.

The second dataset included all PSE reports submitted to the IRS, encompassing both committee-reviewed reports with severity labels (i.e., Entire Committee-adjudicated dataset) and reports that were not escalated for committee review and therefore did not receive severity labels. Under current institutional workflows, reports that are not escalated are implicitly treated as lower-priority events and are not subject to formal severity adjudication. For analytical purposes, we assume that these unreviewed reports represent non-high-severity events as operationally defined by existing triage practices. This assumption reflects institutional decision-making rather than asserting the absence of patient safety concerns. It enables deployment-relevant evaluation in which the model must prioritize a small number of high-severity PSEs for investigation from a large volume of incoming reports, mirroring real-world safety triage conditions. The second dataset comprised 101,239 reports, of which 1,037 (1.0%) were classified as HSEs, reflecting the low prevalence of high-severity cases in routine triage settings (Table 4-1, Panel B).

Table 4-1 Distribution of PSE reports by severity labels.

Panel A. Committee-adjudicated dataset

Severity label	Number of reports	Percentage (%)
HSE	1,037	38.77
LSE	830	31.03
NME	56	2.09
NSE	752	28.11
Total	2,675	100

Panel B. Full incident reporting system dataset

Severity label	Number of reports	Percentage (%)
HSE	1,037	1.00
LSE	830	0.82
NME	56	0.06
NSE	99,316	98.10
Total	101,239	100

Note: In Panel B, NSE includes both explicitly classified NSEs and reports not escalated for committee review.

4.3.3 Model Development and Validation

In this study, we developed machine learning models to prioritize high-severity PSEs using both ranking and classification approaches. We utilized the incident description section of each PSE report as input and the corresponding severity label as the outcome variable, consistent with prior work [30,34,67]. The incident description is a free-text narrative field in which the event reporter documents the circumstances and outcomes associated with the safety event.

Ranking-based models were developed using a learning-to-rank (LTR) framework, which learns to order items based on their relevance or importance and produces a ranked list [62,63]. In this context, LTR was applied to rank PSEs according to their severity, as indicated by the severity label. The goal was to ensure that high-severity PSEs appeared near the top of the ranked list. For classification-based models, we trained models using supervised learning to predict the severity label for each PSE based on its incident description.

We implemented both fixed-representation-based and transformer-based models for ranking and classification approaches. Fixed-representation-based models refer to traditional machine learning algorithms that leverages fixed representation (i.e., TF-IDF) and are computationally efficient and frequently serve as baselines in NLP tasks [121,122]. Transformer-based models are deep learning architectures that have achieved state-of-the-art performance in a variety of classification and ranking benchmarks [33,60,81]. Their effectiveness is attributed to their ability to capture contextual dependencies in text and leverage knowledge from large-scale pretraining.

We trained two fixed-representation-based models: XGBoost and LightGBM [121,123], both of which are gradient-boosted decision tree algorithms known for their effectiveness in prior ranking applications [62]. We trained six transformer-based models: BERT, ROBERTA, XLM-ROBERTA, BIOBERT, BIOMEDBERT, and GATORTRON [32,33,74,124,125]. BERT, ROBERTA, and XLM-ROBERTA were selected due to their widespread adoption and competitive performance across diverse NLP tasks [26,126,127]. BIOBERT, BIOMEDBERT, and GATORTRON were included for their pretraining on biomedical corpora, optimizing their performance for tasks involving clinical narratives and medical terminology [74,124,125]. Table 4-2 summarizes the models evaluated in this study along with their abbreviations; the suffix “RA” denotes ranking tasks, while “CL” indicates classification tasks. For fixed-representation-based models, free-text narratives were converted into structured feature representations before being used as input. A detailed description of the feature extraction methods are presented in Appendix 4.1 Feature Extraction Procedures for Fixed-Representation-Based Models.

Table 4-2 Summary of models used in this study, including their abbreviations and corresponding full model names. Models with the suffix “RA” were developed for ranking tasks, while those with the suffix “CL” were developed for classification tasks.

Task type	Abbreviations	Full model name
Ranking	XGB-RA	XGBoost
	LGBM-RA	LightGBM
	BERT-RA	bert-base-uncased
	ROBERTA-RA	roberta-base
	XLM-ROBERTA-RA	xlm-roberta-base
	BIOBERT-RA	biobert-v1.1
	BIOMEDBERT-RA	BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext
Classification	GATORTRON-RA	gatortron-base
	XGB-CL	XGBoost
	LGBM-CL	LightGBM
	BERT-CL	bert-base-uncased
	ROBERTA-CL	roberta-base
	XLM-ROBERTA-CL	xlm-roberta-base
	BIOBERT-CL	biobert-v1.1
	BIOMEDBERT-CL	BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext
	GATORTRON-CL	gatortron-base

The model development workflow is illustrated in Figure 4-2. The dataset was partitioned into 80% for training and validation and 20% for testing, with no overlap between splits. For fixed-representation-based models, hyperparameter tuning was conducted using grid search with five-fold cross-validation [100]. The training and validation data were divided into five mutually

exclusive folds; for each hyperparameter configuration, models were trained on four folds and validated on the fifth, with the process repeated across all five folds. The optimal configuration was selected based on the average cross-validated performance, measured by normalized discounted cumulative gain (NDCG). For transformer-based models, the dataset was partitioned into 70% for training, 10% for validation, and 20% for testing. To ensure comparability, the same testing set used for the fixed-representation-based models was retained. Models were trained on the training set, and hyperparameter tuning was conducted on the validation set. A conventional training–validation–test split was employed instead of cross-validation due to the substantial computational cost associated with training transformer-based models. A complete list of tuned hyperparameters for both model types is provided in Appendix 4.2 Hyperparameter Settings, Tuning Ranges, and Final Selected Configurations for Feature-Based and Transformer-Based Models.

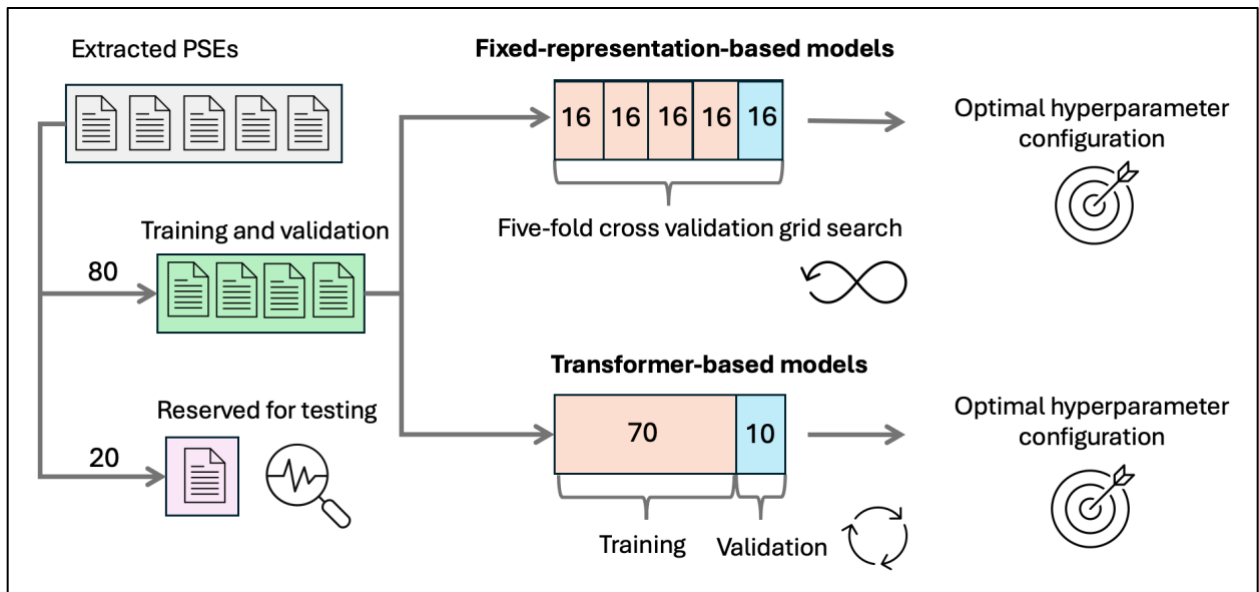


Figure 4-2 Model development workflow for fixed-representation-based and transformer-based models. Numbers indicate the proportion of the dataset allocated to each stage of model training, validation, and testing.

The distribution of severity labels in the dataset was imbalanced, with some classes underrepresented. The imbalance can negatively impact the performance of classification-based models, particularly for minority classes [128]. For fixed-representation-based classifiers, we applied the Synthetic Minority Oversampling Technique (SMOTE) to generate synthetic samples

for the underrepresented classes in the training set, thereby addressing class imbalance [129]. SMOTE was applied only to the training set to preserve the original class distribution in the test set and avoid bias in model evaluation. For transformer-based classifiers, we implemented a loss reweighting strategy by assigning higher weights to minority classes in the loss function [128]. This approach penalizes misclassifications of underrepresented classes more heavily, promoting more balanced performance across classes.

Within the full incident reporting system dataset, the distribution of severity labels is characterized by extreme class imbalance. To mitigate this imbalance during classifier training, we additionally explored random downsampling of the majority class (i.e., NSE). All non-NSE severity reports were retained, while the number of NSEs was varied across a predefined range to construct training sets with differing class imbalance ratios. The downsampling ratio was treated as a hyperparameter and selected based on validation performance. Downsampling was applied exclusively to the training data, and all validation and testing were conducted on datasets reflecting the original class distribution to preserve deployment-relevant prevalence. On the other hand, ranking-based models focus on learning the relative ordering of items rather than predicting discrete labels. These models are optimized based on comparisons between items, making them inherently less sensitive to class imbalance [130].

4.3.4 Model Evaluation and Comparison

Model performance was evaluated using established ranking metrics, including Average Precision, Normalized Discounted Cumulative Gain (NDCG) at multiple cutoff thresholds ($k=10, 25, 50, 100, 200$), as well as precision and recall computed at the same cutoffs. Average Precision quantifies the model's ability to rank high-severity PSEs near the top of the list and is calculated as the mean of precision values at each position where a high-severity PSE is retrieved. NDCG measures the quality of a ranked list by accounting for both the severity of the PSEs and their rank positions, offering a comprehensive evaluation of ranking effectiveness [62]. Precision@ k denotes the proportion of high-severity PSEs ranked in the top k positions, while recall@ k captures the proportion of all high-severity PSEs that appear within the top k . High-severity PSEs were defined as those assigned a severity label of HSE, which necessitates prioritization for institutional investigation. All evaluation metrics range from 0 to 1, with higher values indicating better ranking

performance. The mathematical definitions of all evaluation metrics are provided in Appendix 4.3 Mathematical Definitions of Evaluation Metrics.

For classification models, PSEs were ranked by their predicted probabilities of receiving a severity label and evaluated performance using ranking metrics. This approach more effectively captures the model's ability to prioritize high-severity PSEs than conventional classification metrics. For example, patient safety teams may triage the top 50–100 PSEs that are relatively more severe than others. In such contexts, ranking-based evaluation directly assesses the model's ability to order events by clinical urgency, thereby aligning more closely with clinical workflows.

We evaluated all trained models on the internal testing set, which comprised 20% of the total PSEs. The 95% confidence interval (CI) was computed using bootstrapping with 2,000 random subsamples, drawn with replacement and matched to the testing set's original size, to quantify uncertainty in model performance. We compared the performance of the best-performing ranking and classification using paired t-tests. In addition, the top-performing model was compared with a baseline representing the Reporter-Assigned Severity Baseline, in which safety events from the testing set were ranked according to the reporters' perceived severity. The ranked events were then evaluated against the ground truth severity label using the ranking metrics. To obtain a more robust estimate of baseline performance, the baseline rankings were bootstrapped using the same resampling procedure as the model evaluations. Multiple-comparison adjustment was performed using a Bonferroni correction, setting the significance threshold at $P < 0.0008$ ($0.05 / 64$) to maintain a family-wise error rate of 0.05 across 64 tests ($16 \text{ metrics} \times 4 \text{ comparisons}$).

4.3.5 Interface Design

In the typical incident reporting workflow, frontline healthcare staff are required to submit a narrative description of the adverse event, along with structured metadata such as event type, date, location, and the perceived level of harm. Following this initial submission, the patient safety specialist reviews the reports to determine whether the incident warrants immediate investigation based on its severity. Due to the high volume of reports and the urgency associated with severe events, timely and accurate triage is critical to ensure appropriate institutional response.

To support this process, we developed an interactive interface that integrates the machine learning model to support real-time severity triage. The interface displays a ranked list of PSE reports from

most severe to least severe. For each report, the interface shows its ranking percentile relative to all submitted reports, as well as the structured and unstructured information that normally exists in the report. The interface was developed using Gradio and guided by prior research on the design of incident reporting systems [45,104,105]. The designed interface is presented in Appendix 4.4 Demonstration of the Designed Severity Ranking Interface.

4.3.6 Ethical Considerations

This study conducted a secondary analysis of de-identified PSE data collected through institutional IRSs for quality and safety improvement. The study protocol was approved by the hospital's Research Ethics Board (REB# 23-5237) and was further reviewed to ensure compliance with jurisdictional privacy legislation and institutional quality assurance policies. Following extraction, all data were fully anonymized in accordance with applicable privacy regulations. Informed consent was not required, as the study involved secondary use of anonymized data. No compensation was applicable for the use of these data. Figures and supplementary materials in this manuscript do not contain any identifiable information regarding individual participants or users.

4.4 Results

4.4.1 Performance Comparison of Classification and Ranking Models

Among all ranking models, BERT-RA achieved the highest performance based on Average Precision, while BIOBERT-CL demonstrated the best performance among the classification models. Table 4-3 summarizes the comparative results between these two top-performing models. Average Precision was selected as the primary criterion for model selection, as it provides a comprehensive measure of each model's overall ranking capability by jointly accounting for the accuracy and ordering of high-severity PSEs across the ranked list. Detailed performance metrics for all developed models are presented in Appendix 4.5 Evaluation Results of All Developed Ranking and Classification Models.

Table 4-3 . Comparative performance of the best-performing classification model (BIOBERT-CL) and ranking model (BERT-RA) across two evaluation settings.

Metrics	BIOBERT-CL, mean (95% CI)	BERT-RA, mean (95% CI)	Relative improvement (%)	<i>P</i> value
Panel A. Committee-adjudicated dataset				

Average Precision	0.74±0.06	0.91±0.03	23.0	< 0.0008
NDCG@10	1.00±0.00	1.00±0.00	0.0	1
NDCG@20	1.00±0.00	1.00±0.00	0.0	1
NDCG@50	0.96±0.04	0.99±0.02	3.1	< 0.0008
NDCG@100	0.91±0.05	0.99±0.04	8.8	< 0.0008
NDCG@200	0.81±0.04	0.93±0.03	14.8	< 0.0008
Precision@10	1.00±0.00	1.00±0.00	0.0	1
Precision@20	1.00±0.00	1.00±0.00	0.0	1
Precision@50	0.92±0.07	0.98±0.05	6.5	< 0.0008
Precision@100	0.84±0.08	0.97±0.04	15.5	< 0.0008
Precision@200	0.66±0.07	0.83±0.07	25.8	< 0.0008
Recall@10	0.05±0.01	0.05±0.01	0.0	1
Recall@20	0.10±0.01	0.10±0.01	0.0	1
Recall@50	0.22±0.03	0.24±0.03	9.1	< 0.0008
Recall@100	0.41±0.04	0.47±0.04	14.6	< 0.0008
Recall@200	0.63±0.05	0.80±0.05	27.0	< 0.0008

Panel B. Full incident reporting system dataset

Average Precision	0.68±0.06	0.87±0.04	27.9	< 0.0008
NDCG@10	0.93±0.13	1.00±0.00	7.5	< 0.0008
NDCG@20	0.95±0.09	1.00±0.00	5.3	< 0.0008
NDCG@50	0.93±0.05	0.99±0.01	6.5	< 0.0008
NDCG@100	0.90±0.04	0.96±0.02	6.7	< 0.0008
NDCG@200	0.79±0.06	0.94±0.02	19.0	< 0.0008
Precision@10	0.90±0.18	1.00±0.00	11.1	< 0.0008
Precision@20	0.95±0.10	1.00±0.00	5.3	< 0.0008
Precision@50	0.89±0.06	0.97±0.01	9.0	< 0.0008
Precision@100	0.84±0.07	0.90±0.06	7.1	< 0.0008
Precision@200	0.61±0.09	0.78±0.07	27.9	< 0.0008
Recall@10	0.04±0.01	0.05±0.01	25.0	< 0.0008

Recall@20	0.09±0.01	0.10±0.01	11.1	< 0.0008
Recall@50	0.18±0.03	0.23±0.03	27.8	< 0.0008
Recall@100	0.37±0.05	0.43±0.05	16.2	< 0.0008
Recall@200	0.61±0.06	0.75±0.06	23.0	< 0.0008

Note: Panel A is the committee-adjudicated dataset, and Panel B is the full incident reporting system dataset reflecting real-world deployment conditions. Performance values are reported as mean (95% confidence interval) on the held-out testing set, estimated using 2,000 bootstrap resamples. Relative improvement (%) represents the proportional performance gain of the ranking model over the classification model. P values indicate statistical significance based on paired comparisons with Bonferroni correction.

Across all 16 evaluation metrics in Panel A (committee-adjudicated dataset), BERT-RA achieved equal or higher mean performance compared with BIOBERT-CL. BERT-RA demonstrated a significantly higher Average Precision (0.91 ± 0.03) than BIOBERT-CL (0.74 ± 0.06 ; $P < 0.0008$), corresponding to a 23.0% relative improvement. Both models achieved perfect performance at the most restrictive cutoffs ($\text{NDCG@10-20} = 1.00$; $\text{Precision@10-20} = 1.00$), with identical Recall@10 (0.05) and Recall@20 (0.10), indicating comparable ability to identify the most severe events within the top-ranked positions. Performance differences became more pronounced at broader cutoffs. At $k = 100$, BERT-RA improved NDCG by 8.8%, Precision by 15.5%, and Recall by 14.6%. At $k = 200$, relative improvements increased to 14.8% for NDCG, 25.8% for Precision, and 27.0% for Recall (all $P < 0.0008$). Overall, BERT-RA demonstrated statistically significant advantages in 10 of the 16 evaluated metrics, with the largest gains observed at higher cutoffs, reflecting more consistent prioritization of high-severity events across the ranked list. In Panel B (full incident reporting system dataset), the superiority of the ranking model remained consistent despite the substantially lower prevalence of high-severity events. BERT-RA achieved a 27.9% relative improvement in Average Precision (0.87 ± 0.04 vs. 0.68 ± 0.06 ; $P < 0.0008$). Similar to Panel A, gains widened at broader cutoffs, particularly at $k = 200$, where relative improvements reached 19.0% for NDCG, 27.9% for Precision, and 23.0% for Recall (all $P < 0.0008$). As a sensitivity analysis, we also compared the learning-to-rank formulation with a binary classification model trained to distinguish HSE from non-HSE reports directly. The methods and complete

results of this analysis are presented in Appendix 4.6 Sensitivity Analysis of Multiclass and Binary HSE Classification Formulations.

4.4.2 Comparison with the Reporter-Assigned Severity Baseline

Table 4-4 presents the performance comparison between the best-performing ranking model (BERT-RA) and the Reporter-Assigned Severity Baseline across both evaluation settings. In Panel A (committee-adjudicated dataset), BERT-RA consistently achieved significantly higher mean performance than the baseline across all 16 evaluation metrics (all $P < 0.0008$). Average Precision was substantially higher for BERT-RA (0.91 ± 0.03) compared with the baseline (0.45 ± 0.07), corresponding to a 102.2% relative improvement. Performance gains were evident even at the most restrictive cutoffs. For example, NDCG@10 and NDCG@20 improved by 31.6% and 37.0%, respectively, while Precision@10 and Precision@20 improved by 49.3% and 66.7%. At $k = 100$, BERT-RA demonstrated relative improvements of 59.7% in NDCG, 120.5% in Precision, and 123.8% in Recall. At $k = 200$, improvements remained substantial, reaching 52.5% in NDCG, 97.6% in Precision, and 95.1% in Recall. Notably, the ranking model retrieved 80% of high-severity PSEs within the top 200 positions, compared with 41% under the Reporter-Assigned Severity Baseline.

In Panel B (full incident reporting system dataset), which reflects real-world class imbalance and operational deployment conditions, the superiority of BERT-RA was even more pronounced. Average Precision improved from 0.27 ± 0.04 under the baseline to 0.87 ± 0.04 with BERT-RA, corresponding to a 222.2% relative improvement. Substantial gains were observed across all evaluation cutoffs. At $k = 100$, relative improvements reached 166.7% for NDCG, 221.3% for Precision, and 230.8% for Recall. At $k = 200$, BERT-RA achieved improvements of 184.8% in NDCG, 225.0% in Precision, and 257.1% in Recall. These results indicate that, under realistic deployment conditions with a low prevalence of HSEs, reliance on reporter-assigned severity alone performs markedly worse than the ranking-based approach, whereas BERT-RA maintains robust prioritization performance.

Table 4-4 Comparative performance of the best-performing ranking model (BERT-RA) and the Reporter-Assigned Severity Baseline across two evaluation settings.

Metrics	Reporter-Assigned Severity Baseline, mean (95% CI)	BERT-RA, mean (95% CI)	Relative improvement (95% CI)	<i>P</i> value
Panel A. Committee-adjudicated dataset				
Average Precision	0.45±0.07	0.91±0.03	102.2	< 0.0008
NDCG@10	0.76±0.19	1.00±0.00	31.6	< 0.0008
NDCG@20	0.73±0.17	1.00±0.00	37.0	< 0.0008
NDCG@50	0.65±0.11	0.99±0.02	52.3	< 0.0008
NDCG@100	0.62±0.08	0.99±0.04	59.7	< 0.0008
NDCG@200	0.61±0.06	0.93±0.03	52.5	< 0.0008
Precision@10	0.67±0.25	1.00±0.00	49.3	< 0.0008
Precision@20	0.60±0.23	1.00±0.00	66.7	< 0.0008
Precision@50	0.48±0.14	0.98±0.05	104.2	< 0.0008
Precision@100	0.44±0.10	0.97±0.04	120.5	< 0.0008
Precision@200	0.42±0.07	0.83±0.07	97.6	< 0.0008
Recall@10	0.03±0.01	0.05±0.01	66.7	< 0.0008
Recall@20	0.06±0.02	0.10±0.01	66.7	< 0.0008
Recall@50	0.12±0.03	0.24±0.03	100.0	< 0.0008
Recall@100	0.21±0.04	0.47±0.04	123.8	< 0.0008
Recall@200	0.41±0.05	0.80±0.05	95.1	< 0.0008
Panel B. Full incident reporting system dataset				
Average Precision	0.27±0.04	0.87±0.04	222.2	< 0.0008
NDCG@10	0.46±0.16	1.00±0.00	117.4	< 0.0008
NDCG@20	0.44±0.15	1.00±0.00	127.3	< 0.0008
NDCG@50	0.41±0.13	0.99±0.01	141.5	< 0.0008
NDCG@100	0.36±0.08	0.96±0.02	166.7	< 0.0008
NDCG@200	0.33±0.05	0.94±0.02	184.8	< 0.0008
Precision@10	0.45±0.31	1.00±0.00	122.2	< 0.0008
Precision@20	0.44±0.21	1.00±0.00	127.3	< 0.0008
Precision@50	0.38±0.13	0.97±0.01	155.3	< 0.0008

Precision@100	0.28±0.08	0.90±0.06	221.3	< 0.0008
Precision@200	0.24±0.06	0.78±0.07	225.0	< 0.0008
Recall@10	0.02±0.01	0.05±0.01	150.0	< 0.0008
Recall@20	0.04±0.02	0.10±0.01	150.0	< 0.0008
Recall@50	0.08±0.03	0.23±0.03	187.5	< 0.0008
Recall@100	0.13±0.04	0.43±0.05	230.8	< 0.0008
Recall@200	0.21±0.05	0.75±0.06	257.1	< 0.0008

Note: Panel A is the committee-adjudicated dataset, and Panel B is the full incident reporting system dataset reflecting real-world deployment conditions. Performance values are reported as mean (95% confidence interval) on the held-out testing set, estimated using 2,000 bootstrap resamples. Relative improvement (%) represents the proportional performance gain of BERT-RA over the Reporter-Assigned Severity Baseline. P values indicate statistical significance based on paired comparisons with Bonferroni correction.

4.5 Discussion

4.5.1 Overview

Adverse events remain a major threat to patient safety and impose substantial burdens on healthcare systems worldwide [6]. Although some jurisdictions have established legislation or formal frameworks addressing patient safety incident reporting [112–114], only a small proportion of PSEs undergo a comprehensive investigation because the review process is resource-intensive [34]. Patient safety specialists are therefore responsible for manually triaging reported events to identify high-severity cases that warrant institutional investigation [34,47,48]. However, the rapidly increasing volume of reports, while reflecting ongoing efforts to strengthen reporting and safety culture, has rendered manual triage increasingly challenging [120]. Previous studies have applied NLP techniques to facilitate this process; nevertheless, most have relied on classification-based approaches [30,34,67,131], which may not fully align with the operational needs of patient safety specialists who must efficiently prioritize the most critical events requiring investigation. In contrast, ranking-based approaches learn the relative ordering of safety events by severity, making them particularly well-suited for this prioritization task. By directly optimizing the prioritization

step within existing triage workflows, ranking-based models offer a practical and workflow-aligned advancement. To the best of our knowledge, this study is the first to develop and validate ranking-based models for prioritizing high-severity PSEs.

4.5.2 Principal Findings

The best-performing ranking model (BERT-RA) demonstrated a significant advantage over both classification models and the Reporter-Assigned Severity Baseline in prioritizing high-severity PSEs. Compared with the best-performing classification model (BIOBERT-CL), BERT-RA achieved a 23.0% relative improvement in Average Precision and significantly outperformed it in ten of sixteen evaluation metrics. Although both models achieved perfect performance within the top 20 ranked positions, BERT-RA showed progressively greater gains at broader cutoffs, with relative improvements of 14.8% in NDCG@200, 25.8% in Precision@200, and 27.0% in Recall@200. These findings indicate that while both models effectively identified the most severe cases, the ranking-based approach maintained superior prioritization as the evaluation range expanded. Importantly, similar patterns were observed in the full incident reporting system dataset (Panel B), which reflects real-world class imbalance and operational deployment conditions. Despite the substantially lower prevalence of high-severity PSEs, BERT-RA consistently outperformed the classification model across evaluation cutoffs. This consistency across both expert-adjudicated and deployment-representative datasets strengthens the external validity of the ranking-based approach and underscores its applicability in routine triage environments. Consequently, ranking-based approaches may provide a scalable solution for supporting triage workflows, particularly given the increasing volume of PSE reports in modern healthcare systems [29,93,132].

Compared with the Reporter-Assigned Severity Baseline, the BERT-RA model demonstrated significantly superior capability in prioritizing high-severity PSEs across all sixteen evaluation metrics. At the 200-event evaluation cutoff, BERT-RA retrieved 80.0% of all high-severity PSEs, whereas the Reporter-Assigned Severity Baseline identified only 41.0%, corresponding to a 95.1% relative improvement. This substantial gain highlights the model's ability to consistently detect critical events that may otherwise not be prioritized for timely review. The comparatively lower performance of the baseline likely reflects the inherent limitations of the Reporter-Assigned Severity Baseline, which is based on reporters' subjective severity ratings that tend to be

inconsistent due to variations in professional role, clinical experience, and familiarity with severity taxonomies [46,91]. Integrating ranking-based models such as BERT-RA into institutional triage workflows may therefore enhance the efficiency and accuracy of prioritizing high-severity PSEs, enabling more timely identification and mitigation of such events, facilitating earlier escalation to leadership for timely organizational response, and ultimately strengthening patient safety outcomes.

Although numerous studies have investigated the application of machine learning models to assist in the severity triage of safety events, few have proposed concrete designs or demonstrated how such technologies can be effectively embedded within clinical workflows [30,34,67,131]. Severity triage is an inherently high-stakes process, as decisions directly influence institutional response and allocation of resources for organizational improvements. Within this context, machine learning models should function as decision-support tools rather than autonomous decision-makers, and final judgment must remain a human decision to ensure ethical appropriateness [98,133,134]. To demonstrate the practical utility of the proposed ranking model, this study designed an interactive interface that orders reported safety events from most to least severe (Appendix-4). This interface enables patient safety specialists to strategically review and triage events according to available resources (i.e., investigative capacity), facilitating more effective prioritization of high-severity cases. Collectively, this model-assisted triage framework offers a practical pathway for integrating artificial intelligence into institutional safety review processes, with the potential to enhance efficiency, consistency, and responsiveness of institutional severity triage workflows.

4.5.3 Limitations

This study has several limitations. First, it was conducted using data from a single academic health system in Canada, which may limit the generalizability of the findings to other institutional or geographical contexts. Future multicenter studies involving more diverse patient populations and reporting systems are warranted to enhance external validity. Second, the proposed model was evaluated retrospectively, and its prospective performance in real-world clinical settings remains to be determined. Future research should examine the model's implementation feasibility and its influence on triage outcomes in operational environments. Third, this study relied exclusively on incident reporting system data and did not incorporate information from other relevant sources such as electronic medical records, maintenance and facility reports, or audit data, which patient

safety specialists often use to validate and contextualize reported events. Future research could examine how similar prioritization approaches might integrate these data sources to complement the proposed AI-assisted triage workflow. Future research could also explore the application of ranking models directly to EMR systems, which may enable earlier identification of high-risk events and reduce reliance on downstream incident reporting workflows. In addition, future studies could analyze system-level and longitudinal trends in safety events to enable further automation and provide deeper insight into the underlying contributors to adverse events. Fourth, this study did not examine patient preferences or perspectives on the use of artificial intelligence to support incident severity prioritization. Future research should explicitly address these dimensions, particularly in relation to ethical, governance, and patient-centered considerations. Finally, this study did not assess the potential utility of general-purpose large language models (e.g., ChatGPT) for prioritizing high-severity PSEs due to the sensitive nature of patient safety data, which cannot be transmitted via external APIs. Future investigations using secure, institutionally governed computing environments may help explore this direction. Future research should also prospectively evaluate the ranking system under realistic review-capacity constraints. Relevant measures should include the number and proportion of high-severity events identified within a fixed number of reviewed reports, the time required to identify and escalate serious cases, the number of high-severity events missed, and changes in reviewer workload. Comparisons between model-assisted ranking and the current review process would help determine whether the system enables specialists to identify serious events more efficiently while maintaining appropriate human oversight.

4.6 Conclusion

This study developed and validated ranking-based machine learning models to prioritize high-severity patient safety events (PSEs) from healthcare narratives. The proposed BERT-RA model consistently outperformed classification-based approaches and the Reporter-Assigned Severity baseline, demonstrating that ranking-based methods can more effectively identify critical events requiring institutional investigation. These findings highlight the advantage of modeling prioritization directly, rather than relying on categorical severity classification, for supporting real-world patient safety triage. The integration of an interactive interface further illustrates how such models can be operationalized within clinical workflows.

Within the broader context of this dissertation, this work demonstrates that machine learning can enable scalable and systematic prioritization of patient safety risks from narrative data. By transforming unstructured incident reports into actionable rankings, the proposed framework contributes to more efficient resource allocation and timely identification of high-risk events, thereby strengthening institutional patient safety monitoring.

Chapter 5

Data-Efficient Detection of Stigmatizing Language via In-Context Learning

5 Efficient Detection of Stigmatizing Language in Electronic Health Records via In-Context Learning: Comparative Analysis and Validation Study

5.1 Abstract

Background: Stigmatizing language in electronic health records (EHRs) constitutes a patient safety risk by perpetuating bias, shaping clinical decision-making, and negatively affecting patients' psychological well-being. As clinical documentation plays a central role in communication and decision-making within healthcare systems, the ability to systematically identify such language is

important for mitigating potential harm. While prior studies have explored supervised machine learning approaches for detecting stigmatizing language, these methods typically require large annotated datasets, which may not always be readily available. In-context learning (ICL) has emerged as a data-efficient alternative, allowing large language models to adapt to tasks using only instructions and examples.

Objective: We aimed to investigate the efficacy of ICL in detecting stigmatizing language within EHRs under data-scarce conditions.

Methods: We analyzed 5043 sentences from the Medical Information Mart for Intensive Care–IV dataset, which contains EHRs from patients admitted to the emergency department at the Beth Israel Deaconess Medical Center. We compared ICL with zero-shot (textual entailment), few-shot (SetFit), and supervised fine-tuning approaches. The ICL approach used 4 prompting strategies: generic, chain of thought, clue and reasoning prompting, and a newly introduced stigma detection guided prompt. Model fairness was evaluated using the equal performance criterion, measuring true positive rate, false positive rate, and *F1*-score disparities across protected attributes, including sex, age, and race.

Results: In the zero-shot setting, the best-performing ICL model, GEMMA-2, achieved a mean *F1*-score of 0.858 (95% CI 0.854-0.862), showing an 18.7% improvement over the best textual entailment model, DEBERTA-M (mean *F1*-score 0.723, 95% CI 0.718-0.728; $P < .001$). In the few-shot setting, the top ICL model, LLAMA-3, outperformed the leading SetFit models by 21.2%, 21.4%, and 12.3% with 4, 8, and 16 annotations per class, respectively ($P < .001$). Using 32 labeled instances, the best ICL model achieved a mean *F1*-score of 0.901 (95% CI 0.895-0.907), only 3.2% lower than the best supervised fine-tuning model, ROBERTA (mean *F1*-score 0.931, 95% CI 0.924-0.938), which was trained on 3543 labeled instances. Under the conditions tested, fairness evaluation revealed that supervised fine-tuning models exhibited greater bias compared with ICL models in the zero-shot, 4-shot, 8-shot, and 16-shot settings, as measured by true positive rate, false positive rate, and *F1*-score disparities.

Conclusions: ICL offers a robust and flexible solution for detecting stigmatizing language in EHRs, offering a more data-efficient and equitable alternative to conventional machine learning methods. These findings suggest that ICL could enhance bias detection in clinical documentation while reducing the reliance on extensive labeled datasets.

5.2 Introduction

Patient safety concerns the prevention of both physical and non-physical harm arising from healthcare delivery. Stigmatizing language in electronic health records (EHRs) represents a form of non-physical, preventable harm, as it can perpetuate bias, influence clinical decision-making, and adversely affect patient well-being [16,18,24,36]. Given the widespread adoption of EHR systems and their central role in shaping the interpretation and communication of patient information [135–138], empirical studies have documented stigmatizing language within these records [25,36,43,85]. A recent large-scale study reported that stigmatizing language appears in a substantial proportion of clinical notes, with prevalence ranging from 42.8% to 59.1% in history and physical notes, consultation notes, and discharge summaries [43].

Stigmatizing language in healthcare refers to language that carries negative connotations, labels patients negatively, attributes blame, or renders judgment based on their social identities, medical

conditions, or personal experiences [22]. Such language can be explicit, as in the use of derogatory terms like “junkie” or “alcoholic” [139]. However, it also frequently manifests in more implicit forms, including subtle judgements embedded in tone, phrasing, or contextual framing [19]. Research in social psychology has demonstrated that implicit biases, which refer to unconscious attitudes or stereotypes, can be reflected through people’s language [140,141]. In healthcare, implicit biases may surface through the language used to describe patients. For instance, labeling a patient as “non-compliant” may seem clinically objective, but it can implicitly stigmatize their behavior by suggesting poor self-management, rooted in preconceived notions or stereotypes, without considering contextual factors such as mental capacity, available resources, or external circumstances. Implicit biases reflect broader systemic inequities, as institutional practices and cultural norms often perpetuate and normalize the use of stigmatizing language in clinical settings [36,142]. Stigmatizing language has been frequently observed in clinical specialties such as substance use disorders, mental health conditions, diabetes, and obesity, where it has shaped both clinical terminology and professional attitudes [42,139,143,144]. This systemic perpetuation of stigmatizing language reinforces negative stereotypes and compromises care quality, thereby constituting a patient safety risk by undermining patient trust and adversely affecting treatment outcomes. By embedding biased language into clinical notes, these unconscious attitudes influence immediate clinical decisions and may perpetuate negative stereotypes over time [23,36], as subsequent clinicians rely on prior documentation [145]. These patterns were confirmed in a recent analysis of 754 clinical notes containing stigmatizing language, which found that stereotyping accounted for 41.2% of instances, followed by labeling patients as difficult (15.3%), disapproval (9.6%), unilateral decision-making (7.3%), and questioning patient credibility (5.3%) [146]. Such findings underscore the importance of mitigating bias to prevent patient harm and improve patient safety.

The presence of stigmatizing language constitutes a patient safety concern with direct implications for clinical decision-making and care delivery. A study found that clinicians exposed to the stigmatizing term “substance abuser”, as opposed to the more neutral phrase “person with substance use disorder”, were more likely to view the patient as personally culpable and deserving of punitive action rather than treatment [42]. Additionally, a national survey of 655 emergency physicians, those who used the stigmatizing label “Sickler” to describe patients with sickle cell disease, demonstrated lower compliance with national treatment guidelines and were less likely to

prescribe appropriate medications [144]. Complementing these findings, prior research has also shown that physicians who read notes containing stigmatizing language are more likely to develop negative attitudes toward the patient and to provide less aggressive pain management [24]. Notably, only 18% of inpatient progress notes are manually entered, with the majority being imported or copied from prior documentation [145]. The reliance on prior documentation raises concerns about the potential perpetuation of stigmatizing language, which could reinforce negative biases and influence clinical decision-making over time. With the implementation of the 21st Century Cures Act in 2021 in the United States, patients have gained access to their EHRs [147], making these records increasingly integral to the clinician-patient relationship. A large survey study found that 10% of patients feel judged or offended due to labeling and evidence of disrespect in EHRs [148]. Another study reported that 19% of patients indicated they would avoid future medical appointments if their physicians used stigmatizing language toward them, while 21% stated they would seek a new doctor [143]. There is also substantial evidence that stigmatizing language in EHRs can jeopardize the therapeutic relationship, leading to distrust, delayed care, reduced treatment engagement, and increased patient safety risks [19–21].

Given the detrimental effects of stigmatizing language in EHRs on patient safety and the substantial resources required for its manual identification, there is a growing interest in developing automated detection methods. Manual detection typically necessitates annotators possessing extensive clinical expertise to accurately interpret context, as well as linguistic proficiency to discern subtle indicators of stigma [25,36]. In current practice, manual annotation typically involves experienced physicians reading and classifying individual sentences or notes from EHRs, an approach that is labor-intensive and costly, particularly when scaled across large datasets [25,36,37]. To overcome these challenges, researchers have increasingly explored machine learning approaches for the automated detection of stigmatizing language within EHRs. For instance, Sun et al. developed a logistic regression classifier to assess whether sentences in EHRs convey stigmatization, utilizing a proprietary dataset [36], while Harrigian et al. conducted a comparative analysis of logistic regression and BERT classifiers for the detection of stigmatizing language [25]. These studies, however, have all utilized supervised learning approaches, which generally require thousands to tens of thousands of annotated data points to achieve optimal performance. The assumption that such large volumes of annotated data are readily available is often unrealistic, as the acquisition of labels for stigmatizing language detection presents

significant challenges. Stigmatizing language is frequently subtle, covert, and highly contextual. Consequently, acquiring annotated data for stigmatizing language detection can be demanding and resource-intensive. The previous studies' heavy reliance on supervised machine learning approaches has highlighted a significant limitation in addressing scenarios where annotated data may be scarce.

In-context learning (ICL) is an emerging paradigm for addressing natural language processing (NLP) tasks in data-scarce settings [65,149–151]. ICL enables general-purpose large language models (LLMs) to adapt to new tasks by incorporating instructions and either zero or a few examples directly within the input prompt, thereby reducing reliance on extensively annotated datasets [76,79,149]. By leveraging the robust contextual understanding capabilities inherent in general-purpose LLMs, the ICL approach is particularly well-suited for detecting stigmatizing language, a task that requires the identification of subtle nuances and a high level of contextual comprehension. To the best of our knowledge, no study has systematically examined the efficacy of the ICL approach for stigmatizing language detection, underscoring a significant gap in the current research landscape.

The effectiveness of the ICL approach is contingent upon prompt engineering, which involves crafting precise, informative, and contextually relevant prompts to guide the LLM in executing specific tasks [149,152,153]. Therefore, the efficacy of the ICL approach could be enhanced by a tailored prompting strategy for stigmatizing language detection. Previous studies on detecting stigmatizing language typically involve a two-stage process (Figure 5-1): first, the extraction of sentences containing potentially stigmatizing words or phrases in accordance with medical language guidelines, followed by the development of the supervised machine learning model with the annotated dataset [25,36]. The potentially stigmatizing words or phrases identified during the extraction process can provide valuable information that can be incorporated into the prompting strategy. By explicitly prompting the LLM to focus on these terms, we anticipate an improvement in its sensitivity to detecting stigmatizing language. Furthermore, given that stigmatizing language is often covert and subtle [25,154], incorporating common linguistic characteristics associated with such language may further increase the model's accuracy in identifying instances of stigmatization. Customizing prompts to address the unique aspects of stigmatizing language presents a valuable opportunity to enhance detection precision.

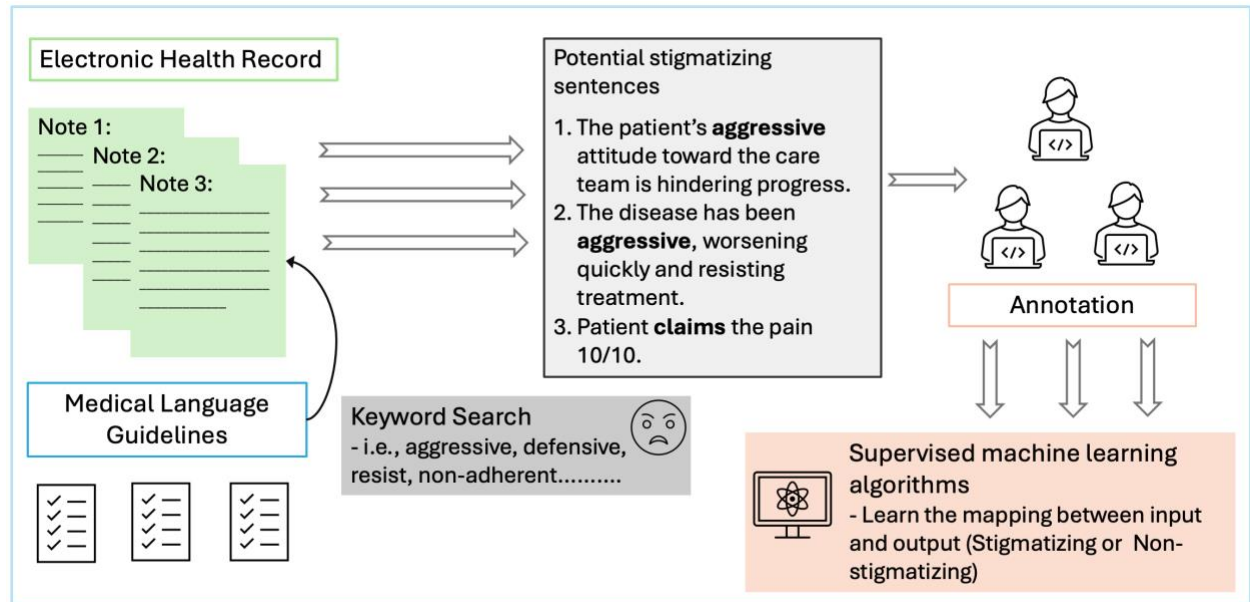


Figure 5-1 Workflow for developing a supervised machine learning model in stigmatizing language detection.

While developing effective machine learning models to detect stigmatizing language is important, it is equally critical to ensure these models do not introduce or perpetuate existing biases [133,155]. Numerous studies have demonstrated that machine learning classifiers can exhibit algorithmic bias, resulting in performance disparities across groups of patients with different protected attributes, such as sex, race, age, and socioeconomic status [133,155]. For example, Seyyed-Kalantari et al. found that a leading machine learning classifier for detecting chest X-ray pathologies exhibited a higher false negative rate for Hispanic female patients [156]. In the context of stigmatizing language detection, algorithmic bias can lead to uneven identification and correction of stigmatizing language, disproportionately affecting certain groups and potentially exacerbating health disparities. To our knowledge, no prior studies have specifically examined the fairness of classifiers in detecting stigmatizing language within EHRs, highlighting the need for further investigation into how these models perform across diverse patient populations. By rigorously evaluating the fairness of these models, we can identify potential disparities and ensure they equitably benefit patients from diverse backgrounds.

5.2.1 Study Objectives

This study investigates the effectiveness of the ICL approach for detecting stigmatizing language in EHRs under data-scarce conditions. First, we evaluated the performance of the ICL approach against established approaches in zero-shot (no annotated data) and few-shot settings (a small amount of annotated data), including the textual entailment [77] and SetFit [78]. For the ICL approach, we employed four distinct prompting strategies, including more sophisticated strategies such as the Chain of Thought (COT) [157] and the Clue and Reasoning Prompting (CARP) [150]. We also introduced a novel prompting strategy, termed ‘Stigma Detection Guided Prompt’, specifically designed to enhance the precision and effectiveness of detecting stigmatizing language. To further contextualize ICL’s performance, we conducted a comparative analysis of supervised fine-tuning and ICL to evaluate the extent to which ICL can achieve competitive performance under data-scarce conditions relative to a well-established method that leverages a large corpus of annotated data. This comparison provides valuable insights for hospitals, enabling them to assess the cost-benefit trade-offs associated with investing in annotated datasets for detecting stigmatizing language in clinical documentation. Furthermore, we assessed the fairness dimension of the classifiers used for detecting stigmatizing language in this study by comparing their performance across diverse patient groups characterized by various protected attributes, including sex, age, and race.

5.3 Methods

5.3.1 Dataset Description

In this study, we utilized an open-source stigmatization detection dataset prepared by Harrigian et al. [25]. The dataset consists of 5,043 sentences, of which 3,249 (64.4%) are labelled as stigmatizing and 1,794 (35.6%) as non-stigmatizing. The original creators randomly sampled these sentences from 4,710 discharge summaries associated with 4,259 patients within the MIMIC-IV database and subsequently annotated them. The MIMIC-IV is an EHR dataset that comprises 331,794 discharge summaries from Beth Israel Deaconess Medical Center in Boston, Massachusetts, United States, from 2008 to 2019 [158]. The observed prevalence of stigmatizing language in this dataset is notably higher than previously reported estimates, such as the 42.8% identified by Weiner et al. [43]. This difference is attributable to the broader scope of stigmatizing

expressions captured in the Harrigian et al. dataset, which encompasses a wider range of terminology than earlier studies.

Discharge summaries were selected because they provide detailed documentation of the patient's clinical course, including diagnoses, treatments, and physician observations, which are particularly relevant for investigating stigmatization in clinical notes. Sentences were extracted based on the presence of potentially stigmatizing words or expressions, such as “non-compliant” and “agitated.” A symmetric context window of 15 tokens on either side of each sentence was applied to preserve surrounding context and enhance model interpretation. If fewer than 15 tokens were available on either side, the context window was shortened as needed to fit the available text. Each sentence was independently annotated by two members of an annotation team comprising one clinician and two research assistants based on a predefined framework [25]. Any disagreements were resolved through discussion among all three annotators. The annotation process achieved a high level of inter-rater agreement, with an average pairwise Cohen's kappa exceeding 0.9. The annotation of stigmatization was determined based on framing and intent of the sentence with its surrounding context (i.e., whether the sentence implies that the patient's behavior is cast in a negative light). Sentences that did not describe patient behavior, or that neutrally described patient behavior, acknowledged structural or clinical barriers, or conveyed provider empathy were not considered stigmatizing. The demographic information of the dataset, including patient sex, age, and race, is presented in Table 5-1. A more detailed description of the dataset is provided in Appendix 5.1 Dataset Description and Stigmatizing Language Examples.

Table 5-1 Demographic information of the stigmatizing dataset used in this study.

Demographic Attributes	# of stigmatizing sentences	# of non-stigmatizing sentences	Total
Male	1731	869	2600
Female	1518	925	2443
0-25	303	165	468
26-50	1184	455	1639

51-75	1249	756	2005
75 plus	513	418	931
White	2040	1165	3205
Black	676	269	945
Hispanic	159	87	246
Asian	69	56	125
Other/Unknown	305	217	522

5.3.2 Comparative Analysis for Stigmatizing Language Detection

Zero-shot and few-shot learning are machine learning paradigms that allow models to perform tasks with either no labeled data or with only a minimal set of labeled examples, respectively [75,159]. Zero-shot learning enables a model to perform a task without any prior examples, relying solely on its pre-trained knowledge. In contrast, few-shot learning operates by providing the model with a limited number of labeled examples, typically ranging from 10 to 100.

In this study, we conducted a comprehensive evaluation of the effectiveness of the ICL approach for detecting stigmatizing language in EHRs under conditions of limited data availability. We compared the performance of ICL against the textual entailment approach [77] within a zero-shot learning context and the SetFit approach [78] within a few-shot learning context, specifically analyzing scenarios involving 4, 8, and 16 annotated examples per class. For the few-shot learning experiments, data points were randomly selected as examples. To ensure reproducibility and mitigate potential variability in performance due to random sampling, the selection process was repeated five times using distinct random seeds. The final performance results reported in this study represent the average across these five runs, providing a more robust and stable assessment of model performance.

Additionally, we extended our comparative analysis to include a fully supervised fine-tuning approach. This broader comparison allows us to rigorously assess the strengths and potential limitations of the ICL approach in the detection of stigmatizing language within EHRs. All approaches employed in this study adhered to the framework established in prior studies (Figure 5-1), wherein sentences were extracted based on potentially stigmatizing words or phrases identified from medical language guidelines, followed by the development of detection approaches (i.e., ICL).

5.3.3 Textual Entailment Approach

The textual entailment approach conceptualizes zero-shot text classification as an entailment task [77]. This method is among the most widely adopted for zero-shot text classification and has demonstrated state-of-the-art performance across various benchmarks [160,161]. In this approach, the input text is interpreted as a premise, and the model assesses whether this premise logically entails the corresponding class label. Specifically, in the context of stigmatizing language detection, the model first considers a sentence from an EHR as the premise and then evaluates whether it entails stigmatization or non-stigmatization. The textual entailment approach relies on the knowledge embedded in pretrained language models, which have been trained on natural language inference datasets [162]. This training effectively enhances the model's capability for sentence comprehension. In our study, we employed a selection of widely recognized pretrained language models for natural language inference to assess the effectiveness of the textual entailment approach in detecting stigmatizing language (Table 5-2). These models were selected due to their widespread use in zero-shot classification tasks [161,163,164] and their representation of a diverse range of transformer architectures, enabling us to fully explore the potential of the textual entailment approach for detecting stigmatizing language.

Table 5-2 Pretrained language models used in the textual entailment approach.

Abbreviation	Full model name
ROBERTA-M	roberta-large-mnli
BART-M	bart-large-mnli
DEBERTA-M	deberta-large-mnli

BERT-M

bert-base-multilingual-cased-multinli_tr

5.3.4 SetFit Approach

SetFit is a popular few-shot text classification approach that has been applied across various domains [161,165,166]. The SetFit approach consists of a two-step process [78]. First, a pretrained sentence transformer is fine-tuned in a contrastive, Siamese manner using pairs of labeled data. The objective of this contrastive fine-tuning is to ensure that embeddings of semantically similar sentences are close together in the feature space while embeddings of semantically dissimilar sentence pairs are maximally distanced. This process aims to generate meaningful representations that accurately capture the semantic relationships between sentences. In the second step, a classification head is trained using the representations produced by the fine-tuned sentence transformer. For this study, we employed six pretrained sentence transformers to evaluate the effectiveness of the SetFit approach in detecting stigmatizing language (Table 5-3). MPNET-V2 and MINILM-V2 were the sentence transformers used in the original work that proposed SetFit [78], while the more contemporary e5 and ROBERTA-V2 were included due to their superior performance on various NLP benchmarks [167,168]. We also incorporated two sentence transformers pretrained on medical domain knowledge: BERT-BIO-ST and BIOLORD. The inclusion of these models was based on the hypothesis that domain-specific pretraining may enhance the model's ability to capture contextually relevant medical terminology and nuances, thereby improving its effectiveness in detecting stigmatizing language [125].

Table 5-3 Pretrained sentence transformers used in the SetFit approach.

Abbreviation	Full model name
MPNET-V2	paraphrase-mpnet-base-v2
E5-V2	e5-base-v2
ROBERTA-V2	stsb-roberta-base-v2
MINILM-V2	paraphrase-multilingual-MiniLM-L12-v2

BERT-BIO-ST	S-BioBert-snli-multinli-stsb
BIOLORD	BioLORD-2023

5.3.5 Fully Supervised Fine-tuning Approach

The fully supervised fine-tuning approach involves adapting a pretrained language model to a specific task by fine-tuning it on a labeled dataset, where each instance is associated with its corresponding label [149]. In this study, we employed five pretrained language models to examine the efficacy of a fully supervised fine-tuning approach for detecting stigmatizing language (Table 5-4). ROBERTA, ROBERTA-XLM, and BERT were selected due to their extensive adoption and strong performance across text classification tasks in different domains [45,100,161,169]. BERT-BIO and BERT-biomed were included for their pretraining on biomedical text corpora, making them particularly well-suited for NLP tasks involving medical terminology [124,125]. We excluded certain contemporary medical pretrained language models, such as Gatortron [42] and ClinicalBERT [43], from our analysis. These models were omitted because they were pretrained on the MIMIC dataset, which could introduce data leakage and potentially lead to an overestimation of the model's effectiveness. We specifically focused on fine-tuning encoder-based language models, such as BERT, rather than decoder-based models (i.e., LLAMA) for two primary reasons. First, prior research has demonstrated that encoder-only models typically achieve superior or comparable performance to decoder-based models in text classification tasks with relatively small labeled datasets, which aligns with our study setting [170,171]. Second, BERT-based models currently represent the state-of-the-art for supervised stigmatizing language detection [25]. Thus, using these models as a benchmark allows for a more rigorous evaluation of ICL performance against established methodologies.

We fine-tuned these pretrained language models on 3,543 labeled sentences from EHRs, representing approximately 70% of the total data. The supervised fine-tuning approach was included to provide a robust benchmark for comparison with the ICL approach in detecting stigmatizing language. Hyperparameters were tuned using a validation dataset of 500 data points, representing approximately 10% of the total dataset. Hyperparameter tuning was conducted for key parameters using grid-search, including learning rate, batch size, and weight decay rate, to

optimize model performance and mitigate overfitting. We employed the AdamW optimizer [172] with default betas set to 0.9 and 0.999 and an epsilon of 1e-06 to ensure numerical stability. The number of epochs was determined using an early stopping criterion, where training was terminated if the validation loss did not decrease for five consecutive epochs. The final hyperparameter settings were selected based on the model’s best performance on the validation dataset before being evaluated on the held-out test dataset.

Table 5-4 Pretrained language models used in the fully supervised fine-tuning approach.

Abbreviation	Full model name
ROBERTA	roberta-base
BERT	bert-base-uncased
BERT-BIO	biobert-v1.1
ROBERTA-XLM	xlm-roberta-base
BERT-BIOMED	BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext

5.3.6 In-Context Learning Approach

The ICL approach leverages general-purpose LLMs through prompting techniques to rapidly adapt to a wide range of NLP tasks [149]. The approach has demonstrated exceptional performance in applications such as question answering, machine translation, text summarization, and text classification [65,76,79,149]. In this study, we employed five widely used open-source general-purpose LLMs to evaluate the efficacy of the ICL approach in detecting stigmatizing language (Table 5-5). We refer to models like LLaMA-3 as general-purpose LLMs to emphasize their broad applicability across various NLP tasks without extensive task-specific fine-tuning. We specifically selected instruction fine-tuned models, as previous research has shown that instruction fine-tuning significantly enhances model performance across a range of NLP tasks, particularly in zero-shot and few-shot settings [173,174]. Additionally, we included BIO-LLAMA-3, a domain-specific model that has undergone instruction fine-tuning on a large biomedical dataset [175], to assess the potential benefits of domain adaptation in stigmatizing language detection. Due to the constraints

of the MIMIC-IV data use agreement, we were unable to utilize newer cloud-based models, such as GPT-4 or Claude 3.5, as their use would necessitate sharing data with external parties. Previous studies have demonstrated that general-purpose LLMs are highly sensitive to subtle variations in prompt formatting [88,176]. To optimize the performance of the ICL approach, we devised a validation dataset for prompt format tuning, which included adjustments such as the wording and positioning of instructions. A random sample of 30 data points was selected to construct this validation set, designed to replicate a data-scarce scenario consistent with prior research [177,178].

Table 5-5 General-purpose LLMs used in the ICL approach.

Abbreviation	Full model name	Context window length (tokens)
LLAMA-3	Meta-Llama-3-8B-Instruct	8000
FLAN-T5	flan-t5-large	512
GEMMA-2	gemma-2-9b-it	8192
MISTRAL-0.2	Mistral-7B-Instruct-v0.2	32000
BIO-LLAMA-3	Bio-Medical-Llama-3-8B	8000

5.3.7 Prompting Strategies for the ICL Approach

The effectiveness of the ICL approach relies heavily on the design of a well-structured prompting strategy that delivers clear and precise task instructions. A prompting strategy refers to the specific approach used to organize instructions, examples, and guiding principles within a prompt to elicit accurate and reliable outputs from the LLM. These prompts are often accompanied by either zero or a few illustrative examples to guide the model in performing downstream tasks effectively [65,149]. In this study, we evaluated four distinct prompting strategies: Generic, Chain of Thought (COT) [157], Clue and Reasoning Prompting (CARP) [150], and a new customized prompting strategy we developed, referred to as the Stigma Detection Guided Prompt. The prompting template for each strategy is provided in Table 5-6. Figure 5-2 provides examples of the prompting strategies utilized in this study.

The generic prompting strategy involves providing the model with minimal context, thereby relying primarily on its pre-trained knowledge to perform the task [88]. In this approach, the prompt consists of a brief instruction about the task at hand, without incorporating reasoning steps or specific domain knowledge. For example, in the context of detecting stigmatizing language, the prompt might simply query the model with: "Choose your answer: Based on the above sentence, does the text convey stigmatization? Yes/No.". This strategy serves as a baseline for comparing with the performance of more advanced prompting techniques, such as COT, CARP, and the Stigma Detection Guided Prompt. In the few-shot scenario, a limited number of example sentences accompanied by their corresponding labels are embedded within the generic prompt to guide the model in making informed predictions.

The COT prompting strategy is designed to enhance the reasoning capabilities of general-purpose LLMs by guiding them to generate intermediate reasoning steps before arriving at a final prediction [157]. This approach leverages the LLM's capacity to systematically reason through complex tasks, thereby improving the reliability and accuracy of its outputs. For instance, the prompt can be appended with "Let's think step by step" to induce the model to break down the problem into logical and incremental steps. By explicitly instructing the model to articulate its reasoning, the COT strategy encourages more thoughtful predictions. In the zero-shot setting, the COT strategy involves prompting the model twice before making a prediction [76]. The first prompt encourages the model to produce intermediate reasoning by appending "Let's think step by step" to the end of the prompt. This intermediate reasoning is then incorporated into a second prompt, which is used to generate the final prediction. In the few-shot setting, the model was first prompted to generate reasoning steps for randomly selected data points. These data points, along with their corresponding reasoning steps, were incorporated into the prompt as examples to enhance the model's performance.

The CARP prompting strategy is specifically designed for text classification tasks. It utilizes a progressive reasoning approach to handle the linguistic complexities inherent in such tasks. This strategy first prompts the model to identify superficial clues within the input text, such as keywords, tone, semantic relations, and references, which provide initial indicators for classification. Once these clues are identified, the model engages in a diagnostic reasoning process to infer the final prediction in a logical manner [150]. For example, the presence of a negative tone combined with specific keywords may collectively suggest the presence of stigmatizing language.

In the few-shot setting, the model was first guided to generate clues and reasoning for randomly selected examples. These examples, along with their associated clues, reasoning, and labels, were then incorporated into the prompt to enhance the model’s performance [150]. The detailed code implementation for the prompting strategies is provided in Appendix 5.2 Detailed Code Implementation for the Prompting Strategies.

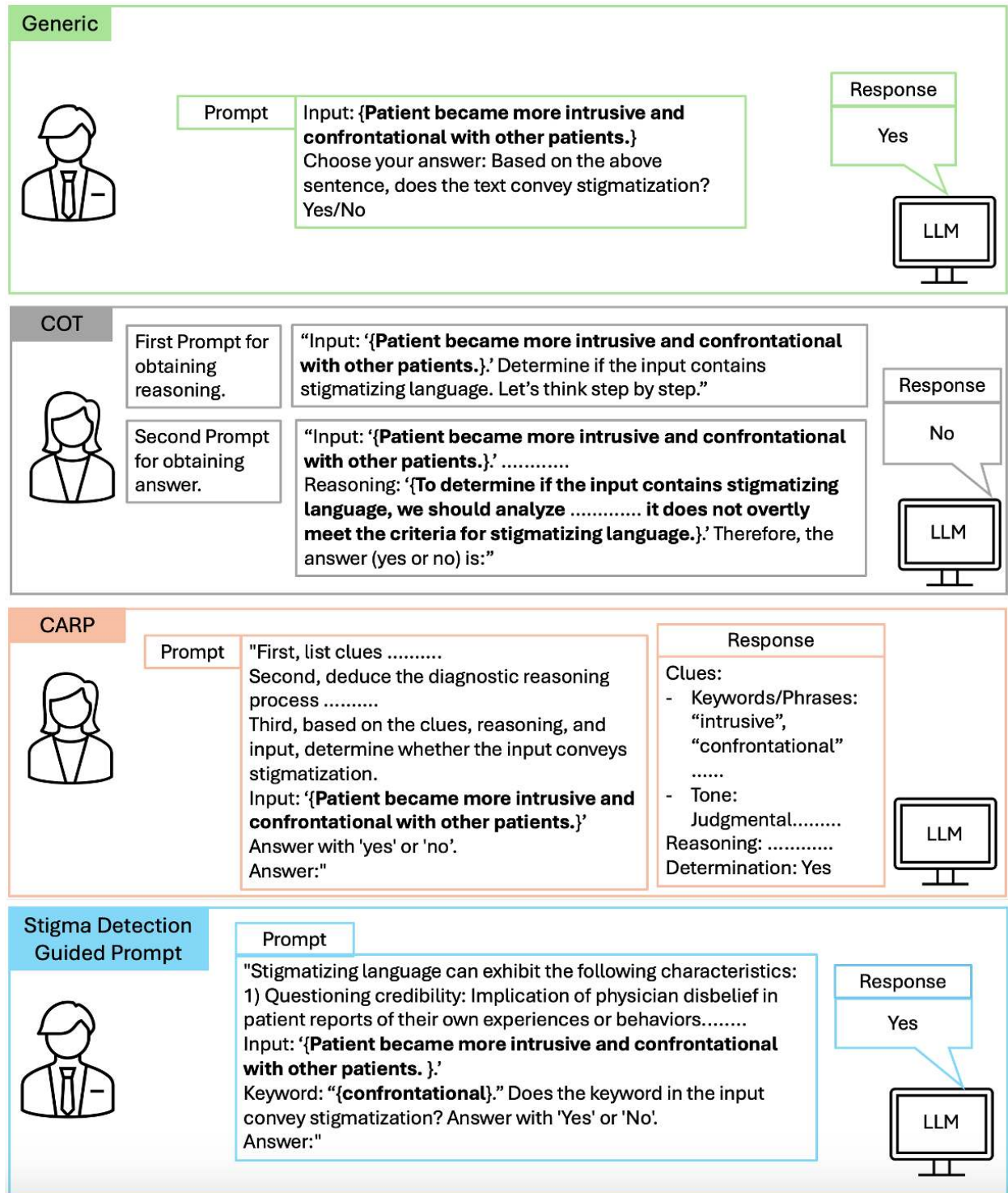


Figure 5-2 Examples of four distinct prompting strategies utilized in this study.

Table 5-6 Prompting strategies and templates used in the ICL approach.

Prompting Strategy	Prompt Template
Generic	“Input: ‘{input sentence}.’ Choose your answer: Based on the above sentence, does the text convey stigmatization? Yes/No ”
COT	“Input: ‘{input sentence}.’ Determine if the input contains stigmatizing language. Let’s think step by step. Reasoning: ‘{reasoning}.’ Therefore, the answer (yes or no) is:”
CARP	"First, list clues (e.g., keywords, phrases, contextual information, semantic relations, tones, references) that support the determination of stigmatization in the input. Second, deduce the diagnostic reasoning process from the premises (i.e., clues and input) that support the stigmatization determination. Third, based on the clues, reasoning, and input, determine whether the input conveys stigmatization. Input: ‘{input sentence}.’ Answer with 'yes' or 'no'. Answer:"
Stigma Detection	"Stigmatizing language can exhibit the following characteristics:
Guided Prompt	1) Questioning credibility: Implication of physician disbelief in patient reports of their own experiences or behaviors. 2) Disapproval: Highlights poor reasoning, decision-making, or self-care, often in a way that suggests the patient is unreasonable. 3) Stereotyping: Quoting incorrect grammar or unsophisticated terms. 4) Difficult patient: Inclusion of details with questionable clinical significance that portray the patient as belligerent or otherwise suggest that the physician is annoyed. 5) Unilateral decisions: Language that emphasizes physician authority. Input: ‘{input statement}.’

Keyword: "{keyword}."

Does the keyword in the input convey stigmatization? Answer with 'Yes' or 'No'.

Answer:"

5.3.8 Design of Stigma Detection Guided Prompt

In this study, we introduce a novel prompting strategy for detecting stigmatizing language, termed the Stigma Detection Guided Prompt. While prompt engineering has shown promise in improving LLM performance, there is no universally established framework for systematically designing prompts tailored to specific tasks [65]. The design of the proposed prompting strategy is inspired by recent research indicating that incorporating task-tailored instructions and domain knowledge can enhance the performance of the LLMs on downstream medical tasks [179,180]. Our proposed strategy consists of two key components: 1) explicitly directing the LLM to assess potentially stigmatizing words or phrases; and 2) incorporating linguistic characteristics commonly associated with stigmatizing language to enhance the model's ability to identify subtle instances of stigmatization in clinical text. Below, we outline the rationale behind these components and their contributions to improving stigma detection in clinical documentation.

In the Stigma Detection Guided Prompt, we specifically instruct the LLM to assess potentially stigmatizing words or expressions. For instance, in the sentence "Patient became more intrusive and confrontational with other patients," the model is specifically instructed to assess whether the term "confrontational" conveys stigmatization based on its surrounding context (Figure 5-2). We intentionally designed the prompt to focus on sentences containing expert-identified terms that are frequently used to convey stigmatization in clinical documentation (i.e., aggressive, non-compliant). In this task formulation, the model is not asked to detect stigmatization broadly, but rather to assess whether the use of a known high-risk term conveys stigmatization in context. This reflects a practical and clinically relevant problem, as the same term may be neutral or stigmatizing depending on how it is used. Additionally, this keyword-based approach aligns with prevailing practices in the literature. Prior studies have typically developed machine learning models or conducted analyses using sentences containing specific terms that are frequently flagged as

stigmatizing according to medical language guidelines [36,85]. This approach remains widely adopted, in part because standardized definitions of stigmatizing language are still underdeveloped, and the topic remains sensitive within healthcare contexts. Therefore, focusing on predefined terms in the prompt offers a consistent and pragmatic approach for studying stigmatizing language across datasets and research settings. Conventional prompting strategies do not effectively leverage this information. By explicitly integrating this task-tailored instruction, our approach aims to enhance the LLM’s ability to detect stigmatizing language with greater contextual awareness.

Within the proposed prompting strategy, we also integrate linguistic characteristics commonly associated with stigmatizing language to enhance the model’s ability to detect subtle and implicit stigmatization. Stigmatizing language in EHRs frequently manifests in indirect ways, where certain words or phrases acquire negative connotations depending on the context in which they are used [23,181]. For instance, the term "claims" in the phrase "Patient claims pain is 10/10" may suggest disbelief, thereby contributing to stigma [19]. Prior work has identified five recurring linguistic patterns associated with stigmatizing language in clinical documentation, including 1) questioning patient credibility, 2) expressing disapproval of patient reasoning or self-care, 3) stereotyping based on race or social class, 4) portraying patients as difficult, and 5) emphasizing physician authority over the patient [154]. By incorporating these linguistic markers into the prompting strategy, our approach aims to enhance the LLM’s sensitivity to context-dependent cues, improving its overall accuracy in detecting stigmatizing language. In the few-shot setting, we incorporated examples within the prompt, following a similar approach to that used in the generic prompting strategy.

5.3.9 Performance Evaluation and Comparison

In this study, stigmatizing language detection was framed as a binary text classification problem. To evaluate the performance of various approaches, we randomly sampled 1,000 data points to create a test dataset, ensuring no overlap with training or validation datasets. We employed five standard classification metrics: accuracy, precision, recall, area under the receiver operating characteristic curve (AUC-ROC), and F1 score. Accuracy indicates the percentage of EHR sentences that are correctly classified. Precision measures the proportion of sentences identified as stigmatizing that are stigmatizing, reflecting the model’s ability to avoid false positives. Recall,

also known as sensitivity, assesses the proportion of actual stigmatizing sentences that are correctly identified, highlighting the model’s capacity to capture true positives. The AUC-ROC evaluates the model’s ability to distinguish between stigmatizing and non-stigmatizing sentences across different classification thresholds, with higher values indicating better discriminatory power. The F1 score, which ranges from 0 to 1, is computed as the harmonic mean of precision and recall, offering a balanced measure of classification performance. The F1 score is particularly critical for evaluating stigmatizing language detection due to the highly imbalanced nature of the dataset. Notably, the AUC-ROC metric was not computed for the ICL approach, as it does not produce the probabilistic classification required for the metric.

To obtain a robust estimate of the model’s performance, we computed 95% confidence intervals (CIs) by training, tuning, and evaluating each model on ten different train-validation-test splits, each generated using a different random seed. These test datasets were designed to have no overlap with the training or validation datasets, thereby minimizing potential biases in performance estimation. To compare the top-performing models, we conducted paired t-tests on F1 score, precision, and recall. Since each model was evaluated on the same test splits, the paired t-test was chosen as it accounts for the dependency between paired observations, reducing variance and increasing statistical power compared to an independent t-test. In the zero-shot setting, we compared the highest-performing models based on F1 score in both the textual entailment and ICL approaches. In the few-shot setting, we compared the best models when different numbers of annotations were available in the SetFit and ICL approaches. Additionally, we compared the top-performing supervised fine-tuning model with the best ICL model. To mitigate the risk of inflated Type I error due to multiple comparisons, we applied the Bonferroni correction to adjust the significance threshold. Given that a total of 15 comparisons were conducted, the adjusted significance level was set at 0.003 (0.05/15).

5.3.10 Fairness Evaluation

The notion of machine learning fairness in healthcare can be broadly classified into equal allocation and equal performance considerations [155]. Equal allocation pertains to the equitable distribution of healthcare resources across different patient subgroups. For instance, in the context of kidney transplantation, a fair machine learning model would ensure that transplants are allocated proportionally among patient groups, thereby mitigating disparities in access to care [182]. Equal

performance, on the other hand, requires that a model achieves comparable predictive accuracy across both protected and non-protected groups. This can be quantified using metrics such as the true positive rate (TPR) and false positive rate (FPR), among others [133,155,156]. In the context of stigmatizing language detection, the equal performance criterion is more suitable as it ensures that the model identifies stigmatizing language consistently across different demographic groups.

For the fairness evaluation, we employed the criterion of equal performance, assessing the model across three key performance metrics: TPR, FPR, and F1 score. A classifier is considered fair under this criterion if its performance remains consistent across subgroups defined by protected attributes. The TPR, also commonly referred to as recall, quantifies the proportion of EHR sentences containing stigmatizing language that the model correctly identifies. This metric is critical for fairness evaluation, as a lower TPR for certain demographic subgroups means the model fails to detect stigmatizing language in their clinical records at the same rate as others, potentially allowing harmful biases to persist unnoticed in medical documentation. This could lead to disparities in patient advocacy efforts, as undetected stigmatizing language may influence provider perceptions and contribute to differences in the quality of care received. Over time, systematic under-detection of stigmatizing language in specific groups may reinforce existing biases in healthcare systems, further marginalizing vulnerable populations. FPR quantifies the proportion of non-stigmatizing sentences that are incorrectly classified as stigmatizing. This metric is essential for fairness evaluation, as an elevated FPR for certain demographic subgroups indicates that the model is more likely to incorrectly classify non-stigmatizing language as stigmatizing for those groups, which could lead to unnecessary scrutiny of their clinical documentation. This over-flagging may cause healthcare providers to become more hesitant in describing patient conditions accurately, potentially leading to vague or incomplete medical records. As a result, the quality of clinical communication may be compromised, which could negatively impact care coordination, patient trust, and clinical decision-making. Lastly, the F1 score provides a comprehensive measure of the model's classification performance by balancing precision and recall. Differences in F1 score across demographic subgroups indicate variability in the model's utility—its effectiveness in balancing the detection (recall) and accuracy (precision) of stigmatizing language. A lower F1 score for certain groups suggests that the model is either failing to detect stigmatizing language when it should (low recall) or flagging non-stigmatizing language incorrectly (low precision). This inconsistency can undermine the trust and usability of the model in clinical decision-making,

potentially leading to biased documentation practices and uneven policy enforcement, ultimately reducing the overall fairness and effectiveness of the system.

We assessed the fairness of the models concerning three protected attributes: sex (male and female), age (0-25, 26-50, 51-75, and 75+), and race (White, Black, Hispanic, Asian, and Other/Unknown). We adopted age groupings similar to those used in previous studies examining the fairness of machine learning classifiers [133,156]. Bias at the subgroup level was quantified by measuring performance disparities across demographic subgroups [156,183]. For sex-differentiated groups, performance disparity was calculated as the difference in the model's performance between one sex group and its complementary counterpart. For race- and age-differentiated groups, performance disparity was determined by the difference between the performance of a specific subgroup and the median performance across all subgroups within that attribute. A high-performance disparity indicates significant variation in the model's performance across different groups, suggesting a higher level of unfairness. We measured performance disparity for subgroups defined by sex, age, and race and reported the highest absolute performance disparity for each demographic attribute across models using ICL and supervised fine-tuning approaches.

5.3.11 Ethical Considerations

This study used a publicly available dataset consisting of EHR sentences. All patient records were anonymized by the dataset's original creators, ensuring that no individual identifiable information was accessible to the research team [158]. As such, the study is exempt from research ethics board review in accordance with Article 2.2 of the Tri-Council Policy Statement (TCPS 2) [184]. The original collection of data was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center, Boston, MA, and the Massachusetts Institute of Technology, Cambridge, MA (#2001P001699). The institutional review boards also granted a waiver of informed consent and approved the sharing of the research resource, thereby permitting secondary analysis of the de-identified data without additional consent.

No compensation was provided or required for the use of this secondary data. No identification of individual participants/users in any images of the manuscript or supplementary material is possible. To access this dataset, the lead author fulfilled all necessary regulatory requirements, including the completion of the course 'CITI Data or Specimens Only Research Training'

(certification number: 62353094). Additionally, the lead author signed a data use agreement to ensure the appropriate use of the data in compliance with relevant policies and regulations.

5.4 Results

5.4.1 Overview

This study conducted a comprehensive evaluation of various approaches for detecting stigmatizing language in data-scarce settings. First, we assessed the effectiveness of textual entailment and ICL approaches within a zero-shot setting. Then, we evaluated the SetFit and ICL approaches in few-shot settings, utilizing 4, 8, and 16 annotated data points per class. We compared supervised fine-tuning with ICL to assess how well ICL can perform with limited data compared to an established method that relies on a large amount of annotated data. This section presents the comparative performance of each approach in detecting stigmatizing language.

5.4.2 Performance Evaluation for Zero-shot Approaches

Table 5-7 presents the performance metrics of the various pretrained language models employing the textual entailment approach under the zero-shot setting. Among the models evaluated, DEBERTA-M achieved the highest mean F1 score and accuracy of 0.723 (95% CI [0.718, 0.728]) and 73.8% (95% CI [73.2%, 74.4%]), respectively. Conversely, BERT-M exhibited the lowest performance, with a mean F1 score and accuracy of 0.615 (95% CI [0.606, 0.624]) and 64.2% (95% CI [63.8%, 64.6%]), respectively.

Table 5-7 Performance of pretrained language models utilizing the textual entailment approach.

Pretrained language models	Accuracy (%)	F1	Precision	Recall	AUC-ROC
ROBERTA-M	72.9±0.7	0.713±0.008	0.698±0.007	0.729±0.008	0.753±0.009
BART-M	69.6±0.9	0.691±0.004	0.701±0.006	0.681±0.011	0.736±0.008
DEBERTA-M	73.8±0.6	0.723±0.005	0.717±0.007	0.730±0.006	0.781±0.010
BERT-M	64.2±0.4	0.615±0.009	0.623±0.008	0.608±0.010	0.701±0.007

Mean	70.1	0.686	0.685	0.687	0.743
------	------	-------	-------	-------	-------

The performance of the various general-purpose LLMs using the zero-shot ICL approach is presented in Table 5-8. The Stigma Detection Guided Prompt strategy yielded the highest mean accuracy of 80.7% and an F1 score of 0.838 across all evaluated prompting strategies. Conversely, the generic prompting strategy demonstrated the lowest performance, with a mean accuracy of 69.4% and an F1 score of 0.746. The mean accuracy across the prompting strategies varied by 11.3 percentage points, and the mean F1 scores differed by 0.092, indicating a significant variability in performance based on the prompting strategy employed. The Stigma Detection Guided Prompt paired with GEMMA-2 achieved the highest performance, with a mean accuracy of 83.5% (95% CI [82.9%, 84.1%]) and an F1 score of 0.858 (95% CI [0.854, 0.862]). The zero-shot ICL approach demonstrated superior performance metrics compared to the textual entailment approach. Specifically, the highest F1 score in the ICL approach, achieved by GEMMA-2, was 0.858, exceeding the best F1 score of 0.723 obtained by DEBERTA-M in the textual entailment approach by 0.135, representing an 18.7% improvement. A paired t-test demonstrated that GEMMA-2 significantly outperformed DEBERTA-M in terms of F1, Precision, and Recall ($P<.001$). The detailed statistical results are provided in Appendix 5.3 Paired t-test Comparison of Top-performing Models in Zero-shot, Few-shot, and Supervised Fine-tuning Approaches. Additionally, three out of four prompting strategies within the ICL approach resulted in higher mean accuracy and F1 scores than those observed with the textual entailment method.

Table 5-8 Performance of general-purpose LLMs utilizing the zero-shot ICL approach.

General-purpose LLM	Accuracy (%)	F1	Precision	Recall
	Generic			
LLAMA-3	70.9±0.8	0.811±0.005	0.821±0.006	0.801±0.004
FLAN-T5	67.2±0.6	0.709±0.007	0.731±0.008	0.688±0.005
GEMMA-2	70.4±0.9	0.750±0.004	0.778±0.004	0.724±0.004

MISTRAL-0.2	68.4±0.6	0.705±0.006	0.713±0.006	0.698±0.005
BIO-LLAMA-3	70.1±0.8	0.762±0.006	0.774±0.005	0.750±0.004
Mean	69.4	0.746	0.763	0.732
COT				
LLAMA-3	74.5±0.7	0.781±0.009	0.792±0.007	0.771±0.009
FLAN-T5	73.6±0.3	0.768±0.007	0.779±0.006	0.757±0.005
GEMMA-2	75.2±0.6	0.795±0.005	0.813±0.007	0.777±0.004
MISTRAL-0.2	72.5±0.4	0.756±0.008	0.768±0.004	0.745±0.009
BIO-LLAMA-3	73.1±0.7	0.775±0.006	0.787±0.007	0.764±0.009
Mean	73.8	0.775	0.788	0.763
CARP				
LLAMA-3	78.3±0.6	0.830±0.005	0.842±0.004	0.818±0.005
FLAN-T5	74.9±0.7	0.785±0.008	0.794±0.008	0.776±0.007
GEMMA-2	73.7±0.9	0.803±0.010	0.811±0.005	0.795±0.009
MISTRAL-0.2	70.8±0.5	0.736±0.007	0.752±0.005	0.721±0.007
BIO-LLAMA-3	77.9±0.6	0.815±0.005	0.831±0.007	0.799±0.004
Mean	75.1	0.794	0.806	0.782
Stigma Detection Guided Prompt				
LLAMA-3	82.4±0.6	0.845±0.005	0.822±0.006	0.869±0.008
FLAN-T5	76.6±0.8	0.819±0.005	0.804±0.004	0.834±0.007

GEMMA-2	83.5±0.6	0.858±0.004	0.841±0.005	0.876±0.003
MISTRAL-0.2	80.1±0.5	0.830±0.004	0.813±0.006	0.847±0.006
BIO-LLAMA-3	81.0±0.5	0.839±0.005	0.857±0.007	0.821±0.004
Mean	80.7	0.838	0.827	0.849

5.4.3 Performance Evaluation for Few-shot Approaches

Table 5-9 presents the performance of the various pretrained sentence transformers utilizing the SetFit approach in the few-shot setting across different numbers of labeled data points per class. MINILM-V2 achieved the highest mean F1 score of 0.721 (95% CI [0.713, 0.729]) and the highest mean AUC-ROC of 0.791 (95% CI [0.785, 0.797]) with 4 annotated data points per class. However, E5-V2 outperformed the other models when 8 and 16 annotated data points per class were available, achieving mean F1 scores of 0.735 (95% CI [0.728, 0.742]) and 0.802 (95% CI [0.798, 0.806]), respectively. On average, accuracy increased from 72.1% with 4 annotated data points per class to 78.0% with 16 annotated data points per class, while the average F1 score improved from 0.707 to 0.767. These results highlight the positive impact of increasing the number of annotated data points on model performance within the few-shot SetFit approach.

Table 5-9 Performance of pretrained sentence transformers utilizing the SetFit approach.

Sentence Transformer	Accuracy (%)	F1	Precision	Recall	AUC- ROC
	4 annotated data per class				
MPNET-V2	70.5±0.4	0.701±0.005	0.708±0.007	0.697±0.005	0.752±0.005
E5-V2	73.6±0.6	0.704±0.003	0.705±0.004	0.703±0.003	0.784±0.007
ROBERTA-V2	72.2±0.6	0.708±0.004	0.712±0.005	0.704±0.005	0.765±0.005

MINILM-V2	73.0±0.5	0.721±0.008	0.718±0.007	0.725±0.005	0.791±0.006
BERT-BIO-ST	71.4±0.7	0.701±0.008	0.687±0.004	0.715±0.006	0.770±0.008
BIOLORD	72.0±0.7	0.707±0.008	0.714±0.005	0.701±0.009	0.772±0.005
Mean	72.1	0.707	0.706	0.707	0.772
8 annotated data per class					
MPNET-V2	72.6±0.8	0.715±0.003	0.727±0.006	0.704±0.004	0.773±0.005
E5-V2	77.1±0.4	0.735±0.007	0.750±0.006	0.721±0.008	0.836±0.006
ROBERTA-V2	75.4±0.8	0.723±0.005	0.729±0.006	0.718±0.004	0.814±0.006
MINILM-V2	74.3±0.6	0.732±0.005	0.721±0.004	0.743±0.005	0.801±0.007
BERT-BIO-ST	75.1±0.7	0.726±0.007	0.715±0.008	0.736±0.005	0.817±0.005
BIOLORD	75.9±0.4	0.729±0.008	0.718±0.005	0.740±0.007	0.821±0.004
Mean	75.1	0.727	0.726	0.728	0.810
16 annotated data per class					
MPNET-V2	75.6±0.3	0.755±0.005	0.773±0.005	0.738±0.006	0.815±0.007
E5-V2	80.7±0.5	0.802±0.004	0.813±0.006	0.792±0.003	0.858±0.006
ROBERTA-V2	78.3±0.4	0.775±0.006	0.766±0.007	0.784±0.004	0.838±0.005
MINILM-V2	79.1±0.6	0.785±0.007	0.802±0.006	0.769±0.005	0.847±0.009
BERT-BIO-ST	77.3±0.5	0.745±0.008	0.752±0.007	0.738±0.008	0.832±0.004
BIOLORD	77.1±0.8	0.739±0.004	0.744±0.005	0.734±0.004	0.829±0.006
Mean	78.0	0.767	0.775	0.759	0.837

Table 5-10 presents the F1 score and accuracy of various general-purpose LLMs using the few-shot ICL approach with different prompting strategies. Additional performance metrics, including precision and recall, are provided in Appendix 5.4 Performance Metrics of the Few-shot In-Context Learning Models. Due to limitations in the number of tokens (or words) that models can process at once (known as context window limitation), it was not feasible to conduct experiments with FLAN-T5 for the COT and CARP prompting strategies in the few-shot settings, nor with LLAMA-3 and GEMMA-2 for COT and CARP when 16 annotated data points per class were used. Thus, we don't report the mean performance of COT and CARP prompting strategies across models. LLAMA-3, paired with the Stigma Detection Guided Prompt, demonstrated the highest performance across all scenarios, achieving F1 scores of 0.874 (95% CI [0.864, 0.884]), 0.892 (95% CI [0.886, 0.898]), and 0.901 (95% CI [0.895, 0.907]) with 4, 8, and 16 annotated data points per class, respectively. When using the Generic and the Stigma Detection Guided Prompt, it was observed that both the F1 score and accuracy improved as the number of annotated data points per class increased.

Table 5-10 Performance of general-purpose LLMs utilizing the few-shot ICL approach.

	4 annotated data per class		8 annotated data per class		16 annotated data per class	
General Purpose LLM	Generic					
	Accuracy (%)	F1	Accuracy (%)	F1	Accuracy (%)	F1
LLAMA-3	74.8±0.8	0.789±0.006	79.2±0.9	0.832±0.008	81.6±0.8	0.855±0.006
FLAN-T5	69.3±0.5	0.691±0.007	70.7±0.8	0.701±0.008	75.1±0.5	0.762±0.006
GEMMA-2	71.1±1.2	0.719±0.007	72.1±0.8	0.739±0.009	76.2±0.9	0.773±0.008
MISTRAL-0.2	72.5±0.7	0.730±0.006	70.5±0.9	0.720±0.010	80.1±1.0	0.822±0.007
BIO-LLAMA-3	76.6±0.9	0.820±0.008	77.3±0.5	0.826±0.004	79.0±0.6	0.851±0.006
Mean	72.9	0.75	74.0	0.764	78.4	0.813

COT						
LLAMA-3	72.5±0.8	0.756±0.004	83.1±0.9	0.823±0.007	—	—
FLAN-T5	—	—	—	—	—	—
GEMMA-2	75.1±0.4	0.796±0.006	79.5±0.8	0.826±0.005	—	—
MISTRAL-0.2	71.3±0.3	0.742±0.005	75.1±0.6	0.767±0.007	76.9±0.8	0.781±0.006
BIO-LLAMA-3	71.6±0.7	0.743±0.005	81.4±0.8	0.835±0.006	—	—
CARP						
LLAMA-3	80.4±0.8	0.837±0.005	85.1±0.5	0.876±0.004	—	—
FLAN-T5	—	—	—	—	—	—
GEMMA-2	75.4±0.7	0.795±0.006	79.6±0.4	0.825±0.009	—	—
MISTRAL-0.2	71.8±0.5	0.728±0.008	72.2±0.5	0.747±0.006	71.2±0.6	0.731±0.005
BIO-LLAMA-3	77.2±0.7	0.813±0.009	81.6±0.6	0.849±0.008	—	—
Stigma Detection Guided Prompt						
LLAMA-3	84.3±0.6	0.874±0.010	85.5±0.7	0.892±0.006	87.7±0.7	0.901±0.006
FLAN-T5	74.2±0.4	0.782±0.006	76.8±0.5	0.805±0.008	82.8±0.3	0.842±0.004
GEMMA-2	84.3±0.9	0.872±0.007	84.9±0.4	0.881±0.005	86.1±0.5	0.887±0.006
MISTRAL-0.2	80.5±0.7	0.849±0.005	82.0±0.6	0.857±0.004	86.7±0.7	0.899±0.005
BIO-LLAMA-3	80.4±1.1	0.861±0.010	81.5±0.8	0.870±0.007	82.7±1.0	0.892±0.009

Mean	80.7	0.848	82.1	0.861	85.2	0.884
------	------	-------	------	-------	------	-------

Table 5-11 presents a comparison between the top-performing models using the SetFit approach and those employing the few-shot ICL approach across different numbers of annotated data points per class. In all scenarios, the few-shot ICL models exhibited superior performance compared to their SetFit counterparts. Specifically, with 4 annotated data points per class, LLAMA-3 achieved an F1 score of 0.874 (95% CI [0.864, 0.884]), which is 0.153 points higher than that of MINILM-V2, whose F1 score was 0.721 (95% CI [0.713, 0.729]), representing an approximate 21.2% improvement. For 8 annotated data points per class, LLAMA-3 attained an F1 score of 0.892 (95% CI [0.886, 0.898]) compared to 0.735 (95% CI [0.728, 0.742]) for E5-V2, corresponding to an improvement of about 21.4%. Finally, at 16 annotated data points per class, LLAMA-3 demonstrated an F1 score of 0.901 (95% CI [0.895, 0.907]), which is 0.099 points higher than that of E5-V2, whose F1 score was 0.802 (95% CI [0.798, 0.806]), reflecting a 12.3% improvement. For each setting (4,8, and 16 annotated data points per class), a paired t-test was conducted to compare the top-performing models from the ICL and SetFit approaches (Appendix 5.3 Paired t-test Comparison of Top-performing Models in Zero-shot, Few-shot, and Supervised Fine-tuning Approaches.). In all cases, the results indicated that the ICL models outperformed the SetFit models in terms of F1, precision, and recall ($P<.001$).

Table 5-11 Comparison of top-performing models in SetFit and few-shot ICL approaches.

Approach	4 annotated data per class		8 annotated data per class		16 annotated data per class	
	Model	F1	Model	F1	Model	F1
SetFit	MINILM-V2	0.721±0.008	E5-V2	0.735±0.007	E5-V2	0.802±0.004
Few-shot ICL	LLAMA-3	0.874±0.010	LLAMA-3	0.892±0.006	LLAMA-3	0.901±0.006

5.4.4 Performance Comparison of the ICL with the Fully Supervised Fine-tuning Approach

We presented the performance of the various pretrained language models using the fully supervised fine-tuning approach in Table 5-12. Among the models evaluated, ROBERTA demonstrated the highest performance, achieving a ROC-AUC of 0.973 (95% CI [0.968, 0.978]) and an F1 score of 0.931 (95% CI [0.924, 0.938]). In contrast, BERT-BIOMED exhibited the lowest performance, with a mean ROC-AUC of 0.926 (95% CI [0.918, 0.934]) and an F1 score of 0.894 (95% CI [0.888, 0.900]).

Table 5-12 Performance of pretrained language models for the fully supervised fine-tuning approach.

Pretrained language model	Accuracy (%)	F1	Precision	Recall	ROC-AUC
ROBERTA	94.5±0.6	0.931±0.007	0.950±0.008	0.913±0.009	0.973±0.005
BERT	94.2±0.7	0.922±0.005	0.943±0.007	0.901±0.011	0.962±0.004
BERT-BIO	93.7±0.8	0.910±0.009	0.929±0.006	0.893±0.006	0.948±0.003
ROBERTA-XLM	94.0±0.5	0.911±0.006	0.945±0.007	0.879±0.008	0.943±0.009
BERT-BIOMED	91.5±0.7	0.894±0.006	0.909±0.004	0.879±0.008	0.926±0.008
Mean	93.6	0.914	0.935	0.893	0.950

Table 5-13 presents the performance of the top-performing models in the zero-shot, few-shot settings, and fully supervised fine-tuning approach. In the zero-shot setting, the best-performing model, GEMMA-2, achieved an F1 score of 0.858 (95% CI [0.854, 0.862]), which is 0.073 points lower than the F1 score of fully supervised ROBERTA at 0.931 (95% CI [0.924, 0.938]). This difference corresponds to an approximate 8.5% improvement in F1 for the fully supervised approach. In the few-shot setting, LLAMA-3 emerged as the top-performing model, achieving F1

scores of 0.874 (95% CI [0.864, 0.884]), 0.892 (95% CI [0.886, 0.898]), and 0.901 (95% CI [0.895, 0.907]) with 4, 8, and 16 annotated data points per class, respectively. Although the F1 score with 16 annotated data points per class was about 3.2% lower than that of fully supervised ROBERTA, it is notable that this performance was attained using only 32 labeled instances, whereas ROBERTA was trained with 3,543 instances—more than 100 times the amount of annotated data.

Table 5-13 Performance of top-performing models in zero-shot, few-shot, and fully supervised fine-tuning approaches.

Zero-shot		Few-shot: 4 annotated data per class		Few-shot: 8 annotated data per class		Few-shot: 16 annotated data per class		Fully supervised fine- tuning	
Model	F1	Model	F1	Model	F1	Model	F1	Model	F1
GEMM	0.858±0.0	LLAM	0.874±0.0	LLAM	0.892±0.0	LLAM	0.901±0.0	ROBERT	0.931±0.0
A-2	04	A-3	10	A-3	06	A-3	06	A	07

5.4.5 Statistical Comparison of Model Performance

Table 5-14 summarizes the key findings from paired t-tests comparing the top-performing models in the zero-shot and few-shot settings, as well as performance differences between the best-performing ICL and supervised fine-tuning models. Additional details on the statistical comparisons are provided in Appendix 5.3 Paired t-test Comparison of Top-performing Models in Zero-shot, Few-shot, and Supervised Fine-tuning Approaches. In the zero-shot setting, paired t-tests revealed that GEMMA-2 significantly outperformed DeBERTa-M in terms of F1 score, precision, and recall (all $P < .001$). In the few-shot setting, ICL models demonstrated significantly higher performance across all three metrics compared to SetFit models. When comparing the top-performing models from the ICL and supervised fine-tuning approaches, RoBERTa achieved significantly higher F1 score and precision ($P < .001$), while LLAMA-3 demonstrated significantly higher recall ($P < .001$).

Table 5-14 Paired t-test results comparing the top-performing models in the zero-shot and few-shot settings, and between the best-performing models in ICL and supervised fine-tuning approaches.

Setting	Comparison	Metric	Mean difference	P-value
Zero-shot	Deberta-M vs. GEMMA-2 coupled with Stigma Detection Guided Prompt	F1	-0.136	<.001
		Precision	-0.125	<.001
		Recall	-0.146	<.001
Few-shot (4 annotations per label)	MINILM-V2 vs. LLAMA 3 coupled with Stigma Detection Guided Prompt	F1	-0.155	<.001
		Precision	-0.138	<.001
		Recall	-0.165	<.001
Few-shot (8 annotations per label)	E5-V2 vs. LLAMA 3 coupled with Stigma Detection Guided Prompt	F1	-0.158	<.001
		Precision	-0.124	<.001
		Recall	-0.192	<.001
Few-shot (16 annotations per label)	E5-V2 vs. LLAMA 3 coupled with Stigma Detection Guided Prompt	F1	-0.101	<.001
		Precision	-0.052	<.001
		Recall	-0.152	<.001
ICL vs. supervised fine-tuning	ROBERTA vs. LAMMA-3 coupled with Stigma Detection Guided Prompt (16 annotated data per class)	F1	0.029	<.001
		Precision	0.086	<.001
		Recall	-0.030	<.001

5.4.6 Fairness Evaluation and Comparison

The performance and disparities of various models developed to detect stigmatizing language are presented in Appendix 5.5 Performance and Performance Disparities of Models across Sex, Age, and Race Demographic Attributes. Table 5-15 highlights the largest absolute performance

disparities across demographic attributes, including sex, age, and race, for each model. Models employing supervised fine-tuning exhibited relatively higher absolute disparities compared to the ICL models under zero-shot, 4-shot, 8-shot, and 16-shot settings. On average, models trained using supervised fine-tuning demonstrated the mean largest absolute TPR disparities of 0.037, 0.077, and 0.046 for sex-, age-, and race-differentiated groups, respectively. These disparities indicate that the TPR can differ by up to 7.7 % in certain demographic subgroups. In contrast, ICL models showed more balanced TPR, with no mean disparity exceeding 0.007, 0.010, and 0.013 for sex-, age-, and race-differentiated groups, respectively. A similar trend was observed for FPR and F1 scores. Supervised fine-tuning models exhibited greater FPR disparities, with mean values of 0.026, 0.027, and 0.031 for sex-, age-, and race-differentiated groups, respectively. Conversely, the ICL models demonstrated lower FPR disparities under different experimental settings, indicating a more consistent FPR across groups. Table 5-16 presents the highest performance disparities of ICL and supervised fine-tuning models across sex, age, and race subgroups. Here, the highest disparities refer to the highest observed differences in performance metrics across demographic groups for each modeling approach. For TPR, supervised models exhibited disparities of up to 0.051 for sex, 0.108 for age, and 0.064 for race. In contrast, both zero-shot and few-shot ICL models showed substantially lower TPR disparities, with all values below 0.016. Similarly, FPR disparities reached up to 0.039 for sex, 0.037 for age, and 0.043 for race in supervised models, whereas ICL models remained below 0.027. Regarding F1 scores, the highest disparity among supervised models reached 0.086, compared to less than 0.025 among ICL models. These findings suggest that ICL models offer more equitable performance across demographic groups.

Table 5-15 The largest absolute performance disparity across demographic attributes for different approaches in detecting stigmatizing language.

Approach	Model	Sex	Age	Race
		Largest absolute TPR disparity		
Supervised Fine-tuning	ROBERTA	0.049	0.108	0.044
	BERT	0.051	0.058	0.036
	BERT-BIO	0.022	0.075	0.037

	ROBERTA-XLM	0.033	0.089	0.052
	BERT-BIOMED	0.032	0.053	0.064
	Mean	0.037	0.077	0.046
Zero-shot ICL	LLAMA-3	0.005	0.012	0.011
	FLAN-T5	0.002	0.004	0.008
	GEMMA-2	0.003	0.009	0.010
	MISTRAL-0.2	0.007	0.005	0.006
	BIO-LLAMA-3	0.006	0.010	0.013
	Mean	0.005	0.008	0.010
Few-shot ICL (4 annotated data per class)	LLAMA-3	0.003	0.008	0.013
	FLAN-T5	0.004	0.009	0.014
	GEMMA-2	0.002	0.007	0.012
	MISTRAL-0.2	0.014	0.006	0.015
	BIO-LLAMA-3	0.008	0.008	0.010
	Mean	0.006	0.008	0.013
Few-shot ICL (8 annotated data per class)	LLAMA-3	0.006	0.008	0.010
	FLAN-T5	0.003	0.005	0.009
	GEMMA-2	0.006	0.007	0.011
	MISTRAL-0.2	0.014	0.004	0.007
	BIO-LLAMA-3	0.005	0.008	0.007
	Mean	0.007	0.006	0.009
Few-shot ICL (16 annotated data per class)	LLAMA-3	0.003	0.007	0.012

	FLAN-T5	0.007	0.008	0.007
	GEMMA-2	0.006	0.012	0.013
	MISTRAL-0.2	0.007	0.011	0.009
	BIO-LLAMA-3	0.006	0.013	0.010
	Mean	0.006	0.010	0.010
Largest absolute FPR disparity				
Supervised Fine-tuning	ROBERTA	0.018	0.022	0.028
	BERT	0.016	0.015	0.023
	BERT-BIO	0.035	0.033	0.027
	ROBERTA-XLM	0.021	0.023	0.043
	BERT-BIOMED	0.039	0.037	0.032
	Mean	0.026	0.027	0.031
Zero-shot ICL	LLAMA-3	0.003	0.008	0.008
	FLAN-T5	0.007	0.011	0.010
	GEMMA-2	0.008	0.018	0.019
	MISTRAL-0.2	0.005	0.016	0.014
	BIO-LLAMA-3	0.001	0.011	0.008
	Mean	0.005	0.013	0.012
Few-shot ICL (4 annotated data per class)	LLAMA-3	0.006	0.013	0.007
	FLAN-T5	0.007	0.015	0.011
	GEMMA-2	0.005	0.013	0.015
	MISTRAL-0.2	0.003	0.016	0.014

	BIO-LLAMA-3	0.005	0.017	0.015
	Mean	0.005	0.015	0.012
Few-shot ICL (8 annotated data per class)	LLAMA-3	0.002	0.007	0.009
	FLAN-T5	0.011	0.026	0.019
	GEMMA-2	0.012	0.013	0.010
	MISTRAL-0.2	0.007	0.021	0.014
	BIO-LLAMA-3	0.005	0.011	0.014
	Mean	0.007	0.016	0.013
Few-shot ICL (16 annotated data per class)	LLAMA-3	0.007	0.015	0.012
	FLAN-T5	0.009	0.011	0.009
	GEMMA-2	0.004	0.013	0.014
	MISTRAL-0.2	0.004	0.016	0.013
	BIO-LLAMA-3	0.014	0.017	0.010
	Mean	0.008	0.014	0.011
Largest absolute F1 disparity				
Supervised Fine-tuning	ROBERTA	0.043	0.066	0.064
	BERT	0.037	0.048	0.030
	BERT-BIO	0.026	0.058	0.031
	ROBERTA-XLM	0.064	0.086	0.032
	BERT-BIOMED	0.041	0.053	0.051
	Mean	0.042	0.063	0.042
Zero-shot ICL	LLAMA-3	0.006	0.015	0.006

	FLAN-T5	0.004	0.018	0.011
	GEMMA-2	0.010	0.013	0.016
	MISTRAL-0.2	0.006	0.015	0.010
	BIO-LLAMA-3	0.005	0.005	0.009
	Mean	0.006	0.013	0.010
Few-shot ICL (4 annotated data per class)	LLAMA-3	0.005	0.010	0.017
	FLAN-T5	0.012	0.011	0.013
	GEMMA-2	0.008	0.013	0.014
	MISTRAL-0.2	0.004	0.016	0.013
	BIO-LLAMA-3	0.010	0.011	0.018
	Mean	0.008	0.012	0.015
Few-shot ICL (8 annotated data per class)	LLAMA-3	0.008	0.009	0.006
	FLAN-T5	0.010	0.023	0.016
	GEMMA-2	0.003	0.013	0.015
	MISTRAL-0.2	0.018	0.017	0.024
	BIO-LLAMA-3	0.015	0.012	0.019
	Mean	0.011	0.015	0.016
Few-shot ICL (16 annotated data per class)	LLAMA-3	0.006	0.014	0.014
	FLAN-T5	0.013	0.020	0.025
	GEMMA-2	0.008	0.011	0.017
	MISTRAL-0.2	0.015	0.014	0.018
	BIO-LLAMA-3	0.005	0.013	0.011

	Mean	0.009	0.014	0.017
--	------	-------	-------	-------

Table 5-16 The highest observed performance disparities of ICL and supervised fine-tuning models across sex, age, and race subgroups.

Approach	Sex	Age	Race
	Highest TPR disparity		
Supervised fine-tuning	0.051	0.108	0.064
Zero-shot ICL	0.007	0.012	0.013
Few-shot ICL	0.014	0.013	0.015
	Highest FPR disparity		
Supervised fine-tuning	0.039	0.037	0.043
Zero-shot ICL	0.008	0.018	0.019
Few-shot ICL	0.014	0.026	0.019
	Highest F1 disparity		
Supervised fine-tuning	0.064	0.086	0.064
Zero-shot ICL	0.010	0.018	0.016
Few-shot ICL	0.018	0.023	0.025

5.5 Discussion

5.5.1 Overview

The prevalence of stigmatizing language in EHR has raised significant concern due to its detrimental impact on patient safety [19,21,24,36,148,185]. Numerous studies have investigated

the potential of utilizing supervised machine-learning approaches for detecting stigmatizing language [25,36,37]. However, these approaches typically require extensive amounts of labeled data to achieve optimal performance. The creation of such a dataset is particularly challenging, given the difficulty in identifying stigmatizing language, which is often subtle, implicit, and highly contextual [25,154]. The ICL approach has emerged as a promising alternative due to its strong contextual understanding capabilities and minimal reliance on labeled data [65,149–151]. In this study, we extensively investigated the efficacy of ICL in detecting stigmatizing language and compared its performance with established zero-shot and few-shot text classification methods. To the best of our knowledge, this is the first study to explore the use of ICL, as well as zero-shot and few-shot approaches, for detecting stigmatizing language. Furthermore, we proposed a novel prompting strategy designed for detecting stigmatizing language and evaluated its effectiveness against state-of-the-art prompting strategies such as COT and CARP.

5.5.2 Principal Findings

The ICL approaches have demonstrated a significant advantage in detecting stigmatizing language compared to both the popular zero-shot textual entailment approach and the few-shot SetFit approach. In the zero-shot setting, the top-performing ICL model, GEMMA-2, achieved an F1 score of 0.858 (95% CI [0.854, 0.862]), surpassing the best F1 score of 0.723 (95% CI [0.718, 0.728]) obtained by ROBERTA-M in the textual entailment approach by 0.135, representing an 18.7% improvement. Similarly, in the few-shot setting, the top-performing ICL model, LLAMA-3, exhibited F1 score improvements of 21.2%, 21.4%, and 12.3% over the leading SetFit models when utilizing 4, 8, and 16 annotations per class, respectively. The superior performance of the ICL approach can be attributed to the enhanced contextual understanding capability and rich knowledge of general-purpose LLMs [76,79]. These general-purpose LLMs have undergone sophisticated training processes, such as instructional fine-tuning, reinforcement learning from human feedback, and extensive pretraining across a diverse range of NLP tasks, including language modeling, text completion, and semantic understanding on vast amounts of textual data [80,173,186]. These training processes have enabled the general-purpose models to better comprehend and contextualize subtle and complex linguistic patterns, making the ICL approach particularly effective for detecting stigmatizing language in EHRs. Therefore, in settings where annotated data is limited, the ICL approach should be prioritized for detecting stigmatizing language. However, it is important to note that the ICL models do not produce probabilistic

classification outputs, which prevents the computation of threshold-independent metrics such as AUC-ROC. This limitation affects the interpretation of results in two key ways. First, it may obscure comparisons with probabilistic models, which may exhibit strong overall class separation but suffer from performance instability at specific thresholds (i.e., precision, recall, or F1 scores). Second, more importantly, the lack of probability outputs prevents us from adjusting the decision threshold to accommodate different preferences for false positives versus false negatives, which is a common requirement in some clinical applications. Some clinical contexts may prioritize minimizing false positives to avoid over-alerting, while others may emphasize minimizing false negatives to avoid missed harms.

Within the ICL approach, substantial performance variability was observed across different prompting strategies. Specifically, in the zero-shot setting, the mean accuracy for the ICL approach varied between 69.4% and 80.7% across different prompting strategies, indicating a gap of 11.3%. In the few-shot setting, the differences in mean accuracy across prompting strategies were 7.8%, 8.1%, and 6.8% when utilizing 4, 8, and 16 annotations per class, respectively. These performance gaps highlight the critical importance of selecting an appropriate prompting strategy to optimize the effectiveness of the ICL approach in detecting stigmatizing language. The new prompting strategy we propose in this work, the Stigma Detection Guided Prompt, outperformed the established strategies, such as COT and CARP [150,157]. This strategy achieved the highest F1 scores across both zero-shot and few-shot settings, with an F1 score of 0.858 (95% CI [0.854, 0.862]) in the zero-shot setting and F1 scores of 0.874 (95% CI [0.864, 0.884]), 0.892 (95% CI [0.886, 0.898]), and 0.901 (95% CI [0.895, 0.907]) in the few-shot settings using 4, 8, and 16 annotations per class, respectively. The goal of the newly proposed prompt is to encode domain-specific knowledge that can assist the LLM in accurately identifying stigmatizing language. This was inspired by the fact that domain-specific information has been used in various settings to improve the performance of AI algorithms, such as designing efficient search heuristics for informed search [187] or enhancing machine learning accuracy [188,189]. The prompt incorporated linguistic cues commonly associated with stigmatizing language, such as stereotyping, disapproval, and questioning credibility [154]. These elements enhance the model’s ability to detect nuanced stigmatizing language that might otherwise be overlooked. Consequently, the Stigma Detection Guided Prompt represents a critical advancement in optimizing the ICL approach for this task. Its demonstrated effectiveness underscores the importance of leveraging

medical informatics-driven insights to improve the detection of stigmatizing language, thereby supporting bias-aware clinical documentation practices.

To thoroughly assess the potential of the ICL approach, we compared the top-performing model in the ICL approach with that of the supervised fine-tuning approach, which is widely regarded as the standard method for text classification [61,150]. The leading model in the supervised fine-tuning approach, ROBERTA, achieved an F1 score of 0.931 (95% CI [0.924, 0.938]). In comparison, the best-performing ICL model, LLAMA-3, attained an F1 score of 0.901 (95% CI [0.895, 0.907]) with only 16 annotations per class. While the F1 score difference of 0.030 corresponds to a 3.2% advantage for the fully supervised fine-tuning approach, it is worth noting that ROBERTA was developed using a training dataset of 3,543 annotated sentences and a validation dataset of 500 annotated sentences, whereas LLAMA-3 used a training dataset of 32 annotated sentences and a validation dataset of 30 annotated sentences. This comparison highlighted the remarkable efficiency of the ICL approach.

As part of our analysis, we examined the fairness dimension of the models of few-shot ICL and the supervised fine-tuning approach. The results indicate that models using the supervised fine-tuning approach exhibited greater bias compared with ICL models. Specifically, with respect to TPR, supervised fine-tuning models demonstrated mean TPR disparities of 0.037, 0.077, and 0.046 across sex-, age-, and race-differentiated groups, respectively, suggesting that the TPR could vary by up to 7.7% for certain demographic subgroups. Similarly, with respect to the FPR, supervised fine-tuning models exhibited mean disparities of 0.026, 0.027, and 0.013 for sex-, age-, and race-differentiated groups, respectively. These findings indicate that supervised fine-tuning models are more likely to assign false positive classifications at uneven rates across demographic subgroups, potentially leading to the over-flagging of stigmatizing language in certain populations. In contrast, ICL models tested in zero-shot, 4-shot, 8-shot, and 16-shot settings demonstrated a more balanced performance across subgroups, with no mean TPR disparities exceeding 0.016. Previous studies have shown that the supervised fine-tuning approach is susceptible to bias due to its dependence on training datasets that may have an imbalanced demographic distribution [190–192]. The primary objective of this approach is to optimize overall model performance, which can result in an emphasis on the majority or well-represented groups within the dataset. Consequently, when certain groups are underrepresented in the training data, the model's performance tends to decline for those groups, resulting in disparities. In contrast, the ICL approach generates predictions based

on the specific context presented in the prompt, minimizing the risk of inheriting and amplifying biases present in the training dataset. The adoption of the ICL approach could, therefore, contribute to more equitable detection of stigmatizing language across all patient populations.

5.5.3 Clinical Adoption and Privacy Considerations

The integration of a stigmatization detection model into clinical documentation systems has the potential to enhance patient safety. However, its deployment requires careful ethical consideration. If clinicians perceive that their language is being excessively monitored or that following language guidelines imposes a cognitive burden, they may become resistant to the model or attempt to circumvent its use by omitting clinically relevant details to avoid triggering the model’s flagging system. Such reluctance could limit the model’s effectiveness in fostering more inclusive documentation practices and impact the accuracy and completeness of medical records. Therefore, it is essential to clearly communicate the model’s purpose, emphasizing its role as a supportive, educational tool rather than a punitive mechanism [84]. Additionally, integrating a recommendation system that offers context-sensitive, neutral language recommendations as alternatives to flagged terms may improve clinician acceptance and facilitate smoother adoption into clinical workflows. Furthermore, real-world datasets often exhibit imbalances concerning protected attributes [191], which may result in the model disproportionately flagging certain patient groups (e.g., racial minorities, and individuals with specific medical conditions) while under-detecting stigmatization in others. If left unaddressed, this could inadvertently reinforce stereotypes and exacerbate disparities in clinical documentation [193]. To mitigate these risks, the model should undergo rigorous fairness evaluations across diverse demographic groups and incorporate appropriate bias mitigation strategies before deployment [190,192]. Furthermore, we were unable to evaluate proprietary models such as GPT-4 due to data-sharing restrictions imposed by the MIMIC-IV dataset [158]. While our findings highlight the promise of open-source LLMs in data-scarce settings, the exclusion of proprietary models limits the generalizability of our findings. Proprietary models often incorporate more advanced architectures and are trained on larger, more diverse corpora, which may allow them to outperform open-source alternatives [80]. Future studies leveraging synthetic data or institutional data-sharing agreements may be needed to assess if these proprietary models outperform open-source models in detecting stigmatizing language.

With the growing adoption of AI scribes in clinical settings, important considerations arise regarding the potential introduction or perpetuation of stigmatizing language in automatically generated clinical notes. Current AI scribing tools often operate as stand-alone systems, transcribing clinician–patient conversations and generating documentation with limited real-time oversight [194,195]. The detection framework developed in this study is intended to identify not only overtly stigmatizing terms but also subtler, context-dependent forms of stigmatization commonly found in clinical narratives [25,36]. The detection tool developed in this study can address this challenge in two ways. First, it can be applied during model development to identify and remove implicit forms of stigmatizing content from training data, thereby preventing LLMs from learning and reproducing such language. Second, it can serve as a post-generation filter to flag potentially stigmatizing content in AI-generated clinical notes. In addition, our models can be used to design a human-AI collaborative scribing tool where AI-generated drafts are reviewed and finalized by clinicians [196]. In this context, the tool developed in this study can assist by alerting clinicians to potentially stigmatizing language, thereby promoting more neutral and patient-centered documentation. Future work should focus on validating the tool’s effectiveness in real-world clinical scribing environments and examining clinician acceptance and workflow integration.

Beyond ethical and deployment considerations, the variability of stigmatizing language across documentation types and clinical settings presents a practical consideration for deployment. Prior studies indicate that stigmatizing language is more frequently observed in history and physical notes, consultation notes, and discharge summaries compared to other types of clinical documentation, and tends to be more common in relation to conditions such as substance use disorders, mental health conditions, diabetes, and obesity [43,139]. These patterns underscore the importance of accounting for documentation context and clinical setting when developing stigma detection tools. For instance, the ICL approach could be optimized by incorporating examples drawn from specific note types or clinical domains where stigmatizing language is more prevalent. Such tailoring may improve the model’s sensitivity to context-dependent expressions of bias. Moreover, recognizing that different specialties and note types reflect distinct clinician workflows, linguistic norms, and patient populations, it is essential to validate model performance across these subdomains. These validation efforts are critical to ensuring equitable performance across documentation formats and minimizing unintended disparities during deployment.

In addition to detecting stigmatizing language, LLMs can serve as valuable preprocessing tools for identifying linguistic patterns that may be integrated into downstream statistical models to explore potential associations between language use and care quality or operational decisions [36]. Furthermore, LLMs could provide real-time alerts about the potential impact of specific terms in clinical notes. These systems can draw on findings from prior research on the effects of specific terminology in clinical communication to prompt clinicians to reflect on their language choices. For example, if a term such as “aggressive” is used in a clinical note and identified by LLM as stigmatizing based on its contextual usage, the system could generate a prompt indicating that the term may lead to patient discomfort or feelings of alienation. Such feedback could serve as a reflective checkpoint, prompting clinicians to reconsider their word choices and adopt more neutral, patient-centered language.

From the perspective of patient privacy and confidentiality, while de-identification is a crucial step in protecting patient privacy [184,197], the use of potentially sensitive data, such as stigmatizing language, still presents confidentiality concerns. Specifically, there are two key privacy risks associated with deploying machine learning-based stigmatization detection models in clinical settings. First, despite de-identification, the model may learn associations between specific linguistic patterns and particular patient populations [198]. For example, if the model systematically classifies clinical notes from a specific hospital, demographic group, or medical condition as stigmatizing, it could lead to unintended disclosures about those populations. If this information were to be combined with external data sources, there is a potential risk of re-identifying patient groups or even individuals. To mitigate such risks, implementing strict access control measures is necessary to ensure that only authorized clinicians and relevant personnel can interact with the model [198,199]. Second, the dataset used to train the model was curated based on sentences containing potentially stigmatizing words or expressions, introducing additional privacy concerns. Improper handling of the dataset during preprocessing, storage, or transfer could pose privacy risks [200]. To address this, robust data security measures should be implemented, including secure storage environments with access logging and monitoring, as well as strict protocols for data transfer to prevent unauthorized access. These safeguards are essential to balancing the benefits of stigmatization detection with the need to protect patient confidentiality in real-world clinical practice.

5.5.4 Comparison with Prior Work

LLMs are increasingly being utilized in healthcare for a wide range of applications, demonstrating their capacity to process and interpret complex clinical textual data [153,180,201]. These applications encompass medical question answering, where LLMs assist clinicians and patients by providing answers to health-related queries [201], and medical text summarization, which simplifies lengthy clinical documents into concise summaries for improved decision-making [201]. Additionally, LLMs are employed for biomedical evidence extraction, enabling the identification of relevant research findings and coreference resolution, ensuring clarity and coherence in clinical narratives [65]. Other notable applications include medication status tracking, which monitors patient prescriptions, and attribute extraction, which identifies key patient or clinical characteristics from unstructured text to support data-driven insights in healthcare workflows [65]. Furthermore, LLMs have been utilized to support clinicians in developing diagnostic reasoning, with research demonstrating that, when paired with appropriate prompting strategies, LLMs can effectively mimic common clinical reasoning processes without compromising diagnostic accuracy [202]. While these applications underscore the versatility of LLMs in advancing clinical workflow efficiency and decision-making, our manuscript addresses a distinct and underexplored challenge: detecting stigmatizing language in EHRs. Stigmatizing language within clinical documentation can perpetuate biases, undermine patient trust, and negatively impact the quality of care provided. By focusing on this critical issue, our work aims to promote bias-aware clinical documentation practices and foster equity in healthcare communication, addressing a significant gap in the field of medical informatics.

Sentiment analysis, a well-established area of research in NLP, has informed a wide range of clinical applications, including early studies on patient sentiment and provider communication patterns [203]. It shares both conceptual and methodological similarities with the task of detecting stigmatizing language. From a conceptual standpoint, sentiment analysis aims to classify the emotional tone of language, typically as positive, negative, or neutral. In contrast, stigmatization detection seeks to identify socially harmful or biased expressions. While the two tasks differ in their end goals, it is plausible that stigmatizing language often co-occurs with negative sentiment. Investigating whether sentiment can reliably predict stigmatization represents a promising direction for future research. However, as our dataset does not include sentiment labels, we were unable to empirically examine the potential overlap between negative sentiment and stigmatizing

content. From the perspective of methodological similarity, both tasks are commonly addressed using text classification models. Indeed, the BERT-based models employed in our study are widely used in sentiment analysis [204,205]. The primary methodological distinction lies in the labeling: sentiment analysis requires annotations for emotional tone, while stigmatization detection relies on annotations indicating the presence of stigmatizing content in text.

Extensive literature has documented the adverse effects of stigmatizing language in EHRs on the patient safety [19–21,24,139,143,145,148]. Despite this, the use of machine learning techniques to detect such language has received relatively limited attention. Sun et al. were among the first to apply machine learning in this context, developing a logistic regression classifier in a supervised fashion that achieved an F1 score of 0.935 on a proprietary dataset [36]. However, their study focused primarily on the unequal use of stigmatizing language across different demographic groups, placing less emphasis on the development and optimization of the machine learning model itself. Harrigian et al. conducted a more technical analysis using two EHR datasets, exploring the application of logistic regression and BERT models for detecting stigmatizing language [25]. Barcelona et al. developed and evaluated decision trees, random forests, and support vector machines for detecting stigmatizing language in labor and clinical notes [37].

In contrast, our research focused on evaluating and comparing the effectiveness of various zero-shot and few-shot approaches, which can operate without the need for extensive annotated data. Our findings revealed that the ICL approach significantly outperformed traditional zero-shot and few-shot text classification approaches, such as textual entailment and SetFit, while only slightly underperforming compared to the fully supervised fine-tuning approach, despite utilizing over one hundred times fewer annotated data points for training. By leveraging ICL, the reliance on large labeled datasets can be significantly diminished, enabling the development of more efficient and flexible solutions for detecting stigmatizing language across diverse clinical settings. Furthermore, our study is the first to rigorously assess the fairness dimension of machine learning classifiers specifically for stigmatizing language detection. In comparing the ICL approach to the supervised fine-tuning approach, we found that the latter is more prone to demonstrate unequal performance across specific demographic groups. However, our study did not explicitly examine methods that can be utilized to improve the fairness of models developed with a supervised fine-tuning approach, such as diversifying training datasets [206,207] and incorporating fairness constraints [190].

5.5.5 Limitations

This study is subject to several limitations that should be acknowledged. First, the analysis was conducted using a single dataset obtained from the emergency department of the Beth Israel Deaconess Medical Center in Boston, MA, US. Consequently, the findings are inherently constrained by the specific characteristics of this dataset, including its geographical setting, imbalanced patient demographics, and institutional practices. Therefore, the generalizability of the models' performance to other stigmatizing language datasets remains uncertain. Future research should aim to assess the ICL approach using larger and more diverse datasets from multiple institutions.

Second, the dataset used in this study was curated by selecting sentences that contained pre-identified keywords. Stigmatizing language can manifest both explicitly through overtly negative terms and implicitly through subtle variations in tone, phrasing, or contextual framing. While our approach may excel in the detection of explicit forms of stigmatizing language, it does not fully capture more implicit forms of stigmatization. As such, our evaluation may be limited in its ability to capture the full spectrum of stigmatizing language. We hypothesize that the ICL approach has the potential to detect subtler forms of stigma by leveraging contextual information through carefully crafted prompts, as partially demonstrated in this study. However, due to the nature of the available dataset, we were unable to empirically assess the model's generalizability to more implicit expressions of stigma. Future work should focus on developing datasets that capture a broader range of stigmatizing language. This would support more comprehensive evaluation and guide the improvement of models capable of identifying nuanced and context-dependent stigma. Moreover, the Stigma Detection Guided Prompt was intentionally designed to focus on a predefined keyword. While this strategy leverages important cues for detecting explicit forms of stigmatization, it may restrict the ICL models' ability to capture broader contextual cues or detect more implicit forms of stigma. Future studies should develop models capable of identifying more subtle forms of stigmatization without reliance on pre-identified terms.

Third, our current binary labeling framework does not explicitly address two important nuances. It does not explicitly consider ambiguous cases where the interpretation of stigmatization is context-dependent. The current dataset employs binary labels, stigmatizing or non-stigmatizing, without accounting for sentences that may require additional context for accurate classification.

Future research could address this issue by introducing an additional classification category, such as “Uncertain/Requires Further Context” similar to the CheXpert dataset [208]. This would enable models to flag ambiguous cases for human review or request additional contextual information to support model predictions. In addition, while collapsing distinct categories of stigmatizing language into a binary label (stigmatizing vs. non-stigmatizing) enabled broader applicability and methodological comprehensiveness in this study, it may have obscured important nuances related to the impacts of different types of stigma. Prior literature has highlighted the clinical consequences of stigmatizing language [42,144], but the differential effects of specific stigma dimensions remain underexplored. In future work, we intend to systematically investigate whether specific types of stigma are differentially associated with key clinical outcomes.

Fourth, we did not examine the effectiveness of fine-tuning decoder-based models, such as LLAMA, due to the limited availability of annotated data, and as a result, their efficacy in detecting stigmatizing language remains underexplored. Future research should investigate whether fine-tuning decoder-based models can enhance performance in settings where large-scale annotated datasets are available.

Fifth, the fairness evaluation was limited to the sex, age, and race categories available in the dataset and to disparities in TPR, FPR, and F1-score. Consequently, the findings characterize observed performance differences across these predefined subgroups but should not be interpreted as establishing fairness across all patient populations. The analysis did not examine other potentially relevant attributes or intersectional subgroups defined by combinations of demographic characteristics. Furthermore, because the dataset consisted of discharge-summary sentences from a single clinical data source, the findings may not generalize to other documentation types, institutions, or clinical settings. Future research should incorporate additional patient characteristics, intersectional analyses, complementary fairness measures, and external validation across diverse documentation contexts and healthcare settings.

[207][209]Lastly, we hypothesize that machine learning solutions for detecting stigmatizing language will be employed as decision-support systems within clinical workflows rather than as decision-making systems due to the necessity of human oversight and the importance of ensuring that narratives within EHRs are validated by clinicians. While this study discusses several potential ethical implications for the clinical deployment of the stigmatization detection model. This study

did not address several critical factors essential for effective human-AI collaboration in this context, such as varying levels of automation, system usability, and the potential for automation bias. Future research should explore these factors to ensure the successful integration of such systems into clinical workflows, thereby optimizing decision-making processes and enhancing the efficiency and effectiveness of stigmatizing language detection. In addition, future research should prospectively evaluate the stigmatizing-language detector within clinical documentation workflows. Beyond conventional measures of detection performance, evaluation should examine whether flagged language is considered clinically relevant and useful by clinicians, the frequency of unnecessary or nonactionable alerts, clinician acceptability and perceived alert burden, and the system's ability to distinguish stigmatizing language from wording whose meaning depends strongly on the surrounding clinical context. Comparisons with current documentation-review practices could also determine whether the system supports more appropriate language use without disrupting clinical workflows or encouraging excessive reliance on automated alerts.

5.6 Conclusions

This study demonstrated the effectiveness of ICL as a data-efficient approach for detecting stigmatizing language in electronic health records, an important source of non-physical patient harm. By enabling accurate detection under limited annotation conditions, the proposed approach supports scalable analysis of clinical documentation, addressing a key barrier to monitoring documentation-related patient safety risks. The results showed that ICL outperformed traditional zero-shot and few-shot methods, while achieving performance comparable to supervised fine-tuning with substantially fewer labeled examples. In addition, the proposed Stigma Detection Guided Prompt improved model sensitivity to the contextual and linguistic characteristics of stigmatizing language, highlighting the importance of domain-informed prompting strategies in safety-critical applications. Notably, ICL demonstrated more equitable performance across demographic groups, suggesting its potential to reduce bias amplification in automated monitoring systems. Collectively, these findings establish ICL as a practical and scalable framework for identifying stigmatizing language in clinical narratives. By enabling systematic detection of non-physical harm embedded in healthcare documentation, this work contributes to strengthening institutional patient safety monitoring and supports the development of more equitable and patient-centered care systems.

Chapter 6

Conclusions and Future Directions

6 Conclusions and Future Directions

6.1 Overview and Synthesis of the Thesis

This dissertation investigated the role of machine learning in enabling scalable and systematic analysis of healthcare narratives to strengthen institutional patient safety monitoring. As healthcare systems increasingly rely on narrative data sources, including patient safety event (PSE) reports and clinical documentation, the ability to extract actionable insights from these unstructured data has become essential. However, the growing volume and complexity of healthcare narratives present substantial challenges for traditional manual review processes, limiting the timely identification of safety risks and constraining the effectiveness of institutional safety surveillance.

To address these challenges, this dissertation developed and evaluated machine learning approaches across three key dimensions of patient safety monitoring: classification, prioritization, and detection of patient harm. First, this work demonstrated that transformer-based models leveraging contextual text representations can substantially improve the accuracy and reliability of PSE classification. By enabling more consistent and precise categorization of safety events, these models support the effective routing of incident reports to appropriate patient safety specialists, thereby facilitating timely investigation and intervention.

Second, this dissertation addressed the critical need for prioritizing high-severity events under constrained institutional resources. By reformulating severity assessment as a ranking problem, rather than a categorical classification task, this work introduced a learning-to-rank framework

that more closely aligns with real-world triage workflows. The proposed ranking-based models consistently outperformed traditional classification approaches and existing baseline practices, enabling more effective identification and prioritization of serious safety events for further review.

Third, this dissertation investigated the detection of stigmatizing language in clinical documentation as a form of non-physical, preventable patient harm. Through the development of a data-efficient in-context learning approach, this work demonstrated that large language models can achieve competitive performance under limited annotation conditions. These findings support the feasibility of systematically identifying documentation-related risks that may influence clinical decision-making and patient outcomes.

Taken together, this dissertation demonstrates that machine learning methods can support scalable analysis of healthcare narratives by improving the classification of patient safety events, enabling effective prioritization of high-severity incidents, and facilitating the detection of stigmatizing language in clinical documentation. These contributions align with the central thesis that machine learning can strengthen institutional patient safety monitoring by augmenting the analysis of complex and large-scale healthcare narrative data.

6.2 Summary of Contributions

This dissertation advances the use of machine learning for scalable patient safety monitoring through the analysis of healthcare narratives. The key contributes are summarized below.

6.2.1 Advancing Reliable and Explainable AI for Patient Safety Event Classification

- We demonstrated that contextual text representations derived from transformer-based models significantly improved PSE report classification accuracy compared with static representations (i.e., 75.4% vs 66.7%).
- We showed that classification errors were not purely random, but instead arose from ambiguity in event taxonomy and narrative interpretation, revealing fundamental challenges in defining and distinguishing event types.

- We established the critical role of explainability (via LIME) for practical deployment, as model predictions may rely on spurious or non-causal tokens, necessitating human validation.
- We designed and implemented a prototype interface that integrates machine learning with explainability, enabling the system to recommend top event types while preserving human decision authority within reporting workflows.
- We further demonstrated that incorporating interpretable explanations within a human–AI collaboration framework improves classification accuracy when users are provided with transparent model reasoning (i.e., highlighting relevant word for predicted event type).
- Finally, we provided empirical evidence that fully automated classification was not appropriate for this task, due to model uncertainty and reliance on non-causal features, reinforcing the need for human-in-the-loop decision support systems.

6.2.2 Enhancing Institutional Severity Triage Through Ranking-Based Machine Learning

- We identified a fundamental mismatch between classification-based modeling approaches and real-world patient safety triage workflows, where the primary objective is to prioritize high-severity events under resource constraints rather than assign categorical labels.
- We reformulated severity assessment as a learning-to-rank problem, enabling models to directly learn the relative ordering of patient safety events based on their severity and better align with operational decision-making processes.
- We demonstrated that ranking-based models (BERT-RA) significantly outperformed classification-based approaches, achieving a 23.0% relative improvement in Average Precision (0.91 vs 0.74; $P < 0.0008$) and superior performance across the majority of evaluation metrics.
- We showed that performance gains were substantially larger at broader evaluation cutoffs (e.g., $k = 100\text{--}200$), which are most relevant for real-world triage, highlighting that ranking models more effectively support high-volume review settings.

- We demonstrated that reporter-assigned severity is highly inconsistent and performs substantially worse than model-based prioritization, with the ranking model achieving over 100% relative improvement across multiple evaluation metrics.
- Finally, we designed an interactive triage interface that presents ranked PSEs along with contextual information, illustrating how ranking outputs can be integrated into real-time clinical workflows to support decision-making.

6.2.3 Data-Efficient Detection of Stigmatizing Language via In-Context Learning

- We identified the reliance on large annotated datasets as a key limitation of existing approaches for stigmatizing language detection, motivating the use of in-context learning (ICL) as a data-efficient alternative.
- We systematically evaluated ICL as a data-efficient alternative and demonstrated that it significantly outperforms traditional zero-shot approaches (textual entailment), achieving an 18.7% improvement in F1-score (i.e., 0.858 vs 0.723).
- We further showed that ICL consistently outperforms few-shot learning approaches (SetFit), with performance improvements ranging from 12.3% to 21.4% across different annotation settings, highlighting its effectiveness in low-data regimes.
- We demonstrated that ICL achieves near-supervised performance under data-scarce conditions, attaining an F1-score of 0.901 using only 32 labeled instances, compared with 0.931 for supervised fine-tuning trained on 3543 labeled instances (3.2% difference).
- We introduced a novel prompting strategy, the Stigma Detection Guided Prompt, which incorporates domain-specific linguistic patterns and keywords, and demonstrated that prompt design is a critical determinant of model performance in detecting subtle and context-dependent stigmatizing language.
- Finally, we performed a comprehensive fairness evaluation and showed that supervised fine-tuning models exhibit greater disparities across demographic groups (sex, age, and

race), whereas ICL approaches consistently achieve more balanced performance across protected attributes.

6.3 Limitations and Future Research

This section outlines the main limitations of the present work and identifies directions for future research.

6.3.1 Data Quality, Subjectivity, and Generalizability

The findings of this dissertation are closely tied to the characteristics of the data used across the three studies, particularly in terms of label definition, annotation practices, and institutional context. Because these datasets are derived from real-world healthcare workflows, they inherently reflect the variability and complexity of how patient safety information is generated, interpreted, and documented.

In the context of patient safety event (PSE) classification (Chapter 3), event type labels are assigned through human reporting and review processes within incident reporting systems. While these processes are essential for institutional safety monitoring, they involve interpretation of clinical narratives using predefined taxonomies that may contain overlapping or loosely defined categories [46,91]. As a result, some degree of variability in labeling is expected, especially in cases where event descriptions are ambiguous or span multiple categories. This variability reflects the complexity of real-world safety reporting and may influence model training and evaluation. In severity triage (Chapter 4), the labels used for model development are shaped by institution-specific practices, including reporter-assigned severity and adjudicated classifications. These labels are embedded within local workflows and reflect how individual institutions define and prioritize safety events. Consequently, the patterns learned by the models may be partially influenced by these local standards. Differences in severity definitions, reporting behaviors, and triage protocols across institutions may therefore affect the extent to which the proposed models generalize to other settings. For stigmatizing language detection (Chapter 5), annotation involves an additional layer of complexity due to the context-dependent and nuanced nature of language

[25,36]. The identification of stigmatizing expressions often depends on subtle linguistic cues and clinical interpretation, which may vary across annotators and institutional environments. Furthermore, the dataset used in this study reflects a specific documentation context (i.e., discharge summary), and differences in language use, clinical specialties, and patient populations may influence model performance when applied elsewhere.

These considerations highlight opportunities for further strengthening the robustness and generalizability of machine learning approaches for patient safety monitoring. Developing more consistent and well-defined annotation frameworks could help reduce variability and improve label reliability [2,45,67]. In parallel, multi-center validation studies would enable systematic evaluation of model performance across diverse institutional settings, providing insights into how differences in data generation and clinical workflows affect model behavior [210,211].

6.3.2 Expanding the Detection of Non-Physical Patient Harm

This dissertation focuses on stigmatizing language in clinical documentation as a measurable and operationalizable form of non-physical patient harm. While this provides a concrete entry point for systematic analysis, stigmatizing language represents only one manifestation of how non-physical harm can arise within healthcare [40]. Such harm may also be conveyed through more subtle expressions reflected in tone, attitude, framing, and implicit assumptions across both written and verbal clinical communication. For instance, communication that questions patient credibility, conveys skepticism, or reflects dismissive or judgmental attitudes may influence clinical reasoning and decision-making, even in the absence of explicitly stigmatizing terms. These patterns may occur not only in electronic health records, but also in verbal exchanges, clinical handoffs, and other forms of communication that shape how patient information is interpreted and acted upon [150,212]. As a result, the scope of this work captures only a subset of documentation- and communication-related risks that may affect patient safety.

Expanding the detection of non-physical patient harm therefore represents an important direction for future research. One key area involves developing broader conceptual frameworks that define and categorize different forms of communication-related harm beyond stigmatizing language. This may include identifying patterns such as dismissive language, implicit bias in clinical reasoning, or framing effects that systematically influence how patients are perceived and treated [24,154].

From a methodological perspective, advancing this line of work will require moving beyond text-only analysis toward models that can capture the full context of clinical communication. Non-physical patient harm is often conveyed through tone, speech patterns, and interaction dynamics, which are not fully observable in written documentation. Future research may therefore explore multimodal approaches that integrate textual, audio, and potentially visual signals to better characterize how communication contributes to patient safety risks [213]. For example, incorporating speech-based features, as well as interaction-level patterns in clinical encounters, may provide a more comprehensive understanding of how subtle forms of bias or dismissiveness manifest in practice. Such multimodal models could enable a more complete representation of clinician–patient communication and support more comprehensive monitoring of non-physical patient harm.

In addition, the development of richer datasets and annotation frameworks that reflect a wider range of communication-related risks will be essential [16,18]. Establishing clearer definitions and guidelines for different forms of non-physical harm can improve annotation consistency and support more robust model evaluation. Together, these efforts would extend the current work toward more comprehensive monitoring of how communication practices within healthcare systems may contribute to patient safety risks.

6.3.3 Prospective Evaluation, Deployment, and Continuous Adaptation

The models developed in this dissertation were evaluated using retrospective datasets, which provide a controlled setting for assessing performance but may not fully reflect real-world clinical use. While retrospective evaluation is an essential step in model development, the practical impact of these approaches depends on how they perform when integrated into live clinical workflows. In patient safety monitoring, where decisions are made under time constraints and involve multiple stakeholders, the effectiveness of machine learning models must be assessed in the context of actual use rather than offline evaluation alone.

Prospective evaluation therefore represents an important next step for this line of work. Future studies may examine how the proposed models influence workflow efficiency, decision-making processes, and the identification of high-risk events when deployed within incident reporting systems or clinical documentation environments [131]. For example, evaluating how patient safety specialists interact with classification or ranking recommendations, or how clinicians respond to

model-assisted feedback on documentation, would provide valuable insight into the practical utility of these approaches. Such evaluations would also help identify potential unintended consequences, including over-reliance on model outputs or changes in reporting behavior.

The deployment of these models also raises important considerations regarding integration with existing systems. As demonstrated in this dissertation, effective use of machine learning in patient safety monitoring requires alignment with institutional workflows and the preservation of human decision authority [45]. Future work should therefore focus on designing systems that support seamless integration, including user interfaces that present model outputs in an interpretable and actionable manner, as well as mechanisms that allow users to provide feedback.

The rapid evolution of artificial intelligence, particularly large language models (LLMs), presents both opportunities and challenges for patient safety monitoring systems. New models are frequently introduced with improved capabilities, making it increasingly difficult for healthcare institutions to determine when and how to adopt updated models in practice [80,214,215]. Unlike traditional machine learning systems with relatively stable development cycles, LLM-based approaches evolve at a much faster pace, requiring continuous reassessment of model performance, reliability, and suitability for safety-critical applications.

Future work should therefore focus on developing systematic strategies for incorporating new models while maintaining consistency and trust in deployed systems. One key direction involves establishing standardized evaluation pipelines that allow newly released models to be benchmarked against existing models using consistent datasets and performance metrics. Such pipeline can help determine whether newer models provide meaningful improvements in tasks such as classification, ranking, or stigmatizing language detection before they are integrated into clinical workflows.

In addition, model versioning and controlled deployment strategies will be essential. Rather than replacing models immediately upon the release of newer versions, healthcare systems may adopt staged deployment approaches, such as parallel testing or shadow deployment, where new models are evaluated alongside existing systems without directly influencing decision-making [216,217]. This allows for careful comparison of performance and identification of potential risks prior to full adoption.

6.4 Concluding Remarks

This dissertation has demonstrated that advanced natural language processing and machine learning methods can enable scalable and systematic analysis of healthcare narratives to strengthen institutional patient safety monitoring. By addressing key challenges in classification, prioritization, and detection of patient safety risks, this work provides a comprehensive examination of how machine learning methods can be applied to support safety-related tasks within healthcare systems.

Across the three studies presented, this research shows that machine learning methods are most effective when they are aligned with real-world clinical workflows and designed to augment, rather than replace, human expertise. In the context of patient safety event classification, contextual language models improve predictive performance, but their integration with explainability techniques and human oversight is essential for reliable deployment. For severity triage, the reformulation of the problem as a ranking task demonstrates that aligning model objectives with operational decision-making processes leads to more effective prioritization of high-risk events. For stigmatizing language detection, the findings highlight the importance of incorporating domain-specific knowledge, showing that the guided prompting strategy can improve the identification of subtle and context-dependent expressions of stigmatizing language.

More broadly, this dissertation underscores that the value of machine learning in healthcare extends beyond predictive accuracy. Effective patient safety monitoring requires systems that are interpretable, workflow-compatible, and capable of operating under real-world constraints such as limited resources and evolving data environments. The findings presented in this work illustrate that careful consideration of task formulation, model transparency, and human–AI collaboration is critical for translating methodological advances into meaningful improvements in healthcare practice.

At the same time, this work highlights several important limitations. These include challenges related to annotation quality and subjectivity, uncertainty in generalizability across institutions, and the need to expand detection beyond stigmatizing language to other forms of non-physical harm. In addition, prospective evaluation in real-world settings remains necessary to assess practical impact, and the rapid evolution of large language models introduces ongoing challenges for reliable and sustained deployment.

Across the three applications, several common design lessons emerged. First, the intended healthcare decision should guide the formulation of the machine learning task: event-type assignment was formulated as classification, severity-based triage as ranking, and stigmatizing-language identification as data-efficient detection. Second, evaluation measures should reflect the intended use of each system. Accordingly, classification accuracy and agreement were emphasized for event assignment, ranking metrics and the retrieval of high-severity events were prioritized for severity triage, and detection performance and subgroup disparities were examined for stigmatizing-language identification. Third, the appropriate role of human review depends on the consequences and uncertainty associated with each decision. Across all three applications, the models were therefore positioned as decision-support tools that augment expert judgment through recommendations, prioritization, or alerts, while preserving human authority over final classification, investigation, and documentation decisions.

In conclusion, this dissertation provides empirical and methodological contributions toward the development of scalable, data-driven approaches for patient safety monitoring. By demonstrating how machine learning can be applied to classify patient safety events, prioritize high-severity incidents, and detect stigmatizing language in clinical documentation, this work advances the systematic identification of patient safety risks and strengthens institutional capacity for timely and effective safety monitoring.

References or Bibliography

1. Kohn LT, Corrigan JM, Donaldson MS, editors. To Err is Human: Building a Safer Health System. Washington (DC): National Academies Press (US); 2000. PMID:25077248 ISBN:978-0-309-06837-6
2. Wang Y, Coiera E, Runciman W, Magrabi F. Using multiclass classification to automate the identification of patient safety incident reports by type and severity. BMC Med Inform Decis Mak 2017 June 12;17(1):84. doi: 10.1186/s12911-017-0483-8
3. World Health Organization, Safety WP. Conceptual framework for the international classification for patient safety version 1.1: final technical report January 2009. World Health Organization; 2010. Report No.: WHO/IER/PSP/2010.2. Available from: <https://apps.who.int/iris/handle/10665/70882> [accessed Mar 8, 2023]
4. Makary MA, Daniel M. Medical error—the third leading cause of death in the US. BMJ British Medical Journal Publishing Group; 2016 May 3;353:i2139. PMID:27143499
5. Improvement still needed as rate of hospital harm remains steady | CIHI. Available from: <https://www.cihi.ca/en/news/improvement-still-needed-as-rate-of-hospital-harm-remains-steady> [accessed Jan 12, 2026]
6. Slawomirski L, Klazinga N. The economics of patient safety: From analysis to action. OECD Health Working Papers OECD Publishing; 2022 Aug 12; Available from: <https://ideas.repec.org/p/oec/elsaad/145-en.html> [accessed Apr 11, 2025]
7. Hewitt TA, Chreim S. Fix and forget or fix and report: a qualitative study of tensions at the front line of incident reporting. BMJ Qual Saf BMJ Publishing Group Ltd; 2015 May 1;24(5):303–310. PMID:25749025
8. Ward JK, Armitage G. Can patients report patient safety incidents in a hospital setting? A systematic review. BMJ Qual Saf BMJ Publishing Group Ltd; 2012 Aug 1;21(8):685–699. PMID:22562875
9. Albolino S, Tartaglia R, Bellandi T, Amicosante A, Bianchini E, Biggeri A. Patient safety and incident reporting: survey of Italian healthcare workers. Quality & safety in health care 2010 Oct 1;19 Suppl 3:i8-12. doi: 10.1136/qshc.2009.036376
10. NHS England. Organisation patient safety incident report up to March 2022. 2022. Available from: <https://www.england.nhs.uk/publication/organisation-patient-safety-incident-report-up-to-march-2022/> [accessed Mar 8, 2023]
11. Vincent C. Incident reporting and patient safety. BMJ 2007 Jan 13;334(7584):51. PMID:17218667

12. Singh G, Patel RH, Vaqar S, Boster J. Root Cause Analysis and Medical Error Prevention. StatPearls Treasure Island (FL): StatPearls Publishing; 2025. PMID:34033400
13. Ong M-S, Magrabi F, Coiera E. Automated categorisation of clinical incident reports using statistical text classification. *BMJ Quality & Safety* 2010 Dec 1;19(6):e55–e55. doi: 10.1136/qshc.2009.036657
14. Nebeker JR, Barach P, Samore MH. Clarifying adverse drug events: a clinician's guide to terminology, documentation, and reporting. *Ann Intern Med* 2004 May 18;140(10):795–801. PMID:15148066
15. Chopard D, Treder MS, Corcoran P, Ahmed N, Johnson C, Busse M, Spasic I. Text Mining of Adverse Events in Clinical Trials: Deep Learning Approach. *JMIR Medical Informatics* JMIR Publications Inc., Toronto, Canada; 2021 Dec 24;9(12):e28632. doi: 10.2196/28632
16. Mitchell P, Cribb A, Entwistle V. Patient Safety and the Question of Dignitary Harms. Available from: <https://dx.doi.org/10.1093/jmp/jhac035> [accessed Mar 5, 2026]
17. Non-physical Harm Inclusion Shapes Patient Journey - UHN. Available from: <https://www.uhn.ca/corporate/News/Pages/uhn-patient-safety-journey-evolving-inclusion-non-physical-harm.aspx> [accessed Mar 5, 2026]
18. Sokol-Hessner L, Folcarelli PH, Sands KEF. Emotional harm from disrespect: the neglected preventable harm. *BMJ Qual Saf* BMJ Publishing Group Ltd; 2015 Sept 1;24(9):550–553. PMID:26085331
19. Cox C, Fritz Z. Presenting complaint: use of language that disempowers patients. *BMJ British Medical Journal Publishing Group*; 2022 Apr 27;377:e066720. PMID:35477529
20. Katz IT, Ryu AE, Onuegbu AG, Psaros C, Weiser SD, Bangsberg DR, Tsai AC. Impact of HIV-related stigma on treatment adherence: systematic review and meta-synthesis. *Journal of the International AIDS Society Wiley*; 2013;16(3Suppl 2). PMID:24242258
21. Dickinson JK, Maryniuk MD. Building Therapeutic Relationships: Choosing Words That Put People First. *Clin Diabetes* 2017 Jan;35(1):51–54. PMID:28144046
22. Smith RA. Language of the Lost: An Explication of Stigma Communication. *Communication Theory* 2007 Nov 1;17(4):462–485. doi: 10.1111/j.1468-2885.2007.00307.x
23. Beach MC, Saha S, Park J, Taylor J, Drew P, Plank E, Cooper LA, Chee B. Testimonial Injustice: Linguistic Bias in the Medical Records of Black Patients and Women. *J Gen Intern Med* 2021 June;36(6):1708–1714. PMID:33754318
24. P Goddu A, O'Connor KJ, Lanzkron S, Saheed MO, Saha S, Peek ME, Haywood C, Beach MC. Do Words Matter? Stigmatizing Language and the Transmission of Bias in the Medical Record. *J Gen Intern Med* 2018 May;33(5):685–691. PMID:29374357

25. Harrigian K, Zirikly A, Chee B, Ahmad A, Links A, Saha S, Beach MC, Dredze M. Characterization of Stigmatizing Language in Medical Records. In: Rogers A, Boyd-Graber J, Okazaki N, editors. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) Toronto, Canada: Association for Computational Linguistics; 2023. p. 312–329. doi: 10.18653/v1/2023.acl-short.28
26. Chen H, Alfred M, Cohen E. Efficient Detection of Stigmatizing Language in Electronic Health Records via In-Context Learning: Comparative Analysis and Validation Study. JMIR Medical Informatics JMIR Publications Inc., Toronto, Canada; 2025 Aug 18;13(1):e68955. doi: 10.2196/68955
27. Fong A, Hettinger AZ, Ratwani RM. Exploring methods for identifying related patient safety events using structured and unstructured data. Journal of Biomedical Informatics 2015 Dec 1;58:89–95. doi: 10.1016/j.jbi.2015.09.011
28. Boxley C, Fujimoto M, Ratwani RM, Fong A. A text mining approach to categorize patient safety event reports by medication error type. Sci Rep Nature Publishing Group; 2023 Oct 26;13(1):18354. doi: 10.1038/s41598-023-45152-w
29. Biannual Incident Report - Clinical Excellence Commission. Available from: <https://www.cec.health.nsw.gov.au/Review-incidents/Biannual-Incident-Report> [accessed May 16, 2023]
30. Wang Y, Coiera E, Magrabi F. Using convolutional neural networks to identify patient safety incident reports by type and severity. J Am Med Inform Assoc 2019 Dec 1;26(12):1600–1608. PMID:31730700
31. Fong A, Behzad S, Pruitt Z, Ratwani RM. A Machine Learning Approach to Reclassifying Miscellaneous Patient Safety Event Reports. J Patient Saf 2021 Dec 1;17(8):e829–e833. PMID:32555052
32. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171–4186. doi: 10.18653/v1/N19-1423
33. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv; 2019. doi: 10.48550/arXiv.1907.11692
34. Ong M-S, Magrabi F, Coiera E. Automated identification of extreme-risk events in clinical incident reports. J Am Med Inform Assoc 2012 June;19(e1):e110–e118. PMID:22237865
35. Public Hospitals Act, R.S.O. 1990, c. P.40 | ontario.ca. Available from: <https://www.ontario.ca/laws/statute/90p40> [accessed Apr 14, 2025]

36. Sun M, Oliwa T, Peek ME, Tung EL. Negative Patient Descriptors: Documenting Racial Bias In The Electronic Health Record. *Health Affairs Health Affairs*; 2022 Feb;41(2):203–211. doi: 10.1377/hlthaff.2021.01423
37. Barcelona V, Scharp D, Moen H, Davoudi A, Idnay BR, Cato K, Topaz M. Using Natural Language Processing to Identify Stigmatizing Language in Labor and Birth Clinical Notes. *Matern Child Health J* 2024 Mar 1;28(3):578–586. doi: 10.1007/s10995-023-03857-4
38. Cooper J, Williams H, Hibbert P, Edwards A, Butt A, Wood F, Parry G, Smith P, Sheikh A, Donaldson L, Carson-Stevens A. Classification of patient-safety incidents in primary care. *Bull World Health Organ* 2018 July 1;96(7):498–505. PMID:29962552
39. Chen H, Cohen E, Alfred M. Examining the development, effectiveness, and limitations of computer-aided diagnosis systems for retained surgical items detection: a systematic review. *Ergonomics Taylor & Francis*; 2025 Apr 10; doi: 10.1080/00140139.2025.2487558
40. Street RL, Makoul G, Arora NK, Epstein RM. How does communication heal? Pathways linking clinician-patient communication to health outcomes. *Patient Educ Couns* 2009 Mar;74(3):295–301. PMID:19150199
41. Meier-Diedrich E, Blease C, Heinze M, Wördemann J, Schwarz J. Changes in Documentation After Implementing Open Notes in Mental Health Care: Pre-Post Mixed Methods Study. *Journal of Medical Internet Research JMIR Publications Inc., Toronto, Canada*; 2025 Sept 3;27(1):e72667. doi: 10.2196/72667
42. Kelly JF, Westerhoff CM. Does it matter how we refer to individuals with substance-related conditions? A randomized study of two commonly used terms. *Int J Drug Policy* 2010 May;21(3):202–207. PMID:20005692
43. Weiner SG, Lo Y-C, Carroll AD, Zhou L, Ngo A, Hathaway DB, Rodriguez CP, Wakeman SE. The Incidence and Disparities in Use of Stigmatizing Language in Clinical Notes for Patients with Substance Use Disorder. *J Addict Med* 2023;17(4):424–430. PMID:37579100
44. Herzer KR, Mirrer M, Xie Y, Steppan J, Li M, Jung C, Cover R, Doyle PA, Mark LJ. Patient Safety Reporting Systems: Sustained Quality Improvement Using a Multidisciplinary Team and “Good Catch” Awards. *Jt Comm J Qual Patient Saf* 2012 Aug;38(8):339–347. PMID:22946251
45. Chen H, Cohen E, Wilson D, Alfred M. A Machine Learning Approach with Human-AI Collaboration for Automated Classification of Patient Safety Event Reports: Algorithm Development and Validation Study. *JMIR Human Factors JMIR Publications Inc., Toronto, Canada*; 2024 Jan 25;11(1):e53378. doi: 10.2196/53378
46. Brubacher J, Hunte G, Hamilton L, Taylor A. Barriers to and Incentives for Safety Event Reporting in Emergency Departments. *hcq* 2011 July 15;14(3):57–65. doi: 10.12927/hcq.2011.22491

47. Islam S. Automating Patient Safety Event Report Classification using Natural Language Processing and Machine Learning. 2024 Nov; Available from: <http://hdl.handle.net/1807/140597> [accessed June 24, 2025]
48. Young IJB, Luz S, Lone N. A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. *International Journal of Medical Informatics* 2019 Dec 1;132:103971. doi: 10.1016/j.ijmedinf.2019.103971
49. Quality & Patient Safety - Overview. Available from: https://www.uhn.ca/corporate/AboutUHN/Quality_Patient_Safety [accessed Mar 10, 2026]
50. Carayon P, Xie A, Kianfar S. Human factors and ergonomics as a patient safety practice. *BMJ Qual Saf* BMJ Publishing Group Ltd; 2014 Mar 1;23(3):196–205. PMID:23813211
51. Rodziewicz TL, Houseman B, Vaqar S, Hipskind JE. Medical Error Reduction and Prevention. StatPearls Treasure Island (FL): StatPearls Publishing; 2026. PMID:29763131
52. Siebers P, Janiesch C, Zschech P. A Survey of Text Representation Methods and Their Genealogy. *IEEE Access* 2022;10:96492–96513. doi: 10.1109/ACCESS.2022.3205719
53. 7.2. Feature extraction. scikit-learn. Available from: https://scikit-learn/stable/modules/feature_extraction.html [accessed Mar 13, 2026]
54. TfidfVectorizer. scikit-learn. Available from: https://scikit-learn/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html [accessed Mar 13, 2026]
55. Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* Doha, Qatar: Association for Computational Linguistics; 2014. p. 1532–1543. doi: 10.3115/v1/D14-1162
56. Liu Q, Kusner MJ, Blunsom P. A Survey on Contextual Embeddings. 2020 Mar 16; doi: 10.48550/arXiv.2003.07278
57. Qin L, Chen Q, Feng X, Wu Y, Zhang Y, Li Y, Li M, Che W, Yu PS. Large Language Models Meet NLP: A Survey. *arXiv*; 2025. doi: 10.48550/arXiv.2405.12819
58. Taha K, Yoo PD, Yeun C, Homouz D, Taha A. A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights. *Computer Science Review* 2024 Nov 1;54:100664. doi: 10.1016/j.cosrev.2024.100664
59. Gasparetto A, Marcuzzo M, Zangari A, Albarelli A. A Survey on Text Classification Algorithms: From Text to Predictions. *Information Multidisciplinary Digital Publishing Institute*; 2022 Feb;13(2):83. doi: 10.3390/info13020083

60. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention Is All You Need. arXiv; 2023. doi: 10.48550/arXiv.1706.03762
61. Li Q, Peng H, Li J, Xia C, Yang R, Sun L, Yu PS, He L. A Survey on Text Classification: From Traditional to Deep Learning. ACM Trans Intell Syst Technol 2022 Apr 8;13(2):31:1-31:41. doi: 10.1145/3495162
62. Lin J, Nogueira R, Yates A. Pretrained Transformers for Text Ranking: BERT and Beyond. arXiv; 2021. doi: 10.48550/arXiv.2010.06467
63. Lee J, Bernier-Colborne G, Maharaj T, Vajjala S. Methods, Applications, and Directions of Learning-to-Rank in NLP Research. In: Duh K, Gomez H, Bethard S, editors. Findings of the Association for Computational Linguistics: NAACL 2024 Mexico City, Mexico: Association for Computational Linguistics; 2024. p. 1900–1917. doi: 10.18653/v1/2024.findings-naacl.123
64. Liu T-Y. Learning to Rank for Information Retrieval. Foundations and Trends® in Information Retrieval 2009 June 27;3(3):225–331. doi: 10.1561/15000000016
65. Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing: Algorithm Development and Validation Study. JMIR Medical Informatics JMIR Publications Inc., Toronto, Canada; 2024 Apr 8;12(1):e55318. doi: 10.2196/55318
66. Barcelona V, Scharp D, Idnay BR, Moen H, Cato K, Topaz M. Identifying stigmatizing language in clinical documentation: A scoping review of emerging literature. PLoS One 2024;19(6):e0303653. PMID:38941299
67. Evans HP, Anastasiou A, Edwards A, Hibbert P, Makeham M, Luz S, Sheikh A, Donaldson L, Carson-Stevens A. Automated classification of primary care patient safety incident report content and severity using supervised machine learning (ML) approaches. Health Informatics J 2020 Dec;26(4):3123–3139. PMID:30843455
68. Bottrighi A, Pennisi M, Roveta A, Massarino C, Cassinari A, Betti M, Bolgeo T, Bertolotti M, Rava E, Maconi A. A machine learning approach for predicting high risk hospitalized patients with COVID-19 SARS-Cov-2. BMC Med Inform Decis Mak 2022 Dec 28;22(1):340. PMID:36578017
69. Vyas A, Raman S, Sen S, Ramasamy K, Rajalakshmi R, Mohan V, Raman R. Machine Learning-Based Diagnosis and Ranking of Risk Factors for Diabetic Retinopathy in Population-Based Studies from South India. Diagnostics (Basel) 2023 June 16;13(12):2084. PMID:37370980
70. Verma AA, Stukel TA, Colacci M, Bell S, Ailon J, Friedrich JO, Murray J, Kuzulugil S, Yang Z, Lee Y, Pou-Prom C, Mamdani M. Clinical evaluation of a machine learning–based early warning system for patient deterioration. CMAJ 2024 Sept 16;196(30):E1027–E1037. PMID:39284602

71. 1.10. Decision Trees. scikit-learn. Available from: <https://scikit-learn/stable/modules/tree.html> [accessed Mar 16, 2026]
72. 1.4. Support Vector Machines. scikit-learn. Available from: <https://scikit-learn/stable/modules/svm.html> [accessed Mar 16, 2026]
73. LogisticRegression. scikit-learn. Available from: https://scikit-learn/stable/modules/generated/sklearn.linear_model.LogisticRegression.html [accessed Mar 16, 2026]
74. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, Compas C, Martin C, Flores MG, Zhang Y, Magoc T, Harle CA, Lipori G, Mitchell DA, Hogan WR, Shenkman EA, Bian J, Wu Y. GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records. arXiv; 2022. doi: 10.48550/arXiv.2203.03540
75. Wang W, Zheng VW, Yu H, Miao C. A Survey of Zero-Shot Learning: Settings, Methods, and Applications. *ACM Trans Intell Syst Technol* 2019 Jan 16;10(2):13:1-13:37. doi: 10.1145/3293318
76. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large Language Models are Zero-Shot Reasoners. arXiv; 2023. doi: 10.48550/arXiv.2205.11916
77. Yin W, Hay J, Roth D. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In: Inui K, Jiang J, Ng V, Wan X, editors. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* Hong Kong, China: Association for Computational Linguistics; 2019. p. 3914–3923. doi: 10.18653/v1/D19-1404
78. Tunstall L, Reimers N, Jo UES, Bates L, Korat D, Wasserblat M, Pereg O. Efficient Few-Shot Learning Without Prompts. arXiv; 2022. doi: 10.48550/arXiv.2209.11055
79. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language Models are Few-Shot Learners. arXiv; 2020. doi: 10.48550/arXiv.2005.14165
80. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, Avila R, Babuschkin I, Balaji S, Balcom V, Baltescu P, Bao H, Bavarian M, Belgum J, Bello I, Berdine J, Bernadett-Shapiro G, Berner C, Bogdonoff L, Boiko O, Boyd M, Brakman A-L, Brockman G, Brooks T, Brundage M, Button K, Cai T, Campbell R, Cann A, Carey B, Carlson C, Carmichael R, Chan B, Chang C, Chantzis F, Chen D, Chen S, Chen R, Chen J, Chen M, Chess B, Cho C, Chu C, Chung HW, Cummings D, Currier J, Dai Y, Decareaux C, Degry T, Deutsch N, Deville D, Dhar A, Dohan D, Dowling S, Dunning S, Ecoffet A, Eleti A, Eloundou T, Farhi D, Fedus L, Felix N, Fishman SP, Forte J, Fulford I, Gao L, Georges E, Gibson C, Goel V, Gogineni T, Goh G, Gontijo-Lopes R, Gordon J, Grafstein M, Gray S, Greene R, Gross J, Gu SS, Guo

Y, Hallacy C, Han J, Harris J, He Y, Heaton M, Heidecke J, Hesse C, Hickey A, Hickey W, Hoeschele P, Houghton B, Hsu K, Hu S, Hu X, Huizinga J, Jain S, Jain S, Jang J, Jiang A, Jiang R, Jin H, Jin D, Jomoto S, Jonn B, Jun H, Kaftan T, Kaiser L, Kamali A, Kanitscheider I, Keskar NS, Khan T, Kilpatrick L, Kim JW, Kim C, Kim Y, Kirchner JH, Kiros J, Knight M, Kokotajlo D, Kondraciuk L, Kondrich A, Konstantinidis A, Kosic K, Krueger G, Kuo V, Lampe M, Lan I, Lee T, Leike J, Leung J, Levy D, Li CM, Lim R, Lin M, Lin S, Litwin M, Lopez T, Lowe R, Lue P, Makanju A, Malfacini K, Manning S, Markov T, Markovski Y, Martin B, Mayer K, Mayne A, McGrew B, McKinney SM, McLeavey C, McMillan P, McNeil J, Medina D, Mehta A, Menick J, Metz L, Mishchenko A, Mishkin P, Monaco V, Morikawa E, Mossing D, Mu T, Murati M, Murk O, Mély D, Nair A, Nakano R, Nayak R, Neelakantan A, Ngo R, Noh H, Ouyang L, O’Keefe C, Pachocki J, Paino A, Palermo J, Pantuliano A, Parascandolo G, Parish J, Parparita E, Passos A, Pavlov M, Peng A, Perelman A, Peres F de AB, Petrov M, Pinto HP de O, Michael, Pokorny, Pokrass M, Pong VH, Powell T, Power A, Power B, Proehl E, Puri R, Radford A, Rae J, Ramesh A, Raymond C, Real F, Rimbach K, Ross C, Rotsted B, Roussez H, Ryder N, Saltarelli M, Sanders T, Santurkar S, Sastry G, Schmidt H, Schnurr D, Schulman J, Selsam D, Sheppard K, Sherbakov T, Shieh J, Shoker S, Shyam P, Sidor S, Sigler E, Simens M, Sitkin J, Slama K, Sohl I, Sokolowsky B, Song Y, Staudacher N, Such FP, Summers N, Sutskever I, Tang J, Tezak N, Thompson MB, Tillet P, Tootoonchian A, Tseng E, Tuggle P, Turley N, Tworek J, Uribe JFC, Vallone A, Vijayvergiya A, Voss C, Wainwright C, Wang JJ, Wang A, Wang B, Ward J, Wei J, Weinmann CJ, Welihinda A, Welinder P, Weng J, Weng L, Wiethoff M, Willner D, Winter C, Wolrich S, Wong H, Workman L, Wu S, Wu J, Wu M, Xiao K, Xu T, Yoo S, Yu K, Yuan Q, Zaremba W, Zellers R, Zhang C, Zhang M, Zhao S, Zheng T, Zhuang J, Zhuk W, Zoph B. GPT-4 Technical Report. arXiv; 2024. doi: 10.48550/arXiv.2303.08774

81. Nogueira R, Yang W, Cho K, Lin J. Multi-Stage Document Ranking with BERT. arXiv; 2019. doi: 10.48550/arXiv.1910.14424
82. Kumar L, Sarkar S. ListBERT: Learning to Rank E-commerce products with Listwise BERT. arXiv; 2022. doi: 10.48550/arXiv.2206.15198
83. Wang Y, Magrabi F. Assessing the transferability of BERT to patient safety: classifying multiple types of incident reports. *BMJ Health Care Inform* 2025 Aug 18;32(1):e101146. PMID:40829884
84. Chen H, Cohen E, Wilson D, Alfted M. Improving Patient Safety Event Report Classification with Machine Learning and Contextual Text Representation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* SAGE Publications Inc; 2023 Sept 1;67(1):1063–1069. doi: 10.1177/21695067231193645
85. Himmelstein G, Bates D, Zhou L. Examination of Stigmatizing Language in the Electronic Health Record. *JAMA Network Open* 2022 Jan 27;5(1):e2144967. doi: 10.1001/jamanetworkopen.2021.44967
86. Martin K, Stanford C. An analysis of documentation language and word choice among forensic mental health nurses. *Int J Ment Health Nurs* 2020 Dec;29(6):1241–1252. PMID:32648351

87. Gordon CM, Yu B, Leung F, Legal M, Simons J, Wang E, Gnyra M. Characterizing Stigmatizing and Biased Language in Clinical Pharmacist Documentation. *JACCP: JOURNAL OF THE AMERICAN COLLEGE OF CLINICAL PHARMACY* 2025;8(10):974–982. doi: 10.1002/jac5.70119
88. Sun J, Shaib C, Wallace BC. Evaluating the Zero-shot Robustness of Instruction-tuned Language Models. *arXiv*; 2023. doi: 10.48550/arXiv.2306.11270
89. World Health Organization. Medication Without Harm. 2023. Available from: <https://www.who.int/initiatives/medication-without-harm> [accessed Mar 8, 2023]
90. Ngo J, Lau D, Ploquin J, Receveur T, Stassen K, Del Castilho C. Improving incident reporting among physicians at south health campus hospital. *BMJ Open Qual* 2022 Oct;11(4):e001945. doi: 10.1136/bmjopen-2022-001945
91. Gong Y, Song H-Y, Wu X, Hua L. Identifying barriers and benefits of patient safety event reporting toward user-centered design. *Saf Health BioMed Central*; 2015 Dec;1(1):1–9. doi: 10.1186/2056-5917-1-7
92. Gong Y. Data Consistency in a Voluntary Medical Incident Reporting System. *J Med Syst* 2011 Aug 1;35(4):609–615. doi: 10.1007/s10916-009-9398-y
93. Koike D, Ito M, Horiguchi A, Yatsuya H, Ota A. Implementation strategies for the patient safety reporting system using Consolidated Framework for Implementation Research: a retrospective mixed-method analysis. *BMC Health Services Research* 2022 Mar 28;22(1):409. doi: 10.1186/s12913-022-07822-9
94. Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. Deep Learning Based Text Classification: A Comprehensive Review. *arXiv*; 2021. Available from: <http://arxiv.org/abs/2004.03705> [accessed Mar 8, 2023]
95. Asan O, Bayrak AE, Choudhury A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *J Med Internet Res* 2020 June 19;22(6):e15154. PMID:32558657
96. Sharma M, Savage C, Nair M, Larsson I, Svedberg P, Nygren JM. Artificial Intelligence Applications in Health Care Practice: Scoping Review. *Journal of Medical Internet Research JMIR Publications Inc., Toronto, Canada*; 2022 Oct 5;24(10):e40238. doi: 10.2196/40238
97. Sezgin E. Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers. *Digit Health* 2023 July 2;9:20552076231186520. PMID:37426593
98. Hemmer P, Schemmer M, Riefle L, Rosellen N, Vössing M, Köhl N. Factors that influence the adoption of human-AI collaboration in clinical decision-making. *arXiv*; 2022. doi: 10.48550/arXiv.2204.09082
99. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, the Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC*

- Medical Informatics and Decision Making 2020 Nov 30;20(1):310. doi: 10.1186/s12911-020-01332-6
100. Wang L, Zhang Y, Chignell M, Shan B, Sheehan KA, Razak F, Verma A. Boosting Delirium Identification Accuracy With Sentiment-Based Natural Language Processing: Mixed Methods Study. *JMIR Med Inform* 2022 Dec 20;10(12):e38161. PMID:36538363
 101. Ribeiro MT, Singh S, Guestrin C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. *arXiv*; 2016. doi: 10.48550/arXiv.1602.04938
 102. Palojoki S, Saranto K, Reponen E, Skants N, Vakkuri A, Vuokko R. Classification of Electronic Health Record–Related Patient Safety Incidents: Development and Validation Study. *JMIR Medical Informatics* JMIR Publications Inc., Toronto, Canada; 2021 Aug 31;9(8):e30470. doi: 10.2196/30470
 103. Figma: The Collaborative Interface Design Tool. Figma. Available from: <https://www.figma.com/> [accessed Sept 24, 2023]
 104. Delio J, Catalanotti JS, Marko K, Paul C, Taffel M, Ho G, Berger J. Integrating Adverse Event Reporting Into a Free-Text Mobile Application Used in Daily Workflow Increases Adverse Event Reporting by Physicians. *Am J Med Qual* 2020;35(5):374–379. PMID:31795736
 105. Klemp K, Zwart D, Hansen J, Hellebek T, Luettel D, Verstappen W, Beyer M, Gerlach FM, Hoffmann B, Esmail A. A safety incident reporting system for primary care. A systematic literature review and consensus procedure by the LINNEAUS collaboration on patient safety in primary care. *Eur J Gen Pract* 2015 Sept;21 Suppl(sup1):39–44. PMID:26339835
 106. Hasegawa T, Fujita S. Patient Safety Policies – Experiences, Effects and Priorities; Lessons from OECD Member States –.
 107. Wang Y, Coiera E, Runciman W, Magrabi F. Automating the Identification of Patient Safety Incident Reports Using Multi-Label Classification. *Stud Health Technol Inform* 2017;245:609–613. PMID:29295168
 108. Berretta S, Tausch A, Ontrup G, Gilles B, Peifer C, Kluge A. Defining human-AI teaming the human-centered way: a scoping review and network analysis. *Frontiers in Artificial Intelligence* 2023;6. Available from: <https://www.frontiersin.org/articles/10.3389/frai.2023.1250725> [accessed Oct 4, 2023]
 109. Lai Y, Kankanhalli A, Ong D. Human-AI Collaboration in Healthcare: A Review and Research Agenda. 2021. p. 390. Available from: <http://hdl.handle.net/10125/70657> [accessed Apr 7, 2025]
 110. Reverberi C, Rigon T, Solari A, Hassan C, Cherubini P, Cherubini A. Experimental evidence of effective human–AI collaboration in medical decision-making. *Sci Rep Nature Publishing Group*; 2022 Sept 2;12(1):14952. doi: 10.1038/s41598-022-18751-2

111. Lee MH, Siewiorek DP, Smailagic A, Bernardino A, Bermúdez i Badia S. A Human-AI Collaborative Approach for Clinical Decision Making on Rehabilitation Assessment. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems New York, NY, USA: Association for Computing Machinery; 2021. p. 1–14. doi: 10.1145/3411764.3445472
112. Canada H. Overview of Vanessa’s Law. 2014. Available from: <https://www.canada.ca/en/health-canada/services/drugs-health-products/legislation-guidelines/overview-vanessa-law-protecting-canadians-unsafe-drugs-act-vanessa-law-amendments-food-drugs-act.html> [accessed Apr 11, 2025]
113. Sen. Jeffords JM [I-V. S.544 - 109th Congress (2005-2006): Patient Safety and Quality Improvement Act of 2005. 2005. Available from: <https://www.congress.gov/bill/109th-congress/senate-bill/544> [accessed Apr 11, 2025]
114. Healthcare Excellence Canada. Status of Patient Safety Incident Legislation and Best Practices Across Canada. 2021 Apr.
115. Howe JL, Hettinger AZ, Ratwani RM. Using patient safety-event report data to assess health-IT safety: benefits and challenges. *Lancet Digit Health* 2019 July;1(3):e104–e105. PMID:33323256
116. Mitchell C, Butler L, Holloway AD, Ra JH, Adapa K, Greenberg C, Marks LB, Ivester T, Mazur L. Analysis of patient safety event report categories at one large academic hospital. *Front Health Serv Frontiers*; 2024 Apr 2;4. doi: 10.3389/frhs.2024.1337840
117. Vincent C, Davis R. Patients and families as safety experts. *CMAJ* 2012 Jan 10;184(1):15–16. PMID:22105747
118. Excellent Care for All Act, 2010, S.O. 2010, c. 14 | ontario.ca. Available from: <https://www.ontario.ca/laws/statute/10e14> [accessed Jan 12, 2026]
119. Williams SD, Ashcroft DM. Medication errors: how reliable are the severity ratings reported to the national reporting and learning system? *Int J Qual Health Care* 2009 Oct;21(5):316–320. PMID:19679598
120. England NHS. NHS England » NRLS national patient safety incident reports: commentary (October 2022). 2022. Available from: <https://www.england.nhs.uk/publication/nrls-national-patient-safety-incident-reports-commentary-october-2022/> [accessed Apr 14, 2025]
121. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016. p. 785–794. doi: 10.1145/2939672.2939785
122. MacLaughlin A, Smith D. Content-based Models of Quotation. In: Merlo P, Tiedemann J, Tsarfaty R, editors. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume Online: Association for Computational Linguistics; 2021. p. 2296–2314. doi: 10.18653/v1/2021.eacl-main.195

123. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems* Curran Associates, Inc.; 2017. Available from: https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html [accessed July 8, 2025]
124. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020 Feb 15;36(4):1234–1240. doi: 10.1093/bioinformatics/btz682
125. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans Comput Healthcare* 2021 Oct 15;3(1):2:1-2:23. doi: 10.1145/3458754
126. Akkalyoncu Yilmaz Z, Yang W, Zhang H, Lin J. Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval. In: Inui K, Jiang J, Ng V, Wan X, editors. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* Hong Kong, China: Association for Computational Linguistics; 2019. p. 3490–3496. doi: 10.18653/v1/D19-1352
127. Ahmad M, Batyrshin I, Sidorov G. Sentiment Analysis Using a Large Language Model–Based Approach to Detect Opioids Mixed With Other Substances Via Social Media: Method Development and Validation. *JMIR Infodemiology* JMIR Publications Inc., Toronto, Canada; 2025 June 19;5(1):e70525. doi: 10.2196/70525
128. Chen W, Yang K, Yu Z, Shi Y, Chen CLP. A survey on imbalanced learning: latest research, applications and future directions. *Artif Intell Rev* 2024 May 9;57(6):137. doi: 10.1007/s10462-024-10759-6
129. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Int Res* 2002 June 1;16(1):321–357.
130. Verberne S, van Halteren H, Theijssen D, Raaijmakers S, Boves L. Learning to rank for why-question answering. *Inf Retrieval* 2011 Apr 1;14(2):107–132. doi: 10.1007/s10791-010-9136-6
131. Choudhury A, Asan O. Role of Artificial Intelligence in Patient Safety Outcomes: Systematic Literature Review. *JMIR Medical Informatics* JMIR Publications Inc., Toronto, Canada; 2020 July 24;8(7):e18599. doi: 10.2196/18599
132. Statistics. Statistics » Patient safety data. Available from: <https://www.england.nhs.uk/statistics/statistical-work-areas/patient-safety-data/> [accessed Oct 28, 2025]
133. Chen H, Alfred M, Brown AD, Atinga A, Cohen E. Intersection of Performance, Interpretability, and Fairness in Neural Prototype Tree for Chest X-Ray Pathology

- Detection: Algorithm Development and Validation Study. JMIR Formative Research JMIR Publications Inc., Toronto, Canada; 2024 Dec 5;8(1):e59045. doi: 10.2196/59045
134. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med* 2020 Feb 6;3:17. PMID:32047862
 135. Liang J, Li Y, Zhang Z, Shen D, Xu J, Zheng X, Wang T, Tang B, Lei J, Zhang J. Adoption of Electronic Health Records (EHRs) in China During the Past 10 Years: Consecutive Survey Data Analysis and Comparison of Sino-American Challenges and Experiences. *Journal of Medical Internet Research JMIR Publications Inc., Toronto, Canada*; 2021 Feb 18;23(2):e24813. doi: 10.2196/24813
 136. Woldemariam MT, Jimma W. Adoption of electronic health record systems to enhance the quality of healthcare in low-income countries: a systematic review. *BMJ Health Care Inform* 2023 June 12;30(1):e100704. PMID:37308185
 137. Overhage JM, McCallie D. Physician Time Spent Using the Electronic Health Record During Outpatient Encounters: A Descriptive Study. *Ann Intern Med* 2020 Feb 4;172(3):169–174. PMID:31931523
 138. Gedikci Ondogan A, Sargin M, Canoz K. Use of electronic medical records in the digital healthcare system and its role in communication and medical information sharing among healthcare professionals. *Informatics in Medicine Unlocked* 2023 Jan 1;42:101373. doi: 10.1016/j.imu.2023.101373
 139. Healy M, Richard A, Kidia K. How to Reduce Stigma and Bias in Clinical Communication: a Narrative Review. *J Gen Intern Med* 2022 Aug;37(10):2533–2540. PMID:35524034
 140. Beukeboom CJ, Burgers C. Linguistic Bias. *Oxford Research Encyclopedia of Communication*. Available from: <https://doi.org/10.1093/acrefore/9780190228613.013.439> [accessed Apr 21, 2025] ISBN:978-0-19-022861-3
 141. Pennebaker JW, Mehl MR, Niederhoffer KG. Psychological aspects of natural language use: our words, our selves. *Annu Rev Psychol* 2003;54:547–577. PMID:12185209
 142. Gopal DP, Chetty U, O'Donnell P, Gajria C, Blackadder-Weinstein J. Implicit bias in healthcare: clinical practice, research and decision making. *Future Healthc J* 2021 Mar;8(1):40–48. PMID:33791459
 143. Puhl R, Peterson JL, Luedicke J. Motivating or stigmatizing? Public perceptions of weight-related language used by health providers. *Int J Obes (Lond)* 2013 Apr;37(4):612–619. PMID:22777543
 144. Glassberg J, Tanabe P, Richardson L, Debaun M. Among emergency physicians, use of the term “Sickler” is associated with negative attitudes toward people with sickle cell disease. *Am J Hematol* 2013 June;88(6):532–533. PMID:23526459

145. Wang MD, Khanna R, Najafi N. Characterizing the Source of Text in Electronic Health Record Progress Notes. *JAMA Internal Medicine* 2017 Aug 1;177(8):1212–1213. doi: 10.1001/jamainternmed.2017.1548
146. Scroggins JK, Hulchafo II, Harkins S, Scharp D, Moen H, Davoudi A, Cato K, Tadiello M, Topaz M, Barcelona V. Identifying stigmatizing and positive/preferred language in obstetric clinical notes using natural language processing. *Journal of the American Medical Informatics Association* 2025 Feb 1;32(2):308–317. doi: 10.1093/jamia/ocae290
147. Blease C, Walker J, DesRoches CM, Delbanco T. New U.S. Law Mandates Access to Clinical Notes: Implications for Patients and Clinicians. *Ann Intern Med* 2021 Jan;174(1):101–102. PMID:33045176
148. Fernández L, Fossa A, Dong Z, Delbanco T, Elmore J, Fitzgerald P, Harcourt K, Perez J, Walker J, DesRoches C. Words Matter: What Do Patients Find Judgmental or Offensive in Outpatient Notes? *J Gen Intern Med* 2021 Sept;36(9):2571–2578. PMID:33528782
149. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *arXiv*; 2021. doi: 10.48550/arXiv.2107.13586
150. Sun X, Li X, Li J, Wu F, Guo S, Zhang T, Wang G. Text Classification via Large Language Models. In: Bouamor H, Pino J, Bali K, editors. *Findings of the Association for Computational Linguistics: EMNLP 2023* Singapore: Association for Computational Linguistics; 2023. p. 8990–9005. doi: 10.18653/v1/2023.findings-emnlp.603
151. Edwards A, Camacho-Collados J. Language Models for Text Classification: Is In-Context Learning Enough? In: Calzolari N, Kan M-Y, Hoste V, Lenci A, Sakti S, Xue N, editors. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* Torino, Italia: ELRA and ICCL; 2024. p. 10058–10072. Available from: <https://aclanthology.org/2024.lrec-main.879> [accessed Aug 2, 2024]
152. White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, Elnashar A, Spencer-Smith J, Schmidt DC. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *arXiv*; 2023. doi: 10.48550/arXiv.2302.11382
153. Zaghir J, Naguib M, Bjelogrljic M, Névél A, Tannier X, Lovis C. Prompt Engineering Paradigms for Medical Applications: Scoping Review. *Journal of Medical Internet Research JMIR Publications Inc.*, Toronto, Canada; 2024 Sept 10;26(1):e60501. doi: 10.2196/60501
154. Park J, Saha S, Chee B, Taylor J, Beach MC. Physician Use of Stigmatizing Language in Patient Medical Records. *JAMA Network Open* 2021 July 14;4(7):e2117052. doi: 10.1001/jamanetworkopen.2021.17052
155. Feng Q, Du M, Zou N, Hu X. Fair Machine Learning in Healthcare: A Review. *arXiv*; 2024. doi: 10.48550/arXiv.2206.14397

156. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: Fairness gaps in deep chest X-ray classifiers. *arXiv*; 2020. doi: 10.48550/arXiv.2003.00827
157. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E, Le Q, Zhou D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv*; 2023. doi: 10.48550/arXiv.2201.11903
158. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, Pollard TJ, Hao S, Moody B, Gow B, Lehman LH, Celi LA, Mark RG. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data* 2023 Jan 3;10:1. PMID:36596836
159. Song Y, Wang T, Mondal SK, Sahoo JP. A Comprehensive Survey of Few-shot Learning: Evolution, Applications, Challenges, and Opportunities. *arXiv*; 2022. doi: 10.48550/arXiv.2205.06743
160. Gretz S, Halfon A, Shnayderman I, Toledo-Ronen O, Spector A, Dankin L, Katsis Y, Arviv O, Katz Y, Slonim N, Ein-Dor L. Zero-shot Topical Text Classification with LLMs - an Experimental Study. In: Bouamor H, Pino J, Bali K, editors. *Findings of the Association for Computational Linguistics: EMNLP 2023 Singapore: Association for Computational Linguistics*; 2023. p. 9647–9676. doi: 10.18653/v1/2023.findings-emnlp.647
161. Khandaker F, Senderovich A, Zhao J, Cohen E, Yu E, Carbajales S, Chan A. Transformer models for mining intents and predicting activities from emails in knowledge-intensive processes. *Engineering Applications of Artificial Intelligence Elsevier*; 2024;128:107450.
162. Williams A, Nangia N, Bowman S. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In: Walker M, Ji H, Stent A, editors. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* New Orleans, Louisiana: Association for Computational Linguistics; 2018. p. 1112–1122. doi: 10.18653/v1/N18-1101
163. Gera A, Halfon A, Shnarch E, Perlitz Y, Ein-Dor L, Slonim N. Zero-Shot Text Classification with Self-Training. In: Goldberg Y, Kozareva Z, Zhang Y, editors. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing Abu Dhabi, United Arab Emirates: Association for Computational Linguistics*; 2022. p. 1107–1119. doi: 10.18653/v1/2022.emnlp-main.73
164. Alcoforado A, Ferraz TP, Gerber R, Bustos E, Oliveira AS, Veloso BM, Siqueira FL, Costa AHR. ZeroBERTo: Leveraging Zero-Shot Text Classification by Topic Modeling. In: Pinheiro V, Gamallo P, Amaro R, Scarton C, Batista F, Silva D, Magro C, Pinto H, editors. *Computational Processing of the Portuguese Language Cham: Springer International Publishing*; 2022. p. 125–136. doi: 10.1007/978-3-030-98305-5_12
165. Loukas L, Stogiannidis I, Malakasiotis P, Vassos S. Breaking the Bank with ChatGPT: Few-Shot Text Classification for Finance. *arXiv*; 2023. doi: 10.48550/arXiv.2308.14634

166. Edwards C, Naik A, Khot T, Burke M, Ji H, Hope T. SynerGPT: In-Context Learning for Personalized Drug Synergy Prediction and Drug Design. arXiv; 2023. doi: 10.48550/arXiv.2307.11694
167. Wang L, Yang N, Huang X, Jiao B, Yang L, Jiang D, Majumder R, Wei F. Text Embeddings by Weakly-Supervised Contrastive Pre-training. arXiv; 2022; doi: 10.48550/ARXIV.2212.03533
168. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv; 2019. doi: 10.48550/arXiv.1908.10084
169. Zhang H, Shafiq MO. Survey of transformers and towards ensemble learning using transformers for natural language processing. *Journal of Big Data* 2024 Feb 4;11(1):25. doi: 10.1186/s40537-023-00842-0
170. Tran HTH, Chatterjee N, Pollak S, Doucet A. DeBERTa Beats Behemoths: A Comparative Analysis of Fine-Tuning, Prompting, and PEFT Approaches on LegalLensNER. In: Aletras N, Chalkidis I, Barrett L, Goanță C, Preoțiuc-Pietro D, Spanakis G, editors. *Proceedings of the Natural Legal Language Processing Workshop 2024 Miami, FL, USA: Association for Computational Linguistics*; 2024. p. 371–380. doi: 10.18653/v1/2024.nllp-1.34
171. Yu H, Yang Z, Pelrine K, Godbout JF, Rabbany R. Open, Closed, or Small Language Models for Text Classification? arXiv; 2023. doi: 10.48550/arXiv.2308.10092
172. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. arXiv; 2019. Available from: <http://arxiv.org/abs/1711.05101> [accessed July 1, 2023]
173. Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, Li Y, Wang X, Dehghani M, Brahma S, Webson A, Gu SS, Dai Z, Suzgun M, Chen X, Chowdhery A, Castro-Ros A, Pellat M, Robinson K, Valter D, Narang S, Mishra G, Yu A, Zhao V, Huang Y, Dai A, Yu H, Petrov S, Chi EH, Dean J, Devlin J, Roberts A, Zhou D, Le QV, Wei J. Scaling Instruction-Finetuned Language Models. arXiv; 2022. doi: 10.48550/arXiv.2210.11416
174. Wei J, Bosma M, Zhao VY, Guu K, Yu AW, Lester B, Du N, Dai AM, Le QV. Finetuned Language Models Are Zero-Shot Learners. arXiv; 2022. doi: 10.48550/arXiv.2109.01652
175. ContactDoctor/Bio-Medical-Llama-3-8B · Hugging Face. Available from: <https://huggingface.co/ContactDoctor/Bio-Medical-Llama-3-8B> [accessed Feb 12, 2025]
176. Sclar M, Choi Y, Tsvetkov Y, Suhr A. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. arXiv; 2024. doi: 10.48550/arXiv.2310.11324
177. Lossio-Ventura JA, Weger R, Lee AY, Guinee EP, Chung J, Atlas L, Linos E, Pereira F. A Comparison of ChatGPT and Fine-Tuned Open Pre-Trained Transformers (OPT) Against Widely Used Sentiment Analysis Tools: Sentiment Analysis of COVID-19 Survey Data. *JMIR Mental Health JMIR Publications Inc., Toronto, Canada*; 2024 Jan 25;11(1):e50150. doi: 10.2196/50150

178. Gao T, Fisch A, Chen D. Making Pre-trained Language Models Better Few-shot Learners. arXiv; 2021. doi: 10.48550/arXiv.2012.15723
179. Meskó B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. *J Med Internet Res* 2023 Oct 4;25:e50638. PMID:37792434
180. Wang J, Shi E, Yu S, Wu Z, Ma C, Dai H, Yang Q, Kang Y, Wu J, Hu H, Yue C, Zhang H, Liu Y, Pan Y, Liu Z, Sun L, Li X, Ge B, Jiang X, Zhu D, Yuan Y, Shen D, Liu T, Zhang S. Prompt Engineering for Healthcare: Methodologies and Applications. arXiv; 2024. doi: 10.48550/arXiv.2304.14670
181. Valdez A. Words matter: Labelling, bias and stigma in nursing. *J Adv Nurs* 2021 Nov;77(11):e33–e35. PMID:34245183
182. Alowidi N, Ali R, Sadaqah M, Naemi FMA. Advancing Kidney Transplantation: A Machine Learning Approach to Enhance Donor–Recipient Matching. *Diagnostics (Basel)* 2024 Sept 25;14(19):2119. PMID:39410523
183. Srivastava M, Heidari H, Krause A. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. arXiv.org. 2019. Available from: <https://arxiv.org/abs/1902.04783v4> [accessed July 2, 2023]
184. Government of Canada IAP on RE. Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans – TCPS 2 (2018). 2019. Available from: https://ethics.gc.ca/eng/policy-politique_tcps2-eptc2_2018.html [accessed Aug 6, 2024]
185. Carroll SM. Respecting and Empowering Vulnerable Populations: Contemporary Terminology. *The Journal for Nurse Practitioners* 2019 Mar 1;15(3):228–231. doi: 10.1016/j.nurpra.2018.12.031
186. Zhang S, Dong L, Li X, Zhang S, Sun X, Wang S, Li J, Hu R, Zhang T, Wu F, Wang G. Instruction Tuning for Large Language Models: A Survey. arXiv; 2024. doi: 10.48550/arXiv.2308.10792
187. Stuart J. Russell, Peter Norvig. *Artificial Intelligence: A Modern Approach*,. 4th ed. Available from: <https://aima.cs.berkeley.edu/>
188. Yu T, Simoff S, Jan T. VQSVM: A case study for incorporating prior domain knowledge into inductive machine learning. *Neurocomputing* 2010 Aug 1;73(13):2614–2623. doi: 10.1016/j.neucom.2010.05.007
189. Wagstaff KL. Constrained Clustering. *Encyclopedia of Machine Learning* Springer, Boston, MA; 2011. p. 220–221. doi: 10.1007/978-0-387-30164-8_163 ISBN:978-0-387-30164-8
190. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. arXiv; 2022. doi: 10.48550/arXiv.1908.09635

191. Yang Y, Lin M, Zhao H, Peng Y, Huang F, Lu Z. A survey of recent methods for addressing AI fairness and bias in biomedicine. *ArXiv* 2024 Feb 13;arXiv:2402.08250v1. PMID:38529077
192. Zhang H, Dullerud N, Roth K, Oakden-Rayner L, Pfohl SR, Ghassemi M. Improving the Fairness of Chest X-ray Classifiers. *arXiv*; 2022. doi: 10.48550/arXiv.2203.12609
193. Davoudi A, Chae S, Evans L, Sridharan S, Song J, Bowles KH, McDonald MV, Topaz M. Fairness gaps in Machine learning models for hospitalization and emergency department visit risk prediction in home healthcare patients with heart failure. *International Journal of Medical Informatics* 2024 Nov 1;191:105534. doi: 10.1016/j.ijmedinf.2024.105534
194. Tierney AA, Gayre G, Hoberman B, Mattern B, Ballesca M, Kipnis P, Liu V, Lee K. Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation. *NEJM Catalyst Massachusetts Medical Society*; 2024 Feb 21;5(3):CAT.23.0404. doi: 10.1056/CAT.23.0404
195. Hassan H, Zipursky AR, Rabbani N, You JG-T, Tse G, Orenstein E, Parsons CR, Jessa K, Lawton G, Shin HS, Sung L, Ray M, Yan AP. Special Topic on Burnout: Clinical Implementation of Artificial Intelligence Scribes in Healthcare: A Systematic Review. *Appl Clin Inform* 2025 Apr 30; PMID:40306686
196. Basha I. The Human Factors in the Adoption of Ambient Artificial Intelligence Scribe Technology: Towards Informed and User-centered Implementation of AI in Healthcare. University of Waterloo; 2024 Oct 8; Available from: <https://hdl.handle.net/10012/21130> [accessed June 30, 2025]
197. Sarkar AR, Chuang Y-S, Mohammed N, Jiang X. De-identification is not enough: a comparison between de-identified and synthetic clinical notes. *Sci Rep Nature Publishing Group*; 2024 Nov 29;14(1):29669. doi: 10.1038/s41598-024-81170-y
198. de Aguiar EJ, Traina C, Traina AJM. Security and Privacy in Machine Learning for Health Systems: Strategies and Challenges. *Yearb Med Inform* 2023 Dec 26;32(1):269–281. PMID:38147869
199. Rights (OCR) O for C. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. 2012. Available from: <https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html> [accessed Jan 31, 2025]
200. Smith V, Shamsabadi AS, Ashurst C, Weller A. Identifying and Mitigating Privacy Risks Stemming from Language Models: A Survey. *arXiv*; 2024. doi: 10.48550/arXiv.2310.01424
201. Chen Q, Hu Y, Peng X, Xie Q, Jin Q, Gilson A, Singer MB, Ai X, Lai P-T, Wang Z, Keloth VK, Raja K, Huang J, He H, Lin F, Du J, Zhang R, Zheng WJ, Adelman RA, Lu Z, Xu H. A systematic evaluation of large language models for biomedical natural language processing: benchmarks, baselines, and recommendations. *arXiv*; 2024. doi: 10.48550/arXiv.2305.16326

202. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *npj Digit Med Nature Publishing Group*; 2024 Jan 24;7(1):1–7. doi: 10.1038/s41746-024-01010-1
203. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online. *J Med Internet Res* 2013 Nov 1;15(11):e239. PMID:24184993
204. Bello A, Ng S-C, Leung M-F. A BERT Framework to Sentiment Analysis of Tweets. *Sensors (Basel)* 2023 Jan 2;23(1):506. PMID:36617101
205. Melton CA, White BM, Davis RL, Bednarczyk RA, Shaban-Nejad A. Fine-tuned Sentiment Analysis of COVID-19 Vaccine–Related Social Media Data: Comparative Study. *Journal of Medical Internet Research JMIR Publications Inc., Toronto, Canada*; 2022 Oct 17;24(10):e40408. doi: 10.2196/40408
206. Ktena I, Wiles O, Albuquerque I, Rebuffi S-A, Tanno R, Roy AG, Azizi S, Belgrave D, Kohli P, Cemgil T, Karthikesalingam A, Goyal S. Generative models improve fairness of medical classifiers under distribution shifts. *Nat Med Nature Publishing Group*; 2024 Apr;30(4):1166–1173. doi: 10.1038/s41591-024-02838-6
207. Abramoff MD, Tarver ME, Loyo-Berrios N, Trujillo S, Char D, Obermeyer Z, Eydelman MB, Maisel WH. Considerations for addressing bias in artificial intelligence for health equity. *NPJ Digit Med* 2023 Sept 12;6:170. PMID:37700029
208. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghighi B, Ball R, Shpanskaya K, Seekins J, Mong DA, Halabi SS, Sandberg JK, Jones R, Larson DB, Langlotz CP, Patel BN, Lungren MP, Ng AY. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *arXiv*; 2019. doi: 10.48550/arXiv.1901.07031
209. Serna I, Morales A, Fierrez J, Obradovich N. Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *Artificial Intelligence* 2022 Apr 1;305:103682. doi: 10.1016/j.artint.2022.103682
210. Cho N-J, Jeong I, Ahn S-J, Gil H-W, Kim Y, Park J-H, Kang S, Lee H. Machine Learning to Assist in Managing Acute Kidney Injury in General Wards: Multicenter Retrospective Study. *J Med Internet Res* 2025 Mar 18;27:e66568. PMID:40101226
211. Huang P, Yang J, Zhao D, Ran T, Luo Y, Yang D, Zheng X, Zhou S, Chen C. Machine Learning-Based Prediction of Early Complications Following Surgery for Intestinal Obstruction: Multicenter Retrospective Study. *J Med Internet Res* 2025 Mar 3;27:e68354. PMID:40053794
212. Haldar S, Iv OBDW, Akter S. From Context to Care: Rethinking Stigma Detection in Clinical Language Models. *JMIR Medical Informatics JMIR Publications Inc., Toronto, Canada*; 2025 Oct 2;13(1):e82484. doi: 10.2196/82484

213. Yin S, Fu C, Zhao S, Li K, Sun X, Xu T, Chen E. A Survey on Multimodal Large Language Models. arXiv.org. 2023. doi: 10.1093/nsr/nwae403
214. Team G, Anil R, Borgeaud S, Alayrac J-B, Yu J, Soricut R, Schalkwyk J, Dai AM, Hauth A, Millican K, Silver D, Johnson M, Antonoglou I, Schrittwieser J, Glaese A, Chen J, Pitler E, Lillicrap T, Lazaridou A, Firat O, Molloy J, Isard M, Barham PR, Hennigan T, Lee B, Viola F, Reynolds M, Xu Y, Doherty R, Collins E, Meyer C, Rutherford E, Moreira E, Ayoub K, Goel M, Krawczyk J, Du C, Chi E, Cheng H-T, Ni E, Shah P, Kane P, Chan B, Faruqui M, Severyn A, Lin H, Li Y, Cheng Y, Ittycheriah A, Mahdiah M, Chen M, Sun P, Tran D, Bagri S, Lakshminarayanan B, Liu J, Orban A, Gura F, Zhou H, Song X, Boffy A, Ganapathy H, Zheng S, Choe H, Weisz Á, Zhu T, Lu Y, Gopal S, Kahn J, Kula M, Pitman J, Shah R, Taropa E, Merey MA, Bauml M, Chen Z, Shafey LE, Zhang Y, Sercinoglu O, Tucker G, Piqueras E, Krikun M, Barr I, Savinov N, Danihelka I, Roelofs B, White A, Andreassen A, Glehn T von, Yagati L, Kazemi M, Gonzalez L, Khalman M, Sygnowski J, Frechette A, Smith C, Culp L, Proleev L, Luan Y, Chen X, Lottes J, Schucher N, Lebron F, Rustemi A, Clay N, Crone P, Kocisky T, Zhao J, Perz B, Yu D, Howard H, Bloniarz A, Rae JW, Lu H, Sifre L, Maggioni M, Alcober F, Garrette D, Barnes M, Thakoor S, Austin J, Barth-Maron G, Wong W, Joshi R, Chaabouni R, Fatiha D, Ahuja A, Tomar GS, Senter E, Chadwick M, Kornakov I, Attaluri N, Iturrate I, Liu R, Li Y, Cogan S, Chen J, Jia C, Gu C, Zhang Q, Grimstad J, Hartman AJ, Garcia X, Pillai TS, Devlin J, Laskin M, Casas D de L, Valter D, Tao C, Blanco L, Badia AP, Reitter D, Chen M, Brennan J, Rivera C, Brin S, Iqbal S, Surita G, Labanowski J, Rao A, Winkler S, Parisotto E, Gu Y, Olszewska K, Addanki R, Miech A, Louis A, Teplyashin D, Brown G, Catt E, Balaguer J, Xiang J, Wang P, Ashwood Z, Briukhov A, Webson A, Ganapathy S, Sanghavi S, Kannan A, Chang M-W, Stjerngren A, Djolonga J, Sun Y, Bapna A, Aitchison M, Pejman P, Michalewski H, Yu T, Wang C, Love J, Ahn J, Bloxwich D, Han K, Humphreys P, Sellam T, Bradbury J, Godbole V, Samangooei S, Damoc B, Kaskasoli A, Arnold SMR, Vasudevan V, Agrawal S, Riesa J, Lepikhin D, Tanburn R, Srinivasan S, Lim H, Hodgkinson S, Shyam P, Ferret J, Hand S, Garg A, Paine TL, Li J, Li Y, Giang M, Neitz A, Abbas Z, York S, Reid M, Cole E, Chowdhery A, Das D, Rogozińska D, Nikolaev V, Sprechmann P, Nado Z, Zilka L, Prost F, He L, Monteiro M, Mishra G, Welty C, Newlan J, Jia D, Allamanis M, Hu CH, Liedekerke R de, Gilmer J, Saroufim C, Rijhwani S, Hou S, Shrivastava D, Baddepudi A, Goldin A, Oztirel A, Cassirer A, Xu Y, Sohn D, Sachan D, Amplayo RK, Swanson C, Petrova D, Narayan S, Guez A, Brahma S, Landon J, Patel M, Zhao R, Villela K, Wang L, Jia W, Rahtz M, Giménez M, Yeung L, Keeling J, Georgiev P, Mincu D, Wu B, Haykal S, Saputro R, Vodrahalli K, Qin J, Cankara Z, Sharma A, Fernando N, Hawkins W, Neyshabur B, Kim S, Hutter A, Agrawal P, Castro-Ros A, Driessche G van den, Wang T, Yang F, Chang S, Komarek P, McIlroy R, Lučić M, Zhang G, Farhan W, Sharman M, Natsev P, Michel P, Bansal Y, Qiao S, Cao K, Shakeri S, Butterfield C, Chung J, Rubenstein PK, Agrawal S, Mensch A, Soparkar K, Lenc K, Chung T, Pope A, Maggiore L, Kay J, Jhakra P, Wang S, Maynez J, Phuong M, Tobin T, Tacchetti A, Trebacz M, Robinson K, Katariya Y, Riedel S, Bailey P, Xiao K, Ghelani N, Aroyo L, Slone A, Hounsby N, Xiong X, Yang Z, Gribovskaya E, Adler J, Wirth M, Lee L, Li M, Kagohara T, Pavagadhi J, Bridgers S, Bortsova A, Ghemawat S, Ahmed Z, Liu T, Powell R, Bolina V, Iinuma M, Zablotskaia P, Besley J, Chung D-W, Dozat T, Comanescu R, Si X, Greer J, Su G, Polacek M, Kaufman RL, Tokumine S, Hu H, Buchatskaya E, Miao Y, Elhawaty M, Siddhant A, Tomasev N, Xing J, Greer C, Miller H, Ashraf S, Roy A, Zhang Z, Ma A, Filos A, Besta M, Blevins R, Klimenko T, Yeh C-K, Changpinyo S, Mu J, Chang O, Pajarskas

M, Muir C, Cohen V, Lan CL, Haridasan K, Marathe A, Hansen S, Douglas S, Samuel R, Wang M, Austin S, Lan C, Jiang J, Chiu J, Lorenzo JA, Sjöstrand LL, Cevey S, Gleicher Z, Avrahami T, Boral A, Srinivasan H, Selo V, May R, Aisopos K, Hussienot L, Soares LB, Baumli K, Chang MB, Recasens A, Caine B, Pritzel A, Pavetic F, Pardo F, Gergely A, Frye J, Ramasesh V, Horgan D, Badola K, Kassner N, Roy S, Dyer E, Campos VC, Tomala A, Tang Y, Badawy DE, White E, Mustafa B, Lang O, Jindal A, Vikram S, Gong Z, Caelles S, Hemsley R, Thornton G, Feng F, Stokowiec W, Zheng C, Thacker P, Ünlü Ç, Zhang Z, Saleh M, Svensson J, Bileschi M, Patil P, Anand A, Ring R, Tsihkas K, Vezer A, Selvi M, Shevlane T, Rodriguez M, Kwiatkowski T, Daruki S, Rong K, Dafoe A, FitzGerald N, Gulemberg K, Khan M, Hendricks LA, Pellat M, Feinberg V, Cobon-Kerr J, Sainath T, Rauh M, Hashemi SH, Ives R, Hasson Y, Noland E, Cao Y, Byrd N, Hou L, Wang Q, Sottiaux T, Paganini M, Lespiau J-B, Moufarek A, Hassan S, Shivakumar K, Amersfoort J van, Mandhane A, Joshi P, Goyal A, Tung M, Brock A, Sheahan H, Misra V, Li C, Rakićević N, Dehghani M, Liu F, Mittal S, Oh J, Noury S, Sezener E, Huot F, Lamm M, Cao ND, Chen C, Mudgal S, Stella R, Brooks K, Vasudevan G, Liu C, Chain M, Melinkeri N, Cohen A, Wang V, Seymore K, Zubkov S, Goel R, Yue S, Krishnakumaran S, Albert B, Hurley N, Sano M, Mohananey A, Joughin J, Filonov E, Keça T, Eldawy Y, Lim J, Rishi R, Badiezadegan S, Bos T, Chang J, Jain S, Padmanabhan SGS, Puttagunta S, Krishna K, Baker L, Kalb N, Bedapudi V, Kurzrok A, Lei S, Yu A, Litvin O, Zhou X, Wu Z, Sobell S, Siciliano A, Papir A, Neale R, Bragagnolo J, Toor T, Chen T, Anklin V, Wang F, Feng R, Gholami M, Ling K, Liu L, Walter J, Moghaddam H, Kishore A, Adamek J, Mercado T, Mallinson J, Wandekar S, Cagle S, Ofek E, Garrido G, Lombriser C, Mukha M, Sun B, Mohammad HR, Matak J, Qian Y, Peswani V, Janus P, Yuan Q, Schelin L, David O, Garg A, He Y, Duzhyi O, Älgmyr A, Lottaz T, Li Q, Yadav V, Xu L, Chinien A, Shivanna R, Chuklin A, Li J, Spadine C, Wolfe T, Mohamed K, Das S, Dai Z, He K, Dincklage D von, Upadhyay S, Maurya A, Chi L, Krause S, Salama K, Rabinovitch PG, M PKR, Selvan A, Dektiarev M, Ghiasi G, Guven E, Gupta H, Liu B, Sharma D, Shtacher IH, Paul S, Akerlund O, Aubet F-X, Huang T, Zhu C, Zhu E, Teixeira E, Fritze M, Bertolini F, Marinescu L-E, Bölle M, Paulus D, Gupta K, Latkar T, Chang M, Sanders J, Wilson R, Wu X, Tan Y-X, Thiet LN, Doshi T, Lall S, Mishra S, Chen W, Luong T, Benjamin S, Lee J, Andrejczuk E, Rabiej D, Ranjan V, Styrc K, Yin P, Simon J, Harriott MR, Bansal M, Robsky A, Bacon G, Greene D, Mirylenka D, Zhou C, Sarvana O, Goyal A, Andermatt S, Siegler P, Horn B, Israel A, Pongetti F, Chen C-W “Louis,” Selvatici M, Silva P, Wang K, Tolins J, Guu K, Yogev R, Cai X, Agostini A, Shah M, Nguyen H, Donnaile NÓ, Pereira S, Friso L, Stambler A, Kurzrok A, Kuang C, Romanikhin Y, Geller M, Yan ZJ, Jang K, Lee C-C, Fica W, Malmi E, Tan Q, Banica D, Balle D, Pham R, Huang Y, Avram D, Shi H, Singh J, Hidey C, Ahuja N, Saxena P, Dooley D, Potharaju SP, O’Neill E, Gokulchandran A, Foley R, Zhao K, Dusenberry M, Liu Y, Mehta P, Kotikalapudi R, Safranek-Shrader C, Goodman A, Kessinger J, Globen E, Kolhar P, Gorgolewski C, Ibrahim A, Song Y, Eichenbaum A, Brovelli T, Potluri S, Lahoti P, Baetu C, Ghorbani A, Chen C, Crawford A, Pal S, Sridhar M, Gurita P, Mujika A, Petrovski I, Cedoz P-L, Li C, Chen S, Santo ND, Goyal S, Punjabi J, Kappaganthu K, Kwak C, LV P, Velury S, Choudhury H, Hall J, Shah P, Figueira R, Thomas M, Lu M, Zhou T, Kumar C, Jurdi T, Chikkerur S, Ma Y, Yu A, Kwak S, Ähdel V, Rajayogam S, Choma T, Liu F, Barua A, Ji C, Park JH, Hellendoorn V, Bailey A, Bilal T, Zhou H, Khatir M, Sutton C, Rzadkowski W, Macintosh F, Vij R, Shagin K, Medina P, Liang C, Zhou J, Shah P, Bi Y, Dankovics A, Banga S, Lehmann S, Bredesen M, Lin Z, Hoffmann JE, Lai J, Chung R, Yang K, Balani N, Bražinskas A, Sozanschi A,

Hayes M, Alcalde HF, Makarov P, Chen W, Stella A, Snijders L, Mandl M, Kärrman A, Nowak P, Wu X, Dyck A, Vaidyanathan K, R R, Mallet J, Rudominer M, Johnston E, Mittal S, Udathu A, Christensen J, Verma V, Irving Z, Santucci A, Elsayed G, Davoodi E, Georgiev M, Tenney I, Hua N, Cideron G, Leurent E, Alnahlawi M, Georgescu I, Wei N, Zheng I, Scandinaro D, Jiang H, Snoek J, Sundararajan M, Wang X, Ontiveros Z, Karo I, Cole J, Rajashekhar V, Tumeh L, Ben-David E, Jain R, Uesato J, Datta R, Bunyan O, Wu S, Zhang J, Stanczyk P, Zhang Y, Steiner D, Naskar S, Azzam M, Johnson M, Paszke A, Chiu C-C, Elias JS, Mohiuddin A, Muhammad F, Miao J, Lee A, Vieillard N, Park J, Zhang J, Stanway J, Garmon D, Karmarkar A, Dong Z, Lee J, Kumar A, Zhou L, Evens J, Isaac W, Irving G, Loper E, Fink M, Arkatkar I, Chen N, Shafran I, Petrychenko I, Chen Z, Jia J, Levskaya A, Zhu Z, Grabowski P, Mao Y, Magni A, Yao K, Snaider J, Casagrande N, Palmer E, Suganthan P, Castaño A, Giannoumis I, Kim W, Rybiński M, Sreevatsa A, Prendki J, Soergel D, Goedeckemeyer A, Gierke W, Jafari M, Gaba M, Wiesner J, Wright DG, Wei Y, Vashisht H, Kulizhskaya Y, Hoover J, Le M, Li L, Iwuanyanwu C, Liu L, Ramirez K, Khorlin A, Cui A, LIN T, Wu M, Aguilar R, Pallo K, Chakladar A, Perng G, Abellan EA, Zhang M, Dasgupta I, Kushman N, Penchev I, Repina A, Wu X, Weide T van der, Ponnappalli P, Kaplan C, Simsa J, Li S, Dousse O, Yang F, Piper J, Ie N, Pasumarthi R, Lintz N, Vijayakumar A, Andor D, Valenzuela P, Lui M, Paduraru C, Peng D, Lee K, Zhang S, Greene S, Nguyen DD, Kurylowicz P, Hardin C, Dixon L, Janzer L, Choo K, Feng Z, Zhang B, Singhal A, Du D, McKinnon D, Antropova N, Bolukbasi T, Keller O, Reid D, Finchelstein D, Raad MA, Crocker R, Hawkins P, Dadashi R, Gaffney C, Franko K, Bulanova A, Leblond R, Chung S, Askham H, Cobo LC, Xu K, Fischer F, Xu J, Sorokin C, Alberti C, Lin C-C, Evans C, Dimitriev A, Forbes H, Banarse D, Tung Z, Omernick M, Bishop C, Sterneck R, Jain R, Xia J, Amid E, Piccinno F, Wang X, Banzal P, Mankowitz DJ, Polozov A, Krakovna V, Brown S, Bateni M, Duan D, Firoiu V, Thotakuri M, Natan T, Geist M, Girgin S tan, Li H, Ye J, Roval O, Tojo R, Kwong M, Lee-Thorp J, Yew C, Sinopalnikov D, Ramos S, Mellor J, Sharma A, Wu K, Miller D, Sonnerat N, Vnukov D, Greig R, Beattie J, Caveness E, Bai L, Eisenschlos J, Korchemniy A, Tsai T, Jasarevic M, Kong W, Dao P, Zheng Z, Liu F, Yang F, Zhu R, Teh TH, Sanmiya J, Gladchenko E, Trdin N, Toyama D, Rosen E, Tavakkol S, Xue L, Elkind C, Woodman O, Carpenter J, Papamakarios G, Kemp R, Kafle S, Grunina T, Sinha R, Talbert A, Wu D, Owusu-Afriyie D, Du C, Thornton C, Pont-Tuset J, Narayana P, Li J, Fatehi S, Wieting J, Ajmeri O, Uria B, Ko Y, Knight L, Héliou A, Niu N, Gu S, Pang C, Li Y, Levine N, Stolovich A, Santamaria-Fernandez R, Goenka S, Yustalim W, Strudel R, Elqursh A, Deck C, Lee H, Li Z, Levin K, Hoffmann R, Holtmann-Rice D, Bachem O, Arora S, Koh C, Yeganeh SH, Pöder S, Tariq M, Sun Y, Ionita L, Seyedhosseini M, Tafti P, Liu Z, Gulati A, Liu J, Ye X, Chrzaszcz B, Wang L, Sethi N, Li T, Brown B, Singh S, Fan W, Parisi A, Stanton J, Koverkathu V, Choquette-Choo CA, Li Y, Lu TJ, Ittycheriah A, Shroff P, Varadarajan M, Bahargam S, Willoughby R, Gaddy D, Desjardins G, Cornero M, Robenek B, Mittal B, Albrecht B, Shenoy A, Moiseev F, Jacobsson H, Ghaffarkhah A, Rivière M, Walton A, Crepy C, Parrish A, Zhou Z, Farabet C, Radebaugh C, Srinivasan P, Salm C van der, Fidjeland A, Scellato S, Latorre-Chimoto E, Klimczak-Plucińska H, Bridson D, Cesare D de, Hudson T, Mendolicchio P, Walker L, Morris A, Mauger M, Guseynov A, Reid A, Odooom S, Loher L, Cotruta V, Yenugula M, Grewe D, Petrushkina A, Duerig T, Sanchez A, Yadlowsky S, Shen A, Globerson A, Webb L, Dua S, Li D, Bhupatiraju S, Hurt D, Qureshi H, Agarwal A, Shani T, Eyal M, Khare A, Belle SR, Wang L, Tekur C, Kale MS, Wei J, Sang R, Saeta B, Liechty T, Sun Y, Zhao Y, Lee S, Nayak P, Fritz D, Vuyyuru MR,

- Aslanides J, Vyas N, Wicke M, Ma X, Eltyshv E, Martin N, Cate H, Manyika J, Amiri K, Kim Y, Xiong X, Kang K, Luisier F, Tripuraneni N, Madras D, Guo M, Waters A, Wang O, Ainslie J, Baldridge J, Zhang H, Pruthi G, Bauer J, Yang F, Mansour R, Gelman J, Xu Y, Polovets G, Liu J, Cai H, Chen W, Sheng X, Xue E, Ozair S, Angermueller C, Li X, Sinha A, Wang W, Wiesinger J, Koukoumidis E, Tian Y, Iyer A, Gurumurthy M, Goldenson M, Shah P, Blake MK, Yu H, Urbanowicz A, Palomaki J, Fernando C, Durden K, Mehta H, Momchev N, Rahimtoroghi E, Georgaki M, Raul A, Ruder S, Redshaw M, Lee J, Zhou D, Jalan K, Li D, Hechtman B, Schuh P, Nasr M, Milan K, Mikulik V, Franco J, Green T, Nguyen N, Kelley J, Mahendru A, Hu A, Howland J, Vargas B, Hui J, Bansal K, Rao V, Ghiya R, Wang E, Ye K, Sarr JM, Preston MM, Elish M, Li S, Kaku A, Gupta J, Pasupat I, Juan D-C, Someswar M, M T, Chen X, Amini A, Fabrikant A, Chu E, Dong X, Muthal A, Buthpitiya S, Jauhari S, Hua N, Khandelwal U, Hitron A, Ren J, Rinaldi L, Drath S, Dabush A, Jiang N-J, Godhia H, Sachs U, Chen A, Fan Y, Taitelbaum H, Noga H, Dai Z, Wang J, Liang C, Hamer J, Ferng C-S, Elkind C, Atias A, Lee P, Listik V, Carlen M, Kerkhof J van de, Pikus M, Zaher K, Müller P, Zykova S, Stefanec R, Gatsko V, Hirnschall C, Sethi A, Xu XF, Ahuja C, Tsai B, Stefanoiu A, Feng B, Dhandhanika K, Katyal M, Gupta A, Parulekar A, Pitta D, Zhao J, Bhatia V, Bhavnani Y, Alhadlaq O, Li X, Danenberg P, Tu D, Pine A, Filippova V, Ghosh A, Limonchik B, Urala B, Lanka CK, Clive D, Sun Y, Li E, Wu H, Hongtongsak K, Li I, Thakkar K, Omarov K, Majmundar K, Alverson M, Kucharski M, Patel M, Jain M, Zabelin M, Pelagatti P, Kohli R, Kumar S, Kim J, Sankar S, Shah V, Ramachandruni L, Zeng X, Bariach B, Weidinger L, Vu T, Andreev A, He A, Hui K, Kashem S, Subramanya A, Hsiao S, Hassabis D, Kavukcuoglu K, Sadovsky A, Le Q, Strohman T, Wu Y, Petrov S, Dean J, Vinyals O. Gemini: A Family of Highly Capable Multimodal Models. arXiv; 2025. doi: 10.48550/arXiv.2312.11805
215. Introducing Meta Llama 3: The most capable openly available LLM to date. Meta AI. Available from: <https://ai.meta.com/blog/meta-llama-3/> [accessed Aug 7, 2024]
 216. Makhni S, Rico J, Cerrato P, Hill B, Overgaard S, Wu J, Fairbanks J, Tripp R, Redziniak J, Blundo R, Otley C, Halamka J. A Comprehensive Approach to Responsible AI Development and Deployment. *Mayo Clin Proc Digit Health* 2025 Nov 4;3(4):100294. PMID:41282583
 217. Li Y, Tian J, Xu A, Greiner R, Hayward J, Greenshaw AJ, Cao B. Maturity Framework for Operationalizing Machine Learning Applications in Health Care: Scoping Review. *J Med Internet Res* 2025 Sept 19;27:e66559. PMID:40971600
 218. Kumar P, Bhatnagar R, Gaur K, Bhatnagar A. Classification of Imbalanced Data: Review of Methods and Applications. *IOP Conference Series: Materials Science and Engineering* 2021 Mar 1;1099:012077. doi: 10.1088/1757-899X/1099/1/012077
 219. Wang Y, Hou Y, Che W, Liu T. From static to dynamic word representations: a survey. *Int J Mach Learn & Cyber* 2020 July 1;11(7):1611–1630. doi: 10.1007/s13042-020-01069-8

Appendices

Appendix 3.1 Data Preprocessing Procedures.

Text Normalization

Two types of features were extracted from the free text section of PSE reports including static text representation and contextual text representation. Static text representation requires text normalization to recognize the same word in different forms (i.e., singular vs. plural), this is achieved by a series of data cleaning procedures, including converting all words to lowercase, removing non-alphabetical characters, stop words, as well as applying stemming. In contrast, contextual text presentation does not require text normalization. This is because contextual text representation is generated by transformer-based NN that are trained on a substantial amount of raw text, making it resilient to variations in words such as casing, word tenses, and plurality.

Feature Extraction

Post text normalization, three common static text representations were obtained including the bag of words (BOW), term frequency-inverse document frequency (TF-IDF), and global vectors (GloVe). BOW is a vector representation of the text document where the total number of occurrences of each word in the text document is used as a feature. TF-IDF is a vector format representation of a text document that reflects how relevant a word is to a text document in the entire dataset. The static text representation produced from BOW and TF-IDF in this study both used n-grams ranging from 1 to 3. GloVe represents each word with a vector that captures the semantic relationship between words and the representation of the text document is then obtained by taking the average of each word's vector.

Five contextual text representations were obtained by passing the free text section of PSE reports into pre-trained transformer-based NNs, including BERT-base, PubMed-BERT-base, xlm-RoBERTa-base, RoBERTa-large, and RoBERTa-base. BERT is a transformer-based NN which stands for Bidirectional Encoder Representations from Transformers, it was trained in a self-supervision fashion using a large corpus of raw text with two tasks, masked language modeling and next sentence prediction. BERT is capable of generating meaningful representations for text documents considering the semantic context within the text. RoBERTa is the optimized version of BERT, it builds on BERT and modifies key hyperparameters, removing the next sentence

prediction task. The difference between contextual text representations used in this study is the number of parameters in the transformer-based NN (i.e., RoBERTa-large: 255 million parameters vs. RoBERTa-base: 125 million parameters) and domain of raw text used to train the NN (i.e., PubMed-BERT-base is trained on abstracts from PubMed). All of the text representations have been standardized to ensure that the resulting ML classifier is not overly sensitive to the magnitudes of different features, which can help prevent issues such as overfitting.

Data Splitting

To train the ML classifier for PSE reports classification, we randomly split the data into 80% for training and 20% for testing using stratified sampling, which is the standard methodology for data splitting when training ML classifiers. This approach helps to prevent overfitting and improve the generalizability of ML classifiers. By splitting the data in a stratified fashion, we ensured the testing set preserved the original distribution of event types, thus maintaining representative of the real-world scenario. Instead of creating a separate validation set, we used five-fold cross-validation during hyperparameter tuning to provide the ML classifier with more access to data during training.

Data Augmentation

The distribution of PSE reports' event type in this study is imbalanced in this study. Imbalanced data can lead to poor classification performance in the minority classes [218]. Therefore, we utilized the synthetic minority oversampling technique (SMOTE) to address the imbalance problem. SMOTE is an oversampling technique where synthetic data are created for the minority class to achieve a balanced distribution of classes, thus improving the classifier's sensitivity to the minority class [129]. Only the training data set was augmented with SMOTE, while the testing set maintained its original distribution.

Appendix 3.2 Evaluation Metrics for Examining Machine Learning Classifiers.

Evaluation Metrics Definitions

- Accuracy - the overall percentage of PSE reports that the classifier has correctly classified.
- Top-2 accuracy - a variation of accuracy that measures the percentage of PSE reports where the true event type matches with one of the top two highest probability event types predicted by the classifier.
- Precision - answers how many PSE reports classified as one specific event type belong to that event type.
- Recall (also known as sensitivity) - represents the proportion of PSE reports that are correctly classified as their true event type.
- F1 - the harmonic mean of the precision and recall, which gives a whole picture of classifiers' performance on both precision and recall.
- Area Under the Receiver Operating Characteristic Curve (AUCROC) - measures classifiers' ability to distinguish between event types by measuring the area under the curve which plots sensitivity versus (1-specificity).

Mathematical Definitions

True positive (TP): number of PSE reports correctly classified into a given event type.

False positive (FP): number of PSE reports falsely classified into a given event type.

True negative (TN): number of PSE reports correctly classified not to be in a given event type.

False negative (FN): number of PSE reports falsely classified as not being in a given event type.

N: Total number of PSE reports classified.

$$\text{Accuracy} = (TP + TN) / N$$

$$\text{Top-2 Accuracy} = (\text{Number of data has correct event types in the top 2 highest probability predicted event types}) / N$$

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall} = TP / (TP + FN)$$

$$F1 = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

Appendix 3.3 The Performance of ML Classifiers in Classifying PSE Report Event Type.

Evaluation metrics	Text Representation / ML Classifier	MLR	SVM	XGB	RF	KNN	LGB	MLP
Accuracy (%)	TF-IDF	66.67	65.08	63.49	57.94	46.83	63.49	65.08
	BOW	54.76	55.56	61.11	63.49	36.51	55.56	43.65
	GloVe	53.17	62.70	53.97	58.73	56.35	54.76	60.32
	BERT-base	62.70	60.32	51.59	52.38	42.86	55.56	65.87
	xlm-RoBERTa-base	68.25	69.05	70.63	63.49	53.97	69.05	70.63
	RoBERTa-base	66.67	75.40	66.67	65.87	50.00	69.05	68.25
	RoBERTa-large	69.05	59.52	58.73	56.35	43.65	55.56	66.67
	PubMed-BERT-base	55.56	65.87	59.52	61.11	51.59	61.11	62.70
Top 2 Accuracy (%)	TF-IDF	85.71	84.92	84.13	81.75	61.11	80.95	83.33
	BOW	80.16	72.22	84.92	80.95	49.21	70.63	65.08
	GloVe	74.60	79.37	73.02	79.37	65.87	76.98	82.54
	BERT-base	83.33	78.57	73.02	71.43	55.56	78.57	85.71
	xlm-RoBERTa-base	84.92	82.54	82.54	74.60	70.63	80.16	88.10
	RoBERTa-base	84.13	84.92	81.75	80.95	65.08	80.16	84.92
	RoBERTa-large	80.95	81.75	81.75	76.19	59.52	78.57	81.75
	PubMed-BERT-base	82.54	85.71	76.98	77.78	62.70	83.33	83.33
Precision	TF-IDF	0.640	0.657	0.624	0.579	0.707	0.627	0.654
	BOW	0.619	0.595	0.613	0.671	0.386	0.574	0.684

	GloVe	0.540	0.629	0.571	0.600	0.627	0.569	0.621
	BERT-base	0.641	0.635	0.526	0.524	0.565	0.552	0.658
	xlm-RoBERTa-base	0.705	0.694	0.715	0.668	0.612	0.703	0.733
	RoBERTa-base	0.706	0.757	0.666	0.665	0.642	0.701	0.710
	RoBERTa-large	0.711	0.612	0.576	0.556	0.545	0.555	0.698
	PubMed-BERT-base	0.571	0.661	0.616	0.648	0.652	0.627	0.648
Recall	TF-IDF	0.667	0.651	0.635	0.587	0.476	0.635	0.651
	BOW	0.548	0.563	0.619	0.635	0.373	0.563	0.444
	GloVe	0.532	0.627	0.540	0.595	0.571	0.548	0.611
	BERT-base	0.627	0.611	0.516	0.524	0.429	0.563	0.659
	xlm-RoBERTa-base	0.683	0.698	0.714	0.635	0.540	0.698	0.714
	RoBERTa-base	0.667	0.754	0.667	0.659	0.500	0.698	0.690
	RoBERTa-large	0.698	0.603	0.595	0.571	0.444	0.563	0.667
	PubMed-BERT-base	0.556	0.659	0.603	0.619	0.516	0.619	0.627
F1	TF-IDF	0.631	0.622	0.617	0.537	0.451	0.625	0.620
	BOW	0.570	0.570	0.602	0.630	0.350	0.565	0.427
	GloVe	0.533	0.620	0.549	0.593	0.572	0.553	0.610
	BERT-base	0.629	0.602	0.517	0.523	0.400	0.552	0.657
	xlm-RoBERTa-base	0.685	0.691	0.712	0.640	0.535	0.698	0.716
	RoBERTa-base	0.676	0.753	0.665	0.651	0.509	0.697	0.696
	RoBERTa-large	0.702	0.598	0.581	0.553	0.449	0.550	0.679

Appendix 4.1 Feature Extraction Procedures for Fixed-Representation-Based Models.

Fixed-representation models require data to be transformed into text representations before being used as model inputs. To this end, we extracted two categories of features from the incident descriptions: static and contextual text representations. Static representations are derived using traditional vectorization methods that often result in sparse, high-dimensional feature spaces and do not capture word order or semantic relationships [219]. In contrast, contextual representations capture the semantic meaning of words in relation to their surrounding text, using embeddings generated by pretrained transformer-based language models [32,56].

For static text representations, preprocessing is essential to ensure consistent input formatting and to enable the model to recognize words in different forms. We applied standard normalization techniques, including removing non-alphabetic characters and stop words, converting text to lowercase, and applying stemming to reduce words to their root forms. In contrast, contextual text representations do not require such preprocessing, as they are generated by transformer-based models pretrained on large-scale raw text corpora [32,33]. These models inherently account for linguistic variations such as plurality and tense.

A total of six distinct text representation methods were employed in this study (Table 1). For static representations, we utilized two widely adopted techniques: Bag-of-Words (BOW) and Term Frequency Inverse Document Frequency (TF-IDF). BOW represents each document as a vector based on the frequency of individual words, without considering their order or context. TF-IDF assigns weights to terms based on their importance across the corpus, capturing the relative significance of words within documents. In both methods, we incorporated n-gram features ranging from unigrams to trigrams to capture local word patterns. Four contextual text representations were generated by passing the incident descriptions through pretrained transformer-based language models, specifically BERT, ROBERTA, GATORTRON, and BIOBERT [32,33,74,124]. BERT and ROBERTA are general-purpose models known for their robust language understanding capabilities, while GATORTRON and BIOBERT have been pretrained on biomedical and clinical corpora, making them suitable for representing medical narratives. All extracted representations were standardized to mitigate scale differences across features, ensuring balanced feature contribution and reducing overfitting risk.

Table 1. Static and contextual text representations used for developing fixed-representation models in this study. The suffix “_R” indicates a text representation derived from the corresponding method or model.

Representation categories	Full model name	Abbreviations
Static	Bag-of-words	BOW_R
	Term Frequency Inverse Document Frequency	TF-IDF_R
Contextual	bert-base-uncased	BERT_R
	roberta-base	ROBERTA_R
	biobert-v1.1	BIOBERT_R
	gatortron-base	GATORTRON_R

Appendix 4.2 Hyperparameter Settings, Tuning Ranges, and Final Selected Configurations for Feature-Based and Transformer-Based Models.

This appendix presents the hyperparameter settings and tuning ranges used during model development for both feature-based and transformer-based models. Tables 1–4 summarize the settings and tuning ranges for the feature-based models (XGB-RA, LGBM-RA, XGB-CL, and LGBM-CL). Tables 5 and 6 present the corresponding information for transformer-based ranking and classification models. Table 7 reports the final selected hyperparameter configuration for the best-performing model within each major model family.

Table 1. Hyperparameter settings and tuning ranges for the XGB-RA model (XGBoost for ranking).

Parameter Name	Description	Tuning Range or Fixed Value
objective	Specifies the learning objective and loss function used to optimize the ranking model.	['rank:ndcg', 'rank:pairwise']
lambdarank_pair_method	Defines the method used to construct item pairs for pairwise ranking.	['mean']
lambdarank_num_pair_per_sample	Specifies the number of document pairs generated for each document.	[5, 10]
learning_rate	Controls the contribution of each tree when added to the model. Smaller values improve generalization but require more boosting rounds.	[0.01, 0.05]

n_estimators	Total number of boosting rounds during model training.	[50, 150]
max_depth	Maximum depth of each decision tree. Increasing depth allows the model to capture more complex patterns.	[3, 6]
subsample	Fraction of training instances randomly sampled for each boosting iteration. Helps prevent overfitting.	[0.8, 1.0]
reg_alpha	L1 regularization term which encourages sparsity in the model and helps reduce overfitting.	[0, 1.0]
reg_lambda	L2 regularization term which penalizes large weights to help prevent overfitting.	[0, 1.0]
gamma	Minimum loss reduction required to make a further split on a leaf node. Helps prevent overfitting.	[0, 1.0]

Table 2. Hyperparameter settings and tuning ranges for the LGBM-RA model (LightGBM for ranking).

Parameter Name	Description	Tuning Range or Fixed Value
objective	Specifies the learning objective and loss function used to optimize the ranking model.	['lambda_rank']

boosting_type	Specifies the type of boosting algorithm to use, which determines how new trees are added during training.	['gbdt']
num_boost_round	Total number of boosting rounds during model training.	[50, 150]
learning_rate	Controls the contribution of each tree when added to the model. Smaller values improve generalization but require more boosting rounds.	[0.01, 0.05]
max_depth	Maximum depth of each decision tree. Increasing depth allows the model to capture more complex patterns.	[3, 6]
feature_fraction	Fraction of training features randomly sampled for each boosting iteration. Helps prevent overfitting.	[0.8, 1.0]
reg_alpha	L1 regularization term which encourages sparsity in the model and helps reduce overfitting.	[0, 1.0]
reg_lambda	L2 regularization term which penalizes large weights to help prevent overfitting.	[0, 1.0]

num_leaves	Maximum number of leaves in each tree; controls the complexity of the model.	[10, 25]
min_data_in_leaf	Minimum number of samples required in a leaf node. Larger values can prevent the model from creating overly specific leaf nodes, reducing overfitting.	[10, 20]

Table 3. Hyperparameter settings and tuning ranges for the XGB-CL model (XGBoost for classification).

Parameter Name	Description	Tuning Range or Fixed Value
objective	Specifies the learning objective and loss function used to optimize the classification model.	['multi:softprob']
learning_rate	Controls the contribution of each tree when added to the model. Smaller values improve generalization but require more boosting rounds.	[0.01, 0.05]
n_estimators	Total number of boosting rounds during model training.	[50, 150]
max_depth	Maximum depth of each decision tree. Increasing depth allows the model to capture more complex patterns.	[3, 6]

subsample	Fraction of training instances randomly sampled for each boosting iteration. Helps prevent overfitting.	[0.8, 1.0]
reg_alpha	L1 regularization term which encourages sparsity in the model and helps reduce overfitting.	[0, 1.0]
reg_lambda	L2 regularization term which penalizes large weights to help prevent overfitting.	[0, 1.0]
gamma	Minimum loss reduction required to make a further split on a leaf node. Helps prevent overfitting.	[0, 1.0]

Table 4. Hyperparameter settings and tuning ranges for the LGBM-CL model (LightGBM for classification).

Parameter Name	Description	Tuning Range or Fixed Value
objective	Specifies the learning objective and loss function used to optimize the classification model.	['multiclass']
boosting_type	Specifies the type of boosting algorithm to use, which determines how new trees are added during training.	['gbdt']

num_boost_round	Total number of boosting rounds during model training.	[50, 150]
learning_rate	Controls the contribution of each tree when added to the model. Smaller values improve generalization but require more boosting rounds.	[0.01, 0.05]
max_depth	Maximum depth of each decision tree. Increasing depth allows the model to capture more complex patterns.	[3, 6]
feature_fraction	Fraction of training features randomly sampled for each boosting iteration. Helps prevent overfitting.	[0.8, 1.0]
reg_alpha	L1 regularization term which encourages sparsity in the model and helps reduce overfitting.	[0, 1.0]
reg_lambda	L2 regularization term which penalizes large weights to help prevent overfitting.	[0, 1.0]
num_leaves	Maximum number of leaves in each tree; controls the complexity of the model.	[10, 25]
min_data_in_leaf	Minimum number of samples required in a leaf node. Larger values can prevent the model from	[10, 20]

	creating overly specific leaf nodes, reducing overfitting.	
--	--	--

Table 5. Hyperparameter settings and tuning ranges for transformer-based ranking models.

Parameter Name	Description	Tuning Range or Fixed Value
objective	Specifies the learning objective and loss function used to optimize the ranking model.	['lambdarank', 'listnet']
learning rate	Step size controlling how much the model parameters are updated at each optimization step.	[1e-5, 5e-5]
batch_size	Number of training examples processed together in a single forward/backward pass before updating model parameters.	[8, 16]
warmup_ratio	Fraction of the total training steps used to gradually increase the learning rate from zero to the target value, helping to stabilize early training.	[0.10]
num_train_epochs	Number of complete passes through the entire training dataset during model training.	[3, 5]
weight_decay	L2 regularization coefficient applied to model weights to reduce overfitting by penalizing large parameter values.	[0.01]

Table 6. Hyperparameter settings and tuning ranges for transformer-based classification models.

Parameter Name	Description	Tuning Range or Fixed Value
objective	Specifies the learning objective and loss function used to optimize the classification model (e.g., cross-entropy loss for multi-class classification).	['Cross-entropy loss']
learning rate	Step size controlling how much the model parameters are updated at each optimization step.	[1e-5, 5e-5]
batch_size	Number of training examples processed together in a single forward/backward pass before updating model parameters.	[8, 16]
warmup_ratio	Fraction of the total training steps used to gradually increase the learning rate from zero to the target value, helping to stabilize early training.	[0.10]
num_train_epochs	Number of complete passes through the entire training dataset during model training.	[3, 5]
weight_decay	L2 regularization coefficient applied to model weights to reduce overfitting by penalizing large parameter values.	[0.01]

Table 7. Final selected hyperparameter configurations for the best-performing model–representation combination within each model family across the two evaluation datasets.

Model Family	Best-Performing Model	Final Selected Hyperparameters
Panel A. Committee-adjudicated dataset		
Feature-based ranking (XGB-RA)	XGB-RA + GATORTRON	objective=rank:ndcg lambdarank_pair_method=mean lambdarank_num_pair_per_sample=10 learning_rate=0.01 n_estimators=150; max_depth=6 subsample=0.8 reg_alpha=0 reg_lambda=1.0 gamma=0
Feature-based ranking (LGBM-RA)	LGBM-RA + GATORTRON	objective=lambdarank boosting_type=gbdt num_boost_round=50 learning_rate=0.05 max_depth=6 feature_fraction=0.8 reg_alpha=0

		reg_lambda=1.0 num_leaves=25 min_data_in_leaf=20
Feature-based classification (XGB-CL)	XGB-CL + BAG	objective=multi:softprob learning_rate=0.01 n_estimators=150 max_depth=6 subsample=0.8 reg_alpha=1.0 reg_lambda=1.0 gamma=0
Feature-based classification (LGBM-CL)	LGBM-CL + ROBERTA	objective=multiclass boosting_type=gbdt num_boost_round=150 learning_rate=0.01 max_depth=6 feature_fraction=0.8 reg_alpha=1.0 reg_lambda=1.0 num_leaves=25

		min_data_in_leaf=20
Transformer-based ranking	LLAMA3.1_RA	objective=lambdarank learning_rate=5e-5 batch_size=8 warmup_ratio=0.10 num_train_epochs=3 weight_decay=0.01
Transformer-based classification	LLAMA3.1_CL	objective=cross-entropy loss learning_rate=1e-5 batch_size=8 warmup_ratio=0.10 num_train_epochs=3 weight_decay=0.01
Panel B. Full incident reporting system dataset		
Feature-based ranking (XGB-RA)	XGB-RA + GATORTRON	objective=rank:ndcg lambdarank_pair_method=mean lambdarank_num_pair_per_sample=10 learning_rate=0.05 n_estimators=150 max_depth=6

		subsample=0.8 reg_alpha=1.0 reg_lambda=1.0 gamma=0
Feature-based ranking (LGBM-RA)	LGBM-RA + GATORTRON	objective=lambdarank boosting_type=gbdt num_boost_round=150 learning_rate=0.01 max_depth=6 feature_fraction=1.0 reg_alpha=1.0 reg_lambda=1.0 num_leaves=25 min_data_in_leaf=20
Feature-based classification (XGB-CL)	XGB-CL + BERT	objective=multi:softprob learning_rate=0.05 n_estimators=150 max_depth=3 subsample=0.8 reg_alpha=1.0

		reg_lambda=1.0 gamma=1.0
Feature-based classification (LGBM-CL)	LGBM-CL + BERT	objective=multiclass boosting_type=gbdt num_boost_round=50 learning_rate=0.05 max_depth=6 feature_fraction=0.8 reg_alpha=1.0 reg_lambda=1.0 num_leaves=25 min_data_in_leaf=20
Transformer-based ranking	LLAMA3.1_RA	objective=lambdarank learning_rate=5e-5 batch_size=8 warmup_ratio=0.10 num_train_epochs=5 weight_decay=0.01
Transformer-based classification	LLAMA3.1_CL	objective=cross-entropy loss learning_rate=5e-5

		<code>batch_size=8</code> <code>warmup_ratio=0.10</code> <code>num_train_epochs=3</code> <code>weight_decay=0.01</code>
--	--	--

Appendix 4.3 Mathematical Definitions of Evaluation Metrics.

In this study, we evaluated model performance using established ranking metrics, including average precision (AP), normalized discounted cumulative gain (NDCG) at multiple cutoff thresholds ($k = 10, 25, 50, 100, 200$), as well as the precision and recall computed at the same cutoffs. The mathematical definitions of these metrics are provided below.

Average Precision (AP):

Average Precision is calculated as the mean of the precision values at each rank position where a relevant item (i.e., high-severity PSE) is retrieved. This metric quantifies the model's ability to place relevant items near the top of the ranked list, thus reflecting the effectiveness of the ranking.

Formula:

$$AP = \frac{1}{n} \sum_{i=1}^N Precision(i) * relevance(i)$$

Definitions:

- AP: Average Precision.
- n : Total number of relevant items (i.e., high severity PSEs) in the dataset.
- N : Total number of items in the ranked list.
- $Precision(i)$: Precision at rank position i .
- $relevance(i)$: Binary relevance indicator (1 if the item at rank i is relevant, 0 otherwise).

Normalized Discounted Cumulative Gain (NDCG@ k):

NDCG@ k quantifies the quality of a ranked list (length= k) by accounting for both the relevance of items (i.e., severity of PSEs) and their positions in the ranking. Items ranked higher contribute more to the overall score, with their relevance scores discounted logarithmically based on rank. The score is normalized by the Ideal Discounted Cumulative Gain (IDCG@ K), which represents the maximum possible DCG@ k obtained by ranking all relevant items at the top of the list.

Formula:

$$NDCG@k = \frac{DCG@k}{IDCG@k}$$

$$DCG@k = \sum_{i=1}^k \frac{2^{r_i} - 1}{\log_2(i + 1)}$$

$$IDCG@k = \sum_{i=1}^k \frac{2^{r_i^*} - 1}{\log_2(i + 1)}$$

Definitions:

- $NDCG@k$: Normalized Discounted Cumulative Gain at rank k .
- $DCG@k$: Discounted Cumulative Gain at rank k .
- $IDCG@k$: Ideal Discounted Cumulative Gain at rank k .
- k : Cutoff threshold (number of top-ranked items considered).
- r_i : relevance score at rank i in the predicted order.
- r_i^* : relevance score at rank i in the ideal order (sorted by the highest relevance first)

Precision@k:

Precision at rank k quantifies the proportion of relevant items (i.e., high-severity PSEs) among the top k positions in the ranked list. This metric evaluates how accurately the model identifies relevant items within the top k positions.

Formula:

$$Precision@k = \frac{1}{k} \sum_{i=1}^k relevance(i)$$

Definitions:

- $Precision@k$: Precision at rank k .
- k : Cutoff threshold (number of top-ranked items considered).
- $relevance(i)$: Binary relevance indicator (1 if the item at rank i is relevant, 0 otherwise).

Recall@k:

Recall at rank k measures the proportion of all relevant items (i.e., high-severity PSEs) that are retrieved within the top k positions in the ranked list. The metric reflects the model's ability to retrieve all relevant items at different cutoff thresholds.

Formula:

$$Recall@k = \frac{1}{n} \sum_{i=1}^k relevance(i)$$

Definitions:

- Recall@k: Recall at rank k.
- n: Total number of relevant items in the dataset.
- k: Cutoff threshold (number of top-ranked items considered).
- relevance(i): Binary relevance indicator (1 if the item at rank i is relevant, 0 otherwise).

Appendix 4.4 Demonstration of the Designed Severity Ranking Interface.

We designed a Severity Ranking Interface that integrates the best-performing prioritization model (BIOBERT-RA) to assist patient safety specialists in triaging high-severity events for institutional review. Figure 1 below illustrates the layout of the interface. The left panel displays safety events ranked by severity percentile, which is derived from the model's normalized severity scores across all reported events. This percentile-based ranking allows specialists to efficiently focus on the most critical cases.

When a specific event is selected, the interface presents both structured information (i.e., event type, date, location) and unstructured information (i.e., incident description) typically documented in a PSE report. Based on this information, specialists can assess whether the event represents a potential high-severity case requiring further investigation. A feedback mechanism, located in the lower-right corner, enables users to indicate agreement or disagreement with the model's ranking, providing valuable input for future model refinement. The upper section of the interface summarizes key operational indicators, including the number of unreviewed events and the current investigative capacity of the multidisciplinary review committee, thereby facilitating data-driven workload management and prioritization.

Severity Ranking Interface		Number of unreviewed events: 29																																																	
		Institutional investigation capacity for July 20XX: 11 out of 13																																																	
<table border="1"> <thead> <tr> <th colspan="2">Severity ranking percentile</th> </tr> </thead> <tbody> <tr><td>PSE-1</td><td>96th</td></tr> <tr><td>PSE-2</td><td>95th</td></tr> <tr><td>PSE-3</td><td>93rd</td></tr> <tr><td>PSE-4</td><td>87th</td></tr> <tr><td>PSE-5</td><td>82nd</td></tr> <tr><td>PSE-6</td><td>74th</td></tr> <tr><td>PSE-7</td><td>70th</td></tr> <tr><td>PSE-8</td><td>68th</td></tr> <tr><td>.....</td><td>.....</td></tr> <tr><td>.....</td><td>.....</td></tr> <tr><td>PSE-27</td><td>14th</td></tr> <tr><td>PSE-28</td><td>9th</td></tr> <tr><td>PSE-29</td><td>5th</td></tr> </tbody> </table>		Severity ranking percentile		PSE-1	96th	PSE-2	95th	PSE-3	93rd	PSE-4	87th	PSE-5	82nd	PSE-6	74th	PSE-7	70th	PSE-8	68th					PSE-27	14th	PSE-28	9th	PSE-29	5th	<table border="1"> <thead> <tr> <th colspan="2">Patient safety event information</th> </tr> </thead> <tbody> <tr> <td>Incident date: 20XX-July-1st</td> <td>Incident description: A 63-year-old male with chronic ischemia of the left lower extremity was scheduled for below-knee amputation. Due to a documentation error and failure to perform adequate surgical site verification, a right-sided amputation was performed. The error was identified postoperatively. The patient sustained irreversible loss of a healthy limb, resulting in permanent functional impairment.</td> </tr> <tr> <td>Incident time: 9:34 pm</td> <td></td> </tr> <tr> <td>Department: Surgery - Vascular</td> <td></td> </tr> <tr> <td>Patient Medical Record Number: 1E49U7K9</td> <td></td> </tr> <tr> <td>Reporter Employee ID: N879123</td> <td>Is this a high-severity safety event that needs formal institutional investigation?</td> </tr> <tr> <td>Reporter Role: Operating room nurse</td> <td> Yes No </td> </tr> <tr> <td>Event type: Surgical procedure error</td> <td>Do you agree with the AI's ranking: top 4 % most severe?</td> </tr> <tr> <td></td> <td> Agree Disagree </td> </tr> <tr> <td></td> <td> Submit </td> </tr> </tbody> </table>		Patient safety event information		Incident date: 20XX-July-1st	Incident description: A 63-year-old male with chronic ischemia of the left lower extremity was scheduled for below-knee amputation. Due to a documentation error and failure to perform adequate surgical site verification, a right-sided amputation was performed. The error was identified postoperatively. The patient sustained irreversible loss of a healthy limb, resulting in permanent functional impairment.	Incident time: 9:34 pm		Department: Surgery - Vascular		Patient Medical Record Number: 1E49U7K9		Reporter Employee ID: N879123	Is this a high-severity safety event that needs formal institutional investigation?	Reporter Role: Operating room nurse	Yes No	Event type: Surgical procedure error	Do you agree with the AI's ranking: top 4 % most severe?		Agree Disagree		Submit
Severity ranking percentile																																																			
PSE-1	96th																																																		
PSE-2	95th																																																		
PSE-3	93rd																																																		
PSE-4	87th																																																		
PSE-5	82nd																																																		
PSE-6	74th																																																		
PSE-7	70th																																																		
PSE-8	68th																																																		
.....																																																			
.....																																																			
PSE-27	14th																																																		
PSE-28	9th																																																		
PSE-29	5th																																																		
Patient safety event information																																																			
Incident date: 20XX-July-1st	Incident description: A 63-year-old male with chronic ischemia of the left lower extremity was scheduled for below-knee amputation. Due to a documentation error and failure to perform adequate surgical site verification, a right-sided amputation was performed. The error was identified postoperatively. The patient sustained irreversible loss of a healthy limb, resulting in permanent functional impairment.																																																		
Incident time: 9:34 pm																																																			
Department: Surgery - Vascular																																																			
Patient Medical Record Number: 1E49U7K9																																																			
Reporter Employee ID: N879123	Is this a high-severity safety event that needs formal institutional investigation?																																																		
Reporter Role: Operating room nurse	Yes No																																																		
Event type: Surgical procedure error	Do you agree with the AI's ranking: top 4 % most severe?																																																		
	Agree Disagree																																																		
	Submit																																																		

Figure 1. The layout of the Severity Ranking Interface is designed to support real-time triage and prioritization of PSEs. All event information displayed in this example is fabricated for illustrative purposes only and does not represent real patient data. No personally identifiable or institution-specific information is shown.

Appendix 4.5 Evaluation Results of All Developed Ranking and Classification Models.

The tables below summarize the performance of all developed ranking and classification models. Tables 1–4 report the performance and corresponding 95% confidence intervals for the Fixed-Representation-Based models, including XGB-CL, XGB-RA, LGBM-CL, and LGBM-RA. Tables 5 and 6 present the performance results for the transformer-based models, with Table 5 showing the ranking models and Table 6 showing the classification models.

Table 1. Performance of the XGB-CL model across multiple evaluation metrics with 95% confidence intervals.

Metric	XGB_TFIDF	XGB_BAG	XGB_BIOERT	XGB_BERT	XGB_ROBERTA	XGB_GATORTRON
Panel A. Committee-adjudicated dataset						
Average Precision	0.62±0.07	0.68±0.06	0.64±0.08	0.62±0.08	0.65±0.07	0.67±0.07
NDCG@10	0.98±0.12	1.00±0.00	0.83±0.17	0.76±0.26	1.00±0.14	0.92±0.16
NDCG@20	0.90±0.12	1.00±0.01	0.89±0.12	0.81±0.17	0.88±0.13	0.95±0.10
NDCG@50	0.88±0.08	0.94±0.05	0.91±0.07	0.85±0.11	0.88±0.08	0.94±0.07
NDCG@100	0.84±0.06	0.89±0.06	0.84±0.07	0.84±0.08	0.87±0.07	0.89±0.06
NDCG@200	0.77±0.04	0.78±0.05	0.77±0.05	0.78±0.05	0.79±0.05	0.81±0.04
Precision@10	0.90±0.22	1.00±0.00	0.70±0.30	0.70±0.32	1.00±0.20	0.90±0.19
Precision@20	0.80±0.20	1.00±0.04	0.85±0.17	0.80±0.18	0.80±0.20	0.95±0.11
Precision@50	0.76±0.13	0.88±0.10	0.86±0.10	0.84±0.11	0.82±0.10	0.90±0.08
Precision@100	0.72±0.09	0.81±0.11	0.74±0.12	0.75±0.11	0.78±0.10	0.76±0.11
Precision@200	0.58±0.07	0.60±0.07	0.59±0.08	0.57±0.08	0.59±0.08	0.60±0.08
Recall@10	0.04±0.01	0.05±0.00	0.03±0.01	0.03±0.02	0.05±0.01	0.04±0.01
Recall@20	0.08±0.02	0.10±0.01	0.08±0.02	0.08±0.02	0.08±0.02	0.09±0.01

Recall@50	0.18±0.03	0.21±0.03	0.21±0.03	0.20±0.03	0.20±0.03	0.22±0.03
Recall@100	0.35±0.04	0.39±0.05	0.36±0.05	0.36±0.05	0.38±0.05	0.37±0.04
Recall@200	0.56±0.05	0.58±0.06	0.57±0.06	0.55±0.06	0.57±0.06	0.57±0.05
Panel B. Full incident reporting system dataset						
Average Precision	0.43±0.07	0.49±0.07	0.54±0.06	0.55±0.07	0.52±0.08	0.50±0.07
NDCG@10	0.75±0.24	1.00±0.10	1.00±0.02	0.95±0.07	0.90±0.07	0.66±0.28
NDCG@20	0.81±0.16	0.92±0.11	0.97±0.07	0.92±0.07	0.90±0.08	0.73±0.20
NDCG@50	0.81±0.11	0.90±0.09	0.93±0.06	0.91±0.06	0.86±0.08	0.78±0.13
NDCG@100	0.79±0.09	0.82±0.07	0.81±0.07	0.85±0.09	0.81±0.08	0.76±0.10
NDCG@200	0.59±0.08	0.65±0.06	0.69±0.05	0.68±0.06	0.71±0.12	0.69±0.07
Precision@10	0.70±0.28	1.00±0.15	1.00±0.03	0.80±0.23	0.80±0.22	0.60±0.25
Precision@20	0.80±0.16	0.90±0.14	0.95±0.11	0.80±0.16	0.80±0.17	0.60±0.18
Precision@50	0.80±0.11	0.86±0.11	0.86±0.11	0.82±0.11	0.76±0.12	0.74±0.13
Precision@100	0.69±0.10	0.67±0.10	0.68±0.09	0.72±0.13	0.68±0.10	0.66±0.10
Precision@200	0.45±0.08	0.48±0.08	0.52±0.07	0.52±0.07	0.55±0.08	0.53±0.07
Recall@10	0.03±0.01	0.05±0.01	0.05±0.01	0.04±0.01	0.04±0.01	0.03±0.01
Recall@20	0.08±0.02	0.09±0.02	0.09±0.01	0.08±0.02	0.08±0.02	0.06±0.02
Recall@50	0.19±0.03	0.21±0.03	0.21±0.03	0.20±0.04	0.18±0.03	0.18±0.04
Recall@100	0.33±0.05	0.32±0.05	0.33±0.04	0.35±0.05	0.33±0.05	0.32±0.05
Recall@200	0.43±0.06	0.46±0.07	0.50±0.06	0.50±0.05	0.52±0.06	0.51±0.06

Table 2. Performance of the XGB-RA model across multiple evaluation metrics with 95% confidence intervals.

Metric	XGB_TFIDF	XGB_BAG	XGB_BIOERT	XGB_BERT	XGB_ROBERTA	XGB_GATORTRON
Panel A. Committee-adjudicated dataset						
Average Precision	0.68±0.06	0.65±0.07	0.69±0.07	0.65±0.07	0.69±0.07	0.70±0.07
NDCG@10	0.93±0.17	0.89±0.17	0.91±0.17	0.72±0.26	0.93±0.10	0.91±0.18
NDCG@20	0.89±0.13	0.93±0.12	0.90±0.12	0.82±0.18	0.95±0.08	0.93±0.12
NDCG@50	0.92±0.08	0.89±0.09	0.91±0.08	0.88±0.10	0.90±0.06	0.93±0.07
NDCG@100	0.85±0.07	0.83±0.07	0.87±0.06	0.84±0.08	0.87±0.05	0.89±0.07
NDCG@200	0.78±0.05	0.75±0.05	0.81±0.06	0.76±0.06	0.78±0.04	0.80±0.05
Precision@10	0.90±0.23	0.80±0.21	0.80±0.28	0.60±0.33	0.90±0.20	0.90±0.18
Precision@20	0.85±0.17	0.90±0.13	0.85±0.16	0.80±0.18	0.95±0.12	0.90±0.12
Precision@50	0.90±0.09	0.82±0.11	0.88±0.09	0.86±0.09	0.86±0.10	0.88±0.08
Precision@100	0.76±0.09	0.74±0.09	0.81±0.08	0.76±0.10	0.78±0.08	0.82±0.09
Precision@200	0.62±0.07	0.57±0.08	0.66±0.08	0.61±0.08	0.62±0.08	0.63±0.08
Recall@10	0.04±0.01	0.04±0.01	0.04±0.01	0.03±0.02	0.04±0.01	0.04±0.01
Recall@20	0.08±0.02	0.09±0.02	0.08±0.02	0.08±0.02	0.09±0.01	0.09±0.02
Recall@50	0.22±0.03	0.20±0.03	0.21±0.03	0.21±0.03	0.21±0.03	0.21±0.03
Recall@100	0.37±0.04	0.36±0.05	0.39±0.04	0.37±0.04	0.38±0.04	0.40±0.05
Recall@200	0.59±0.05	0.55±0.06	0.64±0.05	0.58±0.05	0.60±0.06	0.60±0.05
Panel B. Full incident reporting system dataset						
Average Precision	0.47±0.07	0.56±0.06	0.60±0.08	0.57±0.08	0.55±0.07	0.61±0.07
NDCG@10	0.91±0.13	1.00±0.07	0.86±0.24	0.88±0.21	0.98±0.05	0.93±0.14

NDCG@20	0.87±0.11	0.92±0.10	0.90±0.16	0.88±0.14	0.95±0.07	0.96±0.09
NDCG@50	0.87±0.08	0.93±0.06	0.93±0.09	0.91±0.08	0.92±0.06	0.96±0.06
NDCG@100	0.78±0.08	0.86±0.06	0.90±0.09	0.87±0.07	0.86±0.07	0.92±0.06
NDCG@200	0.64±0.06	0.68±0.05	0.69±0.07	0.67±0.07	0.68±0.07	0.72±0.07
Precision@10	0.80±0.23	1.00±0.11	0.90±0.19	0.80±0.25	0.90±0.16	0.90±0.18
Precision@20	0.80±0.17	0.85±0.17	0.90±0.13	0.80±0.18	0.90±0.12	0.95±0.10
Precision@50	0.80±0.10	0.88±0.09	0.92±0.08	0.86±0.10	0.88±0.10	0.96±0.05
Precision@100	0.66±0.12	0.80±0.08	0.84±0.12	0.81±0.09	0.76±0.10	0.86±0.09
Precision@200	0.49±0.08	0.54±0.07	0.56±0.09	0.55±0.09	0.52±0.08	0.58±0.09
Recall@10	0.04±0.01	0.05±0.01	0.04±0.01	0.04±0.01	0.04±0.01	0.04±0.01
Recall@20	0.08±0.02	0.08±0.02	0.09±0.02	0.08±0.02	0.09±0.02	0.09±0.01
Recall@50	0.19±0.04	0.21±0.03	0.22±0.03	0.21±0.04	0.21±0.03	0.23±0.03
Recall@100	0.32±0.05	0.39±0.05	0.40±0.05	0.39±0.05	0.37±0.04	0.41±0.04
Recall@200	0.47±0.06	0.52±0.06	0.53±0.06	0.53±0.07	0.50±0.06	0.56±0.05

Table 3. Performance of the LGBM-CL model across multiple evaluation metrics with 95% confidence intervals.

Metric	LGBM_TFIDF	LGBM_BAG	LGBM_BIOBERT	LGBM_BERT	LGBM_ROBERTA	LGBM_GATORTRON
Panel A. Committee-adjudicated dataset						
Average Precision	0.62±0.08	0.60±0.07	0.63±0.07	0.59±0.07	0.66±0.06	0.62±0.07
NDCG@10	0.89±0.15	0.98±0.05	0.84±0.24	0.98±0.05	0.97±0.07	1.00±0.02
NDCG@20	0.92±0.10	0.97±0.09	0.83±0.17	0.92±0.08	0.98±0.05	0.99±0.06

NDCG@50	0.90±0.08	0.86±0.08	0.89±0.10	0.88±0.08	0.93±0.05	0.88±0.08
NDCG@100	0.83±0.07	0.84±0.07	0.84±0.08	0.81±0.06	0.83±0.06	0.82±0.07
NDCG@200	0.73±0.06	0.73±0.05	0.76±0.06	0.74±0.05	0.78±0.05	0.75±0.06
Precision@10	0.80±0.29	0.90±0.19	0.80±0.25	0.90±0.21	0.90±0.21	1.00±0.07
Precision@20	0.85±0.18	0.90±0.17	0.80±0.16	0.75±0.19	0.95±0.11	0.95±0.15
Precision@50	0.86±0.11	0.76±0.12	0.88±0.08	0.72±0.13	0.82±0.10	0.76±0.13
Precision@100	0.74±0.09	0.71±0.11	0.75±0.10	0.66±0.09	0.70±0.08	0.70±0.09
Precision@200	0.56±0.08	0.52±0.07	0.58±0.07	0.54±0.07	0.63±0.08	0.58±0.08
Recall@10	0.04±0.01	0.04±0.01	0.04±0.01	0.04±0.01	0.04±0.01	0.05±0.01
Recall@20	0.08±0.02	0.09±0.02	0.08±0.02	0.07±0.02	0.09±0.01	0.09±0.01
Recall@50	0.21±0.03	0.18±0.03	0.21±0.03	0.17±0.03	0.20±0.03	0.18±0.03
Recall@100	0.36±0.04	0.34±0.05	0.36±0.05	0.32±0.04	0.34±0.04	0.34±0.04
Recall@200	0.54±0.05	0.59±0.05	0.56±0.05	0.52±0.05	0.60±0.05	0.56±0.06

Panel B. Full incident reporting system dataset

Average Precision	0.39±0.07	0.44±0.09	0.46±0.07	0.49±0.06	0.45±0.08	0.44±0.05
NDCG@10	0.97±0.12	0.83±0.20	0.93±0.14	0.91±0.13	0.57±0.33	0.61±0.23
NDCG@20	0.92±0.10	0.88±0.14	0.94±0.10	0.90±0.11	0.69±0.22	0.50±0.16
NDCG@50	0.90±0.08	0.90±0.14	0.85±0.09	0.89±0.07	0.77±0.12	0.46±0.12
NDCG@100	0.76±0.08	0.74±0.09	0.85±0.09	0.78±0.06	0.71±0.10	0.47±0.08
NDCG@200	0.57±0.08	0.60±0.07	0.74±0.08	0.64±0.05	0.59±0.08	0.43±0.06
Precision@10	0.90±0.18	0.80±0.27	0.90±0.19	0.80±0.23	0.60±0.32	0.70±0.35
Precision@20	0.90±0.13	0.85±0.17	0.90±0.14	0.80±0.17	0.75±0.18	0.55±0.23

Precision@50	0.84±0.12	0.88±0.11	0.76±0.13	0.82±0.10	0.76±0.13	0.52±0.15
Precision@100	0.64±0.11	0.64±0.11	0.63±0.09	0.67±0.08	0.64±0.12	0.56±0.10
Precision@200	0.41±0.08	0.45±0.09	0.44±0.08	0.50±0.07	0.48±0.09	0.50±0.07
Recall@10	0.04±0.01	0.04±0.01	0.04±0.01	0.04±0.01	0.03±0.02	0.02±0.01
Recall@20	0.09±0.02	0.08±0.02	0.09±0.02	0.08±0.02	0.07±0.02	0.03±0.02
Recall@50	0.20±0.03	0.21±0.03	0.18±0.04	0.20±0.03	0.18±0.03	0.07±0.02
Recall@100	0.31±0.05	0.31±0.05	0.30±0.05	0.32±0.04	0.31±0.05	0.15±0.02
Recall@200	0.40±0.06	0.43±0.07	0.42±0.07	0.48±0.06	0.46±0.07	0.27±0.03

Table 4. Performance of the LGBM-RA model across multiple evaluation metrics with 95% confidence intervals.

Metric	LGBM_TFIDF	LGBM_BAG	LGBM_BIOBERT	LGBM_BERT	LGBM_ROBERTA	LGBM_GATORTRON
Panel A. Committee-adjudicated dataset						
Average Precision	0.64±0.07	0.67±0.06	0.68±0.07	0.65±0.07	0.67±0.06	0.70±0.05
NDCG@10	0.80±0.22	1.00±0.01	0.71±0.36	0.73±0.31	1.00±0.02	0.98±0.05
NDCG@20	0.85±0.17	0.99±0.04	0.78±0.23	0.83±0.21	0.98±0.03	0.98±0.05
NDCG@50	0.87±0.11	0.92±0.06	0.87±0.13	0.83±0.13	0.93±0.06	0.94±0.06
NDCG@100	0.85±0.08	0.88±0.05	0.84±0.09	0.81±0.09	0.85±0.05	0.90±0.05
NDCG@200	0.74±0.05	0.76±0.04	0.76±0.06	0.76±0.06	0.79±0.04	0.79±0.05
Precision@10	0.80±0.22	1.00±0.02	0.80±0.27	0.70±0.28	1.00±0.08	0.90±0.19
Precision@20	0.90±0.15	0.95±0.08	0.85±0.16	0.85±0.18	0.90±0.12	0.95±0.11
Precision@50	0.86±0.10	0.88±0.09	0.90±0.09	0.82±0.11	0.84±0.11	0.88±0.10

Precision@100	0.81±0.10	0.81±0.09	0.82±0.08	0.76±0.09	0.74±0.09	0.81±0.09
Precision@200	0.56±0.08	0.56±0.06	0.61±0.08	0.62±0.08	0.60±0.08	0.63±0.07
Recall@10	0.04±0.01	0.05±0.00	0.04±0.01	0.03±0.01	0.05±0.01	0.04±0.01
Recall@20	0.09±0.02	0.09±0.01	0.08±0.02	0.08±0.02	0.09±0.02	0.09±0.01
Recall@50	0.21±0.03	0.21±0.03	0.22±0.03	0.20±0.03	0.20±0.03	0.21±0.03
Recall@100	0.38±0.04	0.39±0.05	0.40±0.04	0.37±0.04	0.36±0.04	0.39±0.05
Recall@200	0.53±0.05	0.54±0.05	0.59±0.05	0.60±0.06	0.58±0.06	0.61±0.05

Panel B. Full incident reporting system dataset

Average Precision	0.42±0.07	0.49±0.07	0.55±0.07	0.53±0.08	0.54±0.07	0.60±0.07
NDCG@10	1.00±0.00	1.00±0.00	1.00±0.00	0.93±0.07	1.00±0.07	0.98±0.03
NDCG@20	1.00±0.00	1.00±0.00	1.00±0.01	0.95±0.05	0.95±0.07	0.98±0.03
NDCG@50	0.95±0.07	0.92±0.06	0.88±0.08	0.96±0.06	0.93±0.05	0.95±0.06
NDCG@100	0.76±0.10	0.79±0.08	0.82±0.06	0.84±0.07	0.84±0.08	0.86±0.05
NDCG@200	0.52±0.07	0.60±0.06	0.66±0.05	0.65±0.08	0.64±0.06	0.74±0.06
Precision@10	1.00±0.00	1.00±0.00	1.00±0.00	0.70±0.25	1.00±0.10	1.00±0.14
Precision@20	1.00±0.00	1.00±0.00	1.00±0.01	0.85±0.15	0.90±0.11	0.95±0.10
Precision@50	0.92±0.13	0.88±0.09	0.80±0.12	0.92±0.10	0.88±0.08	0.90±0.11
Precision@100	0.63±0.14	0.68±0.11	0.71±0.09	0.76±0.10	0.74±0.12	0.72±0.08
Precision@200	0.38±0.07	0.45±0.07	0.51±0.07	0.50±0.10	0.48±0.07	0.60±0.08
Recall@10	0.05±0.01	0.05±0.01	0.05±0.01	0.03±0.01	0.05±0.01	0.05±0.01
Recall@20	0.10±0.01	0.10±0.01	0.10±0.01	0.08±0.02	0.09±0.01	0.09±0.01
Recall@50	0.22±0.03	0.21±0.03	0.19±0.03	0.22±0.03	0.21±0.03	0.22±0.03

Recall@100	0.30±0.06	0.33±0.05	0.34±0.05	0.37±0.04	0.36±0.05	0.34±0.05
Recall@200	0.36±0.06	0.43±0.06	0.49±0.05	0.48±0.07	0.46±0.06	0.57±0.06

Table 5. Performance of transformer-based classification models (BERT-CL, RoBERTa-CL, XLM-RoBERTa-CL, BIOBERT-CL, BIOMEDBERT-CL, and GATORTRON-CL) across multiple evaluation metrics with 95% confidence intervals.

Metric	BERT	ROBERTA	XLM-ROBERTA	BIOBERT	PUBMEDBERT	GATORTRON
Panel A. Committee-adjudicated dataset						
Average Precision	0.72±0.06	0.73±0.06	0.69±0.06	0.74±0.06	0.71±0.06	0.70±0.07
NDCG@10	1.00±0.04	1.00±0.00	1.00±0.01	1.00±0.00	0.96±0.07	0.91±0.20
NDCG@20	0.97±0.07	1.00±0.00	0.99±0.02	1.00±0.00	0.98±0.05	0.93±0.13
NDCG@50	0.94±0.06	0.96±0.05	0.95±0.05	0.96±0.04	0.94±0.05	0.93±0.08
NDCG@100	0.91±0.05	0.91±0.05	0.88±0.06	0.91±0.05	0.91±0.05	0.89±0.06
NDCG@200	0.80±0.04	0.80±0.04	0.81±0.05	0.81±0.04	0.80±0.05	0.80±0.05
Precision@10	1.00±0.06	1.00±0.00	1.00±0.06	1.00±0.00	0.90±0.19	0.90±0.20
Precision@20	0.95±0.10	1.00±0.03	0.95±0.10	1.00±0.00	0.95±0.10	0.90±0.13
Precision@50	0.88±0.09	0.92±0.10	0.86±0.10	0.92±0.07	0.90±0.08	0.90±0.10
Precision@100	0.84±0.08	0.83±0.07	0.80±0.09	0.84±0.08	0.83±0.09	0.82±0.08
Precision@200	0.65±0.07	0.64±0.07	0.65±0.08	0.66±0.07	0.66±0.07	0.65±0.09
Recall@10	0.05±0.01	0.05±0.01	0.05±0.01	0.05±0.01	0.04±0.01	0.04±0.01
Recall@20	0.09±0.01	0.10±0.01	0.09±0.01	0.10±0.01	0.09±0.01	0.09±0.02
Recall@50	0.21±0.03	0.22±0.03	0.21±0.03	0.22±0.03	0.22±0.03	0.22±0.03
Recall@100	0.41±0.04	0.40±0.04	0.39±0.04	0.41±0.04	0.40±0.04	0.40±0.04

Recall@200	0.62±0.05	0.61±0.05	0.62±0.05	0.63±0.05	0.63±0.05	0.62±0.05
Panel B. Full incident reporting system dataset						
Average Precision	0.61±0.06	0.65±0.07	0.52±0.07	0.68±0.06	0.63±0.07	0.62±0.06
NDCG@10	0.89±0.14	0.88±0.13	0.94±0.14	0.93±0.13	0.90±0.16	0.92±0.12
NDCG@20	0.91±0.11	0.89±0.10	0.79±0.17	0.95±0.09	0.90±0.13	0.90±0.15
NDCG@50	0.86±0.10	0.89±0.07	0.76±0.10	0.93±0.05	0.84±0.09	0.86±0.08
NDCG@100	0.85±0.06	0.89±0.06	0.76±0.08	0.90±0.04	0.83±0.07	0.88±0.06
NDCG@200	0.75±0.04	0.76±0.05	0.68±0.06	0.79±0.06	0.77±0.05	0.75±0.05
Precision@10	0.80±0.18	0.80±0.28	0.90±0.20	0.90±0.18	0.80±0.28	0.90±0.14
Precision@20	0.75±0.18	0.85±0.16	0.70±0.24	0.95±0.10	0.75±0.23	0.80±0.18
Precision@50	0.74±0.12	0.86±0.10	0.58±0.14	0.89±0.06	0.72±0.13	0.78±0.08
Precision@100	0.71±0.08	0.84±0.09	0.59±0.10	0.84±0.07	0.73±0.09	0.76±0.08
Precision@200	0.60±0.05	0.59±0.08	0.49±0.07	0.61±0.09	0.63±0.07	0.61±0.05
Recall@10	0.04±0.01	0.04±0.01	0.04±0.01	0.04±0.01	0.04±0.01	0.04±0.01
Recall@20	0.07±0.02	0.08±0.02	0.07±0.02	0.09±0.01	0.07±0.02	0.08±0.02
Recall@50	0.17±0.04	0.21±0.03	0.14±0.03	0.18±0.03	0.17±0.03	0.21±0.03
Recall@100	0.33±0.04	0.41±0.04	0.29±0.05	0.37±0.05	0.35±0.05	0.40±0.04
Recall@200	0.60±0.05	0.57±0.05	0.47±0.05	0.61±0.06	0.61±0.05	0.60±0.05

Table 6. Performance of transformer-based ranking models (BERT-RA, RoBERTa-RA, XLM-RoBERTa-RA, BIOBERT-RA, BIOMEDBERT-RA, and GATORTRON-RA) across multiple evaluation metrics with 95% confidence intervals.

Metric	BERT	ROBERTA	XLM-ROBERTA	BIOBERT	PUBMEDBERT	GATORTRON
--------	------	---------	-------------	---------	------------	-----------

Panel A. Committee-adjudicated dataset

Average Precision	0.91±0.03	0.79±0.05	0.67±0.06	0.88±0.05	0.89±0.04	0.81±0.05
NDCG@10	1.00±0.00	1.00±0.01	1.00±0.01	0.95±0.09	1.00±0.00	1.00±0.00
NDCG@20	1.00±0.00	0.99±0.02	0.99±0.03	0.97±0.06	0.99±0.03	1.00±0.01
NDCG@50	0.99±0.02	0.97±0.03	0.92±0.06	0.98±0.03	0.98±0.02	0.98±0.04
NDCG@100	0.99±0.04	0.94±0.04	0.88±0.06	0.98±0.02	0.97±0.02	0.93±0.04
NDCG@200	0.93±0.03	0.87±0.03	0.81±0.04	0.91±0.02	0.94±0.03	0.88±0.04
Precision@10	1.00±0.00	1.00±0.06	1.00±0.04	0.90±0.20	1.00±0.02	1.00±0.00
Precision@20	1.00±0.00	0.95±0.09	0.95±0.11	0.95±0.10	0.95±0.11	1.00±0.01
Precision@50	0.98±0.05	0.92±0.07	0.82±0.11	0.96±0.06	0.94±0.07	0.98±0.05
Precision@100	0.97±0.04	0.84±0.07	0.77±0.09	0.95±0.05	0.95±0.05	0.87±0.07
Precision@200	0.83±0.07	0.74±0.06	0.63±0.07	0.82±0.06	0.85±0.07	0.76±0.08
Recall@10	0.05±0.01	0.05±0.01	0.05±0.01	0.04±0.01	0.05±0.00	0.05±0.01
Recall@20	0.10±0.01	0.09±0.01	0.09±0.01	0.09±0.01	0.09±0.01	0.10±0.01
Recall@50	0.24±0.03	0.22±0.03	0.19±0.03	0.23±0.03	0.23±0.03	0.24±0.03
Recall@100	0.47±0.04	0.41±0.04	0.37±0.04	0.46±0.05	0.46±0.05	0.42±0.04
Recall@200	0.80±0.05	0.72±0.05	0.60±0.06	0.81±0.06	0.81±0.05	0.73±0.05

Panel B. Full incident reporting system dataset

Average Precision	0.87±0.04	0.71±0.06	0.60±0.07	0.74±0.05	0.85±0.04	0.73±0.06
NDCG@10	1.00±0.00	0.97±0.06	0.93±0.14	1.00±0.00	1.00±0.00	1.00±0.03
NDCG@20	1.00±0.00	0.98±0.04	0.94±0.10	1.00±0.01	1.00±0.01	0.99±0.03
NDCG@50	0.99±0.01	0.99±0.02	0.94±0.06	0.99±0.02	0.99±0.01	0.97±0.04

NDCG@100	0.96±0.02	0.97±0.04	0.90±0.05	0.98±0.03	0.98±0.03	0.94±0.05
NDCG@200	0.94±0.02	0.81±0.05	0.71±0.06	0.83±0.06	0.92±0.03	0.83±0.05
Precision@10	1.00±0.00	0.90±0.19	0.90±0.19	1.00±0.00	1.00±0.00	1.00±0.14
Precision@20	1.00±0.00	0.95±0.10	0.90±0.14	1.00±0.06	1.00±0.04	0.95±0.11
Precision@50	0.97±0.01	0.98±0.04	0.90±0.09	0.96±0.06	0.96±0.05	0.92±0.08
Precision@100	0.90±0.06	0.92±0.07	0.81±0.09	0.93±0.07	0.95±0.06	0.90±0.07
Precision@200	0.78±0.07	0.65±0.08	0.55±0.08	0.67±0.10	0.78±0.08	0.67±0.08
Recall@10	0.05±0.01	0.04±0.01	0.04±0.01	0.05±0.01	0.05±0.01	0.05±0.01
Recall@20	0.10±0.01	0.09±0.02	0.09±0.02	0.10±0.01	0.10±0.01	0.09±0.02
Recall@50	0.23±0.03	0.24±0.03	0.22±0.03	0.23±0.03	0.23±0.03	0.22±0.03
Recall@100	0.43±0.05	0.44±0.05	0.39±0.04	0.45±0.05	0.46±0.05	0.43±0.05
Recall@200	0.75±0.06	0.62±0.06	0.52±0.06	0.64±0.06	0.74±0.05	0.64±0.06

Appendix 4.6 Sensitivity Analysis of Multiclass and Binary HSE Classification Formulations.

To examine whether the multiclass outcome formulation influenced the comparison between classification and ranking, we conducted a targeted sensitivity analysis using LLAMA3.1, the best-performing classification architecture. In the primary multiclass formulation, LLAMA3.1-CL was trained to predict the original HSE, LSE, NME, and NSE categories, and reports were ordered according to their predicted probabilities of receiving the HSE label. For the binary formulation, the original labels were collapsed into HSE and non-HSE categories, with LSE, NME, and NSE reports assigned to the non-HSE category. The resulting LLAMA3.1-CL-Binary model was trained to distinguish HSE from non-HSE reports directly, and reports were ordered according to their predicted probabilities of belonging to the HSE class.

The original training, validation, and testing partitions were retained. The binary model underwent the same hyperparameter-selection procedure as the multiclass model, using the unchanged validation set and the same candidate ranges for learning rate, batch size, number of training epochs, warmup ratio, and weight decay. Class imbalance was addressed using class-weighted cross-entropy loss, with weights calculated from the HSE and non-HSE frequencies in the binary training partition. The final hyperparameter configuration was selected based on validation Average Precision. No oversampling was applied to the transformer-based binary model, consistent with the imbalance-mitigation procedure used for the primary multiclass transformer-based classification models.

As shown in Table 1, all three formulations achieved identical performance at the most restrictive cutoffs, with NDCG and precision of 1.00 at $k = 10$ and $k = 20$. Differences emerged as the evaluation range expanded. LLAMA3.1-CL-Binary achieved an Average Precision of 0.79, compared with 0.77 for LLAMA3.1-CL. At $k = 200$, the binary formulation increased precision from 0.71 to 0.77 and recall from 0.69 to 0.74, indicating that directly optimizing the distinction between HSE and non-HSE reports provided a modest improvement in HSE-focused retrieval relative to multiclass classification. Despite its improved HSE-focused retrieval, LLAMA3.1-CL-Binary achieved lower NDCG than LLAMA3.1-CL at the broader cutoffs. At $k = 200$, NDCG was 0.82 for LLAMA3.1-CL-Binary and 0.85 for LLAMA3.1-CL. This pattern is consistent with the

difference between the two outcome formulations. The binary model was trained to distinguish HSE from non-HSE reports but did not explicitly learn distinctions among LSE, NME, and NSE reports. In contrast, the multiclass model retained information about the complete severity structure during training, resulting in stronger graded ordering despite retrieving fewer HSE reports.

LLAMA3.1-RA continued to achieve the highest performance overall, with an Average Precision of 0.94 and NDCG, precision, and recall of 0.96, 0.86, and 0.83, respectively, at $k = 200$. LLAMA3.1-CL-Binary achieved an Average Precision of 0.79 and NDCG@200, Precision@200, and Recall@200 values of 0.82, 0.77, and 0.74, respectively. These findings demonstrate that the performance advantage of LLAMA3.1-RA was not attributable solely to the four-category outcome formulation used for the primary classification comparator. Although binary HSE classification improved HSE-focused retrieval relative to multiclass classification, LLAMA3.1-RA achieved higher performance than both classification formulations on Average Precision, Precision@ k , Recall@ k , and NDCG@ k . These findings provide additional support for the ranking formulation as the strongest evaluated approach for both retrieving HSE reports and maintaining severity-aligned prioritization across the broader ranked list.

Table 1. Comparison of multiclass classification, binary HSE classification, and learning-to-rank formulations using LLAMA3.1 in the committee-adjudicated dataset.

Metric	LLAMA3.1-CL	LLAMA3.1-CL-Binary	LLAMA3.1-RA
Average Precision	0.77	0.79	0.94
NDCG@10	1.00	1.00	1.00
NDCG@20	1.00	1.00	1.00
NDCG@50	0.96	0.95	0.99

NDCG@100	0.94	0.90	0.98
NDCG@200	0.85	0.82	0.96
Precision@10	1.00	1.00	1.00
Precision@20	1.00	1.00	1.00
Precision@50	0.92	0.94	0.98
Precision@100	0.88	0.92	0.96
Precision@200	0.71	0.77	0.86
Recall@10	0.05	0.05	0.05
Recall@20	0.10	0.10	0.10
Recall@50	0.22	0.23	0.24
Recall@100	0.43	0.44	0.47
Recall@200	0.69	0.74	0.83

Appendix 5.1 Dataset Description and Stigmatizing Language Examples.

The dataset utilized in this study was originally curated by Harrigian et al. [25]. Harrigian et al. originally categorized 5,043 sentences from the discharge summaries in the MIMIC-IV database and framed stigmatizing language detection as three independent classification tasks, each targeting different types of stigmatizing language. These tasks included: ‘credibility and obstinacy’ (with labels Disbelief, Difficult, Exclude), ‘compliance’ (with labels Negative, Neutral, Positive), and ‘descriptors’ (with labels Negative, Neutral, Positive, Exclude). Table 1 below presents the original classification taxonomy and associated examples as developed by Harrigian et al. [25]. The taxonomy is developed by the study team of Harrigian et al, drawing upon previous literature [23,36].

In our study, these dimensions were collapsed into a binary classification by treating the labels “Disbelief”, by treating the labels “Disbelief”, “Difficult”, and “Negative” as stigmatizing, and the remaining labels as non-stigmatizing. From a conceptual perspective, we recognize that different dimensions of stigma represent important nuances, however, there is currently no universally accepted taxonomy for categorizing stigmatizing language [25,36,154]. To ensure the broad applicability of our study, we opted for a conventional binary classification framework consistent with prior work [24,36,42]. This alignment also enhances the comparability and interpretability of

results across studies. In addition, a binary formulation enables more consistent and efficient model training, allowing for robust comparisons across a broad range of models across zero-shot, few-shot, and supervised learning.

Table 1. Classification taxonomy of stigmatizing language and associated examples, originally developed by Harrigian et al. [25].

Stigma Type	Class	Definition	Examples
Credibility & Obstinacy	Disbelief	Insinuates doubt about a patient's stated testimony.	adamant he doesn't smoke; claims to see a therapist
	Difficult	Describes patient (or patient's family) perspective as inflexible/difficult/entrenched, typically with respect to their intentions.	insists on being admitted; adamantly opposed to limiting fruit intake
	Exclude	Word/phrase is not used to characterize the patient or describe the patient's behavior; may refer to medical condition or treatment or to another person or context.	patient's friend insisted she go to the hospital; test claims submitted to insurance
Compliance	Negative	Patient not, unlikely to, or questionably following medical advice.	adherence to therapeutic medication is unclear; mother declines vaccines; struggles with medication and follow-up compliance
	Neutral	Not used to describe whether the patient is not following medical advice or rejecting treatment; often used to describe generically some future plan involving a hypothetical. Alternatively, see Exclude (Credibility & Obstinacy).	discussed the medication compliance; school refuses to provide adequate accommodations; feels that her parents' health has declined
	Positive	Patient following medical advice.	continues to be compliant with aspirin regimen; reports excellent adherence
Descriptors	Negative	Patient's demeanor or behavior is cast in a negative light; insinuates the patient is not being forthright or transparent; patient may be falsifying symptoms to get something they want.	drug-seeking behavior; concern for secondary gain; unwilling to meet with case manager; unfortunately a poor historian

	Neutral	Negation of negative descriptors; insinuates the patient was expected to have a negative demeanor or be difficult to interact with.	his mother is the primary historian; interactive and cooperative; not combative or belligerent; dad seems angry with patient at times
	Positive	Patient's demeanor or behavior is described in a positive light; patient is easy to interact with.	lovely 80 year old woman; well-groomed and holds good eye contact; pleasant and appropriate interaction with staff
	Exclude	Patient self-description or description of another individual. Alternatively, see Exclude (Credibility & Obstinacy).	does not want providers to think she's malingering; reports feeling angry before her period; lives on pleasant avenue downtown

Appendix 5.2 Detailed Code Implementation for the Prompting Strategies.

Table 1. The code implementation for the prompting strategies.

Prompting Strategy	Code for Prompting Template
Generic	<pre> Template= [{"role": "system", "content": "You are an expert in detecting stigmatizing language."}, {"role": "user", "content": f"" Input: "{statement}". Choose your answer: Based on the above sentence, does the text convey stigmatization? Yes/No ""}] </pre>
COT	<pre> Template=[{"role": "system", "content": "You are an expert in detecting stigmatizing language."}, {"role": "user", "content": f"" Input: '{statement}' Determine if the input contains stigmatizing language. Let's think step by step. Reasoning: '{reasoning}.' Therefore, the answer (yes or no) is: ""}] </pre>

CARP	Template=[{ "role": "system", "content": "You are an expert in detecting stigmatizing language." }, { "role": "user", "content": f"" First, list clues (e.g., keywords, phrases, contextual information, semantic relations, tones, references) that support the determination of stigmatization in the input. Second, deduce the diagnostic reasoning process from the premises (i.e., clues and input) that support the stigmatization determination. Third, based on the clues, reasoning, and input, determine whether the input conveys stigmatization. Input: ‘{statement}’ Answer with 'yes' or 'no'. Answer: "" "" "" "" ""]
Stigma Detection Guided Prompt	Template=[{ "role": "system", "content": "You are an expert in detecting stigmatizing language." }, { "role": "user", "content": f"" Stigmatizing language can exhibit the following characteristics: 1) Questioning credibility: Implication of physician disbelief in patient reports of their own experiences or behaviors. 2) Disapproval: Highlights poor reasoning, decision-making, or self-care, often in a way that suggests the patient is unreasonable. 3) Stereotyping: Quoting incorrect grammar or unsophisticated terms. 4) Difficult patient: Inclusion of details with questionable clinical significance that portray the patient as belligerent or otherwise suggest that the physician is annoyed.

	5) Unilateral decisions: Language that emphasizes physician authority. Input: ‘{statement}’ Keyword: ‘{keyword}.’ Does the keyword in the input convey stigmatization? Answer with 'Yes' or 'No'. Answer:" """} }
--	---

Appendix 5.3 Paired t-test Comparison of Top-performing Models in Zero-shot, Few-shot, and Supervised Fine-tuning Approaches.

The paired t-test results for comparing top-performing models in zero-shot and few-shot settings, as well as the comparison between the best-performing model in the ICL approach and the best-performing supervised model, are presented in Tables 1, 2, and 3, respectively.

Table 1. Paired t-test results for comparing top-performing models in textual entailment and zero-shot ICL.

Comparison	Metric	t-statistic	Sample size	Mean difference	P-value
Deberta-M vs. GEMMA-2 coupled with Stigma Detection Guided Prompt	F1	-74.186	10	-0.136	<.001
	Precision	-50.286	10	-0.125	<.001
	Recall	-70.761	10	-0.146	<.001

Table 2. Paired t-test results for comparing top-performing models in SetFit and few-shot ICL.

Number of annotations per label	Comparison	Metric	t-statistic	Sample size	Mean difference	P-value
4	MINILM-V2 vs. LLAMA 3 coupled with Stigma Detection Guided Prompt	F1	-40.651	10	-0.155	<.001
		Precision	-40.338	10	-0.138	<.001
		Recall	-61.25	10	-0.165	<.001
8	E5-V2 vs. LLAMA 3 coupled with Stigma Detection Guided Prompt	F1	-60.140	10	-0.158	<.001
		Precision	-45.95	10	-0.124	<.001
		Recall	-59.267	10	-0.192	<.001
16	E5-V2 vs. LLAMA 3 coupled with Stigma Detection Guided Prompt	F1	-44.614	10	-0.101	<.001
		Precision	-17.096	10	-0.052	<.001
		Recall	-80.413	10	-0.152	<.001

Table 3. Paired t-test results for comparing top-performing models in supervised fine-tuning and few-shot ICL.

Comparison	Metric	t-statistic	Sample size	Mean difference	P-value
ROBERTA vs. LAMMA-3 coupled with Stigma Detection	F1	11.03	10	0.029	<.001
	Precision	26.325	10	0.086	<.001

Guided Prompt (16 annotated data per class)	Recall	-9.695	10	-0.030	<.001
--	--------	--------	----	--------	-------

Appendix 5.4 Performance Metrics of the Few-shot In-Context Learning Models.

Table 1. Performance metrics of the few-shot ICL models using Generic prompting strategy.

Prompting Strategy	4 annotated data per class				
Generic	Model	Accuracy	F1	Precision	Recall
	LLAMA-3	0.748±0.008	0.789±0.006	0.855±0.007	0.733±0.008
	FLAN-T5	0.693±0.005	0.691±0.007	0.712±0.005	0.672±0.008
	GEMMA-2	0.711±0.012	0.719±0.007	0.731±0.008	0.708±0.006
	MISTRAL-0.2	0.725±0.007	0.730±0.006	0.752±0.005	0.710±0.004
	BIO-LLAMA-3	0.766±0.009	0.820±0.008	0.832±0.011	0.809±0.008

	8 annotated data per class				
	Model	Accuracy	F1	Precision	Recall
	LLAMA-3	0.792±0.009	0.832±0.008	0.865±0.010	0.802±0.012
	FLAN-T5	0.707±0.008	0.701±0.008	0.712±0.009	0.691±0.009
	GEMMA-2	0.721±0.008	0.739±0.009	0.757±0.011	0.723±0.007
	MISTRAL-0.2	0.705±0.009	0.720±0.010	0.731±0.013	0.709±0.012
	BIO-LLAMA-3	0.773±0.005	0.826±0.004	0.863±0.010	0.792±0.007
	16 annotated data per class				
	Model	Accuracy	F1	Precision	Recall
	LLAMA-3	0.816±0.008	0.855±0.006	0.869±0.009	0.842±0.012
	FLAN-T5	0.751±0.005	0.762±0.006	0.775±0.004	0.749±0.007
	GEMMA-2	0.762±0.009	0.773±0.008	0.765±0.005	0.782±0.008
	MISTRAL-0.2	0.801±0.010	0.822±0.007	0.832±0.008	0.813±0.004
	BIO-LLAMA-3	0.790±0.006	0.851±0.006	0.901±0.011	0.807±0.004

Table 2. Performance metrics of the few-shot ICL models using COT prompting strategy.

Prompting Strategy	4 annotated data per class				
COT	Model	Accuracy	F1	Precision	Recall
	LLAMA-3	0.725±0.008	0.756±0.004	0.809±0.007	0.709±0.004
	FLAN-T5	-	-	-	-
	GEMMA-2	0.751±0.004	0.796±0.006	0.810±0.005	0.783±0.005
	MISTRAL-0.2	0.713±0.003	0.742±0.005	0.753±0.006	0.730±0.006
	BIO-LLAMA-3	0.716±0.007	0.743±0.005	0.782±0.005	0.707±0.007
	8 annotated data per class				
	Model	Accuracy	F1	Precision	Recall
	LLAMA-3	0.831±0.009	0.823±0.007	0.827±0.010	0.819±0.012
	FLAN-T5	-	-	-	-
	GEMMA-2	0.795±0.008	0.826±0.005	0.843±0.007	0.811±0.005
	MISTRAL-0.2	0.751±0.006	0.767±0.007	0.773±0.004	0.762±.008
	BIO-LLAMA-3	0.814±0.008	0.835±0.006	0.840±0.008	0.831±0.005
	16 annotated data per class				
	Model	Accuracy	F1	Precision	Recall

	LLAMA-3	-	-	-	-
	FLAN-T5	-	-	-	-
	GEMMA-2	-	-	-	-
	MISTRAL-0.2	0.769±0.008	0.781±0.006	0.794±0.009	0.771±0.004
	BIO-LLAMA-3	-	-	-	-

Table 3. Performance metrics of the few-shot ICL models using CARP prompting strategy.

Prompting Strategy	4 annotated data per class				
CARP	Model	Accuracy	F1	Precision	Recall
	LLAMA-3	0.804±0.008	0.837±0.005	0.853±0.009	0.822±0.004
	FLAN-T5	-	-	-	-
	GEMMA-2	0.754±0.007	0.795±0.006	0.804±0.005	0.786±0.006
	MISTRAL-0.2	0.718±0.005	0.728±0.008	0.734±0.006	0.722±0.005
	BIO-LLAMA-3	0.772±0.007	0.813±0.009	0.842±0.007	0.786±0.008
	8 annotated data per class				
	Model	Accuracy	F1	Precision	Recall

	LLAMA-3	0.851±0.005	0.876±0.004	0.884±0.006	0.869±0.003
	FLAN-T5	-	-	-	-
	GEMMA-2	0.796±0.004	0.825±0.009	0.834±0.008	0.815±0.006
	MISTRAL-0.2	0.722±0.005	0.747±0.006	0.757±0.005	0.738±0.007
	BIO-LLAMA-3	0.816±0.006	0.849±0.008	0.867±0.005	0.832±0.006
	16 annotated data per class				
	Model	Accuracy	F1	Precision	Recall
	LLAMA-3	-	-	-	-
	FLAN-T5	-	-	-	-
	GEMMA-2	-	-	-	-
	MISTRAL-0.2	0.712±0.006	0.731±0.005	0.743±0.004	0.720±0.005
	BIO-LLAMA-3	-	-	-	-

Table 4. Performance metrics of the few-shot ICL models using Stigma Detection Guided Prompt prompting strategy.

Prompting Strategy	4 annotated data per class
---------------------------	-----------------------------------

Stigma Detection Guided Prompt	Model	Accuracy	F1	Precision	Recall
	LLAMA-3	0.843±0.006	0.874±0.010	0.854±0.009	0.889±0.005
	FLAN-T5	0.742±0.004	0.782±0.006	0.771±0.005	0.794±0.005
	GEMMA-2	0.843±0.009	0.872±0.007	0.862±0.007	0.883±0.005
	MISTRAL-0.2	0.805±0.007	0.849±0.005	0.835±0.006	0.864±0.007
	BIO-LLAMA-3	0.804±0.011	0.861±0.010	0.839±0.007	0.884±0.005
	8 annotated data per class				
	Model	Accuracy	F1	Precision	Recall
	LLAMA-3	0.855±0.007	0.892±0.006	0.873±0.007	0.912±0.008
	FLAN-T5	0.768±0.005	0.805±0.008	0.782±0.007	0.829±0.005
	GEMMA-2	0.849±0.004	0.881±0.005	0.870±0.004	0.891±0.003
	MISTRAL-0.2	0.820±0.006	0.857±0.004	0.853±0.006	0.862±0.005
	BIO-LLAMA-3	0.815±0.008	0.870±0.007	0.848±0.009	0.893±0.005
	16 annotated data per class				
	Model	Accuracy	F1	Precision	Recall
	LLAMA-3	0.877±0.007	0.901±0.006	0.863±0.008	0.943±0.004
	FLAN-T5	0.828±0.003	0.842±0.004	0.833±0.006	0.851±0.005

	GEMMA-2	0.861±0.005	0.887±0.006	0.876±0.006	0.898±0.007
	MISTRAL-0.2	0.867±0.007	0.899±0.005	0.871±0.004	0.929±0.004
	BIO-LLAMA-3	0.827±0.010	0.892±0.009	0.865±0.007	0.921±0.005

Appendix 5.5 Performance and Performance Disparities of Models across Sex, Age, and Race Demographic Attributes.

Table 1. Performance and performance disparities for models utilizing supervised fine-tuning approach.

Model	Attributes	TPR	TPR Disparity	FPR	FPR Disparity	F1	F1 Disparity
ROBERTA	Male	0.951±0.00 7	0.049	0.142±0.00 2	-0.018	0.946±0.00 6	0.043
	Female	0.902±0.00 5	-0.049	0.160±0.00 4	0.018	0.903±0.00 8	-0.043
	0-25	0.835±0.00 9	-0.108	0.162±0.00 3	0.022	0.862±0.00 5	-0.066
	26-50	0.947±0.00 6	0.005	0.142±0.00 3	0.001	0.922±0.00 4	-0.005
	51-75	0.955±0.00 8	0.013	0.138±0.00 2	-0.003	0.957±0.00 7	0.030

	75 plus	0.938±0.00 3	-0.004	0.139±0.00 4	-0.002	0.933±0.00 9	0.006
	White	0.922±0.00 9	0.029	0.140±0.00 1	-0.008	0.937±0.00 3	0.017
	Black	0.891±0.00 7	-0.002	0.148±0.00 3	0.000	0.920±0.00 3	0.000
	Hispanic	0.893±0.00 4	0.000	0.156±0.00 3	0.008	0.932±0.00 8	0.012
	Asian	0.849±0.00 6	-0.044	0.176±0.00 2	0.028	0.856±0.00 3	-0.064
	Other/Uknow n	0.902±0.00 3	0.009	0.143±0.00 4	-0.005	0.912±0.00 9	-0.008
BERT	Male	0.925±0.00 9	0.051	0.134±0.00 2	-0.016	0.930±0.00 5	0.016
	Female	0.874±0.00 8	-0.051	0.150±0.00 3	0.016	0.914±0.00 6	-0.016
	0-25	0.807±0.00 7	-0.058	0.160±0.00 1	0.016	0.851±0.00 4	-0.048
	26-50	0.896±0.00 5	0.032	0.142±0.00 3	-0.002	0.927±0.00 8	0.028
	51-75	0.904±0.00 6	0.040	0.138±0.00 4	-0.006	0.940±0.00 3	0.041
	75 plus	0.833±0.00 3	-0.032	0.146±0.00 3	0.002	0.871±0.00 6	-0.028
	White	0.895±0.00 4	0.016	0.150±0.00 5	0.000	0.928±0.00 9	0.012
	Black	0.877±0.00 9	-0.002	0.138±0.00 2	-0.012	0.886±0.00 4	-0.030
	Hispanic	0.879±0.00 8	0.000	0.154±0.00 3	0.004	0.916±0.00 7	0.000

	Asian	0.876±0.00 7	-0.003	0.158±0.00 1	0.008	0.910±0.00 9	-0.006
	Other/Uknown	0.882±0.00 5	0.003	0.128±0.00 2	-0.022	0.920±0.00 3	0.004
BERT-BIO	Male	0.914±0.00 6	0.022	0.153±0.00 3	-0.035	0.882±0.00 8	0.003
	Female	0.892±0.00 3	-0.022	0.188±0.00 3	0.035	0.879±0.00 5	-0.003
	0-25	0.840±0.00 9	-0.075	0.177±0.00 4	0.022	0.840±0.00 4	-0.058
	26-50	0.926±0.00 8	0.011	0.148±0.00 1	-0.008	0.884±0.00 7	-0.014
	51-75	0.930±0.00 7	0.015	0.122±0.00 1	-0.034	0.912±0.00 9	0.014
	75 plus	0.904±0.00 5	-0.011	0.163±0.00 3	0.008	0.919±0.00 3	0.022
	White	0.938±0.00 6	0.037	0.158±0.00 4	0.000	0.910±0.00 3	0.015
	Black	0.910±0.00 3	0.009	0.158±0.00 2	0.000	0.918±0.00 5	0.023
	Hispanic	0.877±0.00 5	-0.024	0.181±0.00 4	0.023	0.864±0.00 6	-0.031
	Asian	0.901±0.00 4	0.000	0.164±0.00 5	0.006	0.885±0.00 8	-0.010
	Other/Uknown	0.898±0.00 9	-0.003	0.131±0.00 1	-0.027	0.895±0.00 3	0.000
ROBERTA-XLM	Male	0.934±0.00 7	0.033	0.164±0.00 2	-0.022	0.907±0.00 5	0.006
	Female	0.901±0.00 5	-0.033	0.186±0.00 2	0.022	0.900±0.00 6	-0.006

	0-25	0.823±0.00 7	-0.090	0.165±0.00 4	0.023	0.811±0.00 5	-0.086
	26-50	0.933±0.00 5	0.021	0.142±0.00 3	0.000	0.919±0.00 6	0.022
	51-75	0.939±0.00 6	0.026	0.142±0.00 4	0.000	0.927±0.00 9	0.029
	75 plus	0.892±0.00 5	-0.021	0.133±0.00 5	-0.009	0.876±0.00 3	-0.022
	White	0.941±0.00 8	0.052	0.161±0.00 2	0.000	0.896±0.00 9	0.017
	Black	0.910±0.00 7	0.021	0.144±0.00 1	-0.017	0.904±0.00 5	0.026
	Hispanic	0.859±0.00 5	-0.030	0.157±0.00 3	-0.004	0.846±0.00 6	-0.032
	Asian	0.889±0.00 6	0.000	0.205±0.00 1	0.044	0.867±0.00 5	-0.011
	Other/Uknow n	0.872±0.00 3	-0.017	0.164±0.00 1	0.003	0.878±0.00 0	0.000
BERT- BIOMED	Male	0.918±0.00 4	0.032	0.186±0.00 3	-0.039	0.913±0.00 7	0.040
	Female	0.886±0.00 9	-0.032	0.225±0.00 4	0.039	0.873±0.00 6	-0.040
	0-25	0.823±0.00 7	-0.053	0.223±0.00 2	0.038	0.815±0.00 5	-0.053
	26-50	0.917±0.00 5	0.042	0.180±0.00 1	-0.005	0.898±0.00 6	0.030
	51-75	0.910±0.00 6	0.035	0.190±0.00 4	0.005	0.884±0.00 7	0.016
	75 plus	0.841±0.00 3	-0.035	0.155±0.00 3	-0.030	0.852±0.00 6	-0.016

	White	0.906±0.004	0.010	0.198±0.003	0.004	0.888±0.009	0.011
	Black	0.924±0.009	0.028	0.182±0.003	-0.012	0.894±0.008	0.016
	Hispanic	0.879±0.008	-0.017	0.206±0.002	0.012	0.877±0.003	0.000
	Asian	0.832±0.007	-0.064	0.194±0.006	0.000	0.788±0.005	-0.089
	Other/Unknown	0.896±0.005	0.000	0.161±0.003	-0.033	0.856±0.005	-0.021

Table 2. Performance and performance disparities for models utilizing zero-shot ICL approach.

Model	Attributes	TPR	TPR Disparity	FPR	FPR Disparity	F1	F1 Disparity
LLAMA-3	Male	0.805±0.004	-0.005	0.192±0.004	0.003	0.803±0.003	-0.006
	Female	0.810±0.006	0.005	0.190±0.002	-0.003	0.809±0.004	0.006
	0-25	0.814±0.003	-0.004	0.184±0.003	0.002	0.802±0.005	-0.010
	26-50	0.816±0.002	-0.002	0.180±0.003	-0.002	0.808±0.006	-0.004
	51-75	0.829±0.006	0.012	0.177±0.005	-0.005	0.827±0.004	0.015
	75 plus	0.820±0.005	0.002	0.190±0.003	0.008	0.817±0.004	0.004
	White	0.822±0.003	0.003	0.186±0.002	-0.004	0.813±0.006	-0.006
	Black	0.819±0.005	0.000	0.192±0.005	0.002	0.820±0.003	0.001
	Hispanic	0.823±0.002	0.003	0.198±0.004	0.008	0.818±0.005	-0.002
	Asian	0.815±0.006	-0.005	0.190±0.003	0.000	0.820±0.004	0.001
	Other/Unknown	0.808±0.002	-0.012	0.186±0.004	-0.004	0.819±0.009	0.000
FLAN-T5	Male	0.748±0.004	-0.002	0.246±0.006	-0.007	0.742±0.005	0.004

	Female	0.750±0.007	0.002	0.254±0.006	0.007	0.738±0.006	-0.004
	0-25	0.750±0.005	0.001	0.253±0.005	0.005	0.725±0.004	-0.018
	26-50	0.747±0.006	-0.002	0.237±0.004	-0.011	0.742±0.006	-0.001
	51-75	0.747±0.005	-0.001	0.242±0.003	-0.005	0.750±0.005	0.007
	75 plus	0.752±0.003	0.004	0.259±0.002	0.011	0.744±0.006	0.001
	White	0.748±0.007	-0.007	0.235±0.003	-0.006	0.738±0.007	-0.007
	Black	0.756±0.007	0.000	0.241±0.005	0.000	0.756±0.003	0.011
	Hispanic	0.758±0.004	0.002	0.257±0.003	0.016	0.745±0.004	0.000
	Asian	0.744±0.005	-0.012	0.232±0.004	-0.010	0.753±0.005	0.008
	Other/Uknown	0.760±0.003	0.004	0.241±0.003	0.000	0.744±0.002	-0.001
GEMMA-2	Male	0.780±0.004	-0.003	0.201±0.002	0.008	0.774±0.004	-0.010
	Female	0.783±0.006	0.003	0.194±0.004	-0.008	0.784±0.003	0.010
	0-25	0.780±0.007	0.000	0.195±0.003	0.001	0.786±0.002	0.002
	26-50	0.788±0.004	0.009	0.196±0.004	0.002	0.772±0.006	-0.012
	51-75	0.775±0.005	-0.004	0.176±0.006	-0.018	0.796±0.007	0.013
	75 plus	0.779±0.006	0.000	0.193±0.005	-0.001	0.781±0.005	-0.002
	White	0.785±0.003	0.010	0.186±0.002	-0.003	0.765±0.008	-0.016
	Black	0.787±0.006	0.012	0.188±0.003	-0.001	0.781±0.005	-0.001
	Hispanic	0.775±0.007	0.000	0.204±0.004	0.015	0.790±0.002	0.008
	Asian	0.775±0.005	0.000	0.209±0.005	0.019	0.785±0.003	0.004
	Other/Uknown	0.771±0.006	-0.005	0.189±0.005	0.000	0.781±0.006	0.000
MISTRAL-0.2	Male	0.741±0.003	-0.007	0.248±0.007	-0.005	0.741±0.007	-0.006
	Female	0.748±0.006	0.007	0.253±0.008	0.005	0.746±0.004	0.006

	0-25	0.731±0.002	-0.012	0.236±0.003	0.005	0.731±0.005	-0.015
	26-50	0.740±0.002	-0.003	0.235±0.004	0.004	0.748±0.005	0.003
	51-75	0.755±0.004	0.012	0.215±0.006	-0.016	0.748±0.006	0.002
	75 plus	0.746±0.004	0.003	0.228±0.005	-0.004	0.744±0.007	-0.002
	White	0.755±0.005	0.011	0.238±0.004	0.000	0.733±0.008	-0.004
	Black	0.744±0.003	0.000	0.252±0.009	0.013	0.738±0.005	0.001
	Hispanic	0.743±0.006	-0.001	0.253±0.004	0.014	0.729±0.006	-0.009
	Asian	0.742±0.005	-0.002	0.233±0.006	-0.005	0.737±0.004	0.000
	Other/Uknown	0.755±0.003	0.010	0.237±0.007	-0.001	0.748±0.003	0.010
BIO-LLAMA-3	Male	0.794±0.004	0.007	0.175±0.005	0.001	0.799±0.006	0.005
	Female	0.787±0.005	-0.007	0.174±0.006	-0.001	0.794±0.005	-0.005
	0-25	0.782±0.006	-0.007	0.184±0.004	0.005	0.789±0.003	-0.005
	26-50	0.799±0.005	0.010	0.175±0.003	-0.005	0.797±0.007	0.004
	51-75	0.796±0.004	0.007	0.187±0.006	0.007	0.790±0.002	-0.004
	75 plus	0.782±0.003	-0.007	0.169±0.003	-0.011	0.797±0.008	0.004
	White	0.799±0.006	0.013	0.170±0.007	-0.001	0.788±0.005	0.007
	Black	0.786±0.005	0.000	0.164±0.002	-0.008	0.781±0.005	0.000
	Hispanic	0.780±0.003	-0.006	0.179±0.005	0.007	0.790±0.004	0.009
	Asian	0.786±0.006	0.000	0.172±0.004	0.000	0.780±0.005	-0.001
	Other/Uknown	0.789±0.006	0.003	0.174±0.005	0.002	0.773±0.006	-0.009

Table 3. Performance and performance disparities for models utilizing few-shot ICL approach (4 annotated data per class).

Model	Attributes	TPR	TPR Disparity	FPR	FPR Disparity	F1	F1 Disparity
LLAMA-3	Male	0.824±0.004	0.003	0.193±0.003	0.006	0.812±0.005	-0.005
	Female	0.821±0.006	-0.003	0.187±0.004	-0.006	0.816±0.006	0.005
	0-25	0.829±0.004	-0.001	0.187±0.003	0.007	0.821±0.007	-0.004
	26-50	0.829±0.007	-0.001	0.173±0.002	-0.007	0.822±0.008	-0.003
	51-75	0.838±0.003	0.008	0.173±0.003	-0.007	0.834±0.005	0.010
	75 plus	0.831±0.002	0.001	0.193±0.004	0.013	0.828±0.006	0.003
	White	0.840±0.005	0.009	0.185±0.005	-0.003	0.829±0.007	0.000
	Black	0.832±0.005	0.001	0.195±0.003	0.007	0.828±0.004	-0.001
	Hispanic	0.831±0.004	0.000	0.189±0.006	0.001	0.831±0.008	0.002
	Asian	0.830±0.006	-0.001	0.184±0.003	-0.004	0.846±0.006	0.017
	Other/Uknown	0.818±0.004	-0.013	0.188±0.003	0.000	0.828±0.005	-0.001
FLAN-T5	Male	0.769±0.003	0.004	0.247±0.004	-0.007	0.740±0.004	-0.012
	Female	0.765±0.005	-0.004	0.254±0.005	0.007	0.752±0.006	0.012
	0-25	0.759±0.007	0.001	0.250±0.004	0.005	0.747±0.007	-0.011
	26-50	0.754±0.005	-0.004	0.230±0.006	-0.015	0.758±0.005	0.000
	51-75	0.757±0.005	-0.001	0.240±0.007	-0.005	0.761±0.004	0.002
	75 plus	0.767±0.004	0.009	0.253±0.004	0.008	0.759±0.007	0.000
	White	0.766±0.003	-0.001	0.235±0.003	0.000	0.752±0.006	-0.009
	Black	0.773±0.002	0.006	0.241±0.006	0.006	0.774±0.007	0.013
	Hispanic	0.772±0.005	0.005	0.246±0.005	0.011	0.762±0.005	0.001
	Asian	0.753±0.006	-0.014	0.229±0.004	-0.006	0.761±0.004	0.000

	Other/Uknown	0.767±0.006	0.000	0.231±0.006	-0.004	0.760±0.006	-0.001
GEMMA-2	Male	0.800±0.005	0.002	0.201±0.004	0.005	0.799±0.003	0.008
	Female	0.798±0.006	-0.002	0.195±0.003	-0.005	0.791±0.006	-0.008
	0-25	0.786±0.005	-0.004	0.196±0.002	0.007	0.795±0.004	0.000
	26-50	0.799±0.002	0.009	0.189±0.002	0.000	0.783±0.005	-0.013
	51-75	0.793±0.005	0.002	0.175±0.005	-0.013	0.809±0.008	0.013
	75 plus	0.788±0.004	-0.002	0.188±0.003	0.000	0.796±0.005	0.000
	White	0.797±0.006	0.007	0.187±0.003	0.000	0.783±0.007	-0.009
	Black	0.802±0.007	0.012	0.187±0.006	0.000	0.795±0.003	0.003
	Hispanic	0.783±0.004	-0.007	0.202±0.003	0.015	0.806±0.005	0.014
	Asian	0.789±0.003	0.000	0.202±0.004	0.015	0.792±0.007	0.000
	Other/Uknown	0.782±0.006	-0.008	0.184±0.003	-0.003	0.789±0.006	-0.004
MISTRAL-0.2	Male	0.753±0.007	-0.014	0.238±0.002	-0.003	0.759±0.007	-0.004
	Female	0.766±0.005	0.014	0.240±0.005	0.003	0.763±0.009	0.004
	0-25	0.738±0.004	-0.017	0.247±0.006	0.016	0.743±0.005	-0.016
	26-50	0.749±0.003	-0.006	0.236±0.005	0.005	0.755±0.006	-0.003
	51-75	0.762±0.004	0.007	0.217±0.004	-0.014	0.765±0.007	0.007
	75 plus	0.762±0.007	0.006	0.226±0.003	-0.005	0.762±0.007	0.003
	White	0.759±0.006	0.0044	0.231±0.004	-0.004	0.758±0.007	0.007
	Black	0.755±0.007	0.000	0.241±0.006	0.006	0.752±0.005	0.000
	Hispanic	0.748±0.008	-0.007	0.249±0.007	0.014	0.744±0.004	-0.007
	Asian	0.748±0.004	-0.007	0.229±0.005	-0.006	0.745±0.003	-0.006
	Other/Uknown	0.768±0.008	0.013	0.235±0.003	0.000	0.765±0.008	0.013

BIO-LLAMA-3	Male	0.807±0.005	0.008	0.173±0.005	-0.005	0.816±0.005	0.010
	Female	0.799±0.006	-0.008	0.178±0.003	0.005	0.806±0.005	-0.010
	0-25	0.795±0.007	-0.005	0.178±0.002	0.003	0.796±0.008	-0.011
	26-50	0.805±0.004	0.005	0.173±0.004	-0.003	0.816±0.009	0.009
	51-75	0.814±0.003	0.014	0.181±0.006	0.005	0.804±0.004	-0.003
	75 plus	0.794±0.002	-0.006	0.168±0.004	-0.017	0.810±0.003	0.003
	White	0.806±0.005	0.010	0.164±0.002	-0.001	0.798±0.004	0.007
	Black	0.793±0.004	-0.004	0.164±0.003	-0.001	0.792±0.005	0.000
	Hispanic	0.789±0.006	-0.007	0.180±0.004	0.015	0.798±0.006	0.007
	Asian	0.797±0.005	0.000	0.166±0.005	0.000	0.789±0.007	-0.002
	Other/Uknown	0.797±0.007	0.001	0.173±0.006	0.007	0.774±0.003	-0.018

Table 4. Performance and performance disparities for models utilizing few-shot ICL approach (8 annotated data per class).

Model	Attributes	TPR	TPR Disparity	FPR	FPR Disparity	F1	F1 Disparity
LLAMA-3	Male	0.837±0.00 5	-0.006	0.195±0.00 2	0.002	0.839±0.00 6	-0.008
	Female	0.843±0.00 4	0.006	0.193±0.00 3	-0.002	0.847±0.00 9	0.008
	0-25	0.850±0.00 8	-0.003	0.198±0.00 2	0.002	0.843±0.00 5	-0.006
	26-50	0.851±0.00 6	-0.002	0.195±0.00 1	-0.002	0.846±0.00 3	-0.003
	51-75	0.861±0.00 3	0.008	0.192±0.00 4	-0.005	0.858±0.00 8	0.009

	75 plus	0.855±0.00 9	0.002	0.203±0.00 3	0.007	0.851±0.00 7	0.002
	White	0.859±0.00 6	0.01	0.211±0.00 5	-0.005	0.847±0.00 4	-0.004
	Black	0.849±0.00 4	0	0.209±0.00 2	0.003	0.851±0.00 9	0
	Hispanic	0.853±0.00 8	0.004	0.201±0.00 3	0.005	0.850±0.00 5	-0.001
	Asian	0.849±0.00 6	0	0.195±0.00 3	-0.001	0.857±0.00 3	0.006
	Other/Unkno wn	0.845±0.00 3	-0.004	0.196±0.00 4	0	0.851±0.00 8	0
FLAN-T5	Male	0.776±0.00 5	0.003	0.248±0.00 1	-0.011	0.769±0.00 6	-0.001
	Female	0.773±0.00 4	-0.003	0.258±0.00 2	0.011	0.771±0.00 9	0.001
	0-25	0.769±0.00 8	-0.005	0.250±0.00 2	0	0.751±0.00 5	-0.023
	26-50	0.774±0.00 6	0	0.223±0.00 1	-0.026	0.775±0.00 3	0.002
	51-75	0.773±0.00 3	-0.001	0.250±0.00 5	0	0.786±0.00 8	0.012
	75 plus	0.779±0.00 9	0.005	0.251±0.00 3	0.011	0.771±0.00 7	-0.002
	White	0.773±0.00 6	-0.008	0.234±0.00 3	-0.007	0.767±0.00 4	-0.006
	Black	0.781±0.00 4	0	0.242±0.00 2	0.001	0.788±0.00 9	0.016
	Hispanic	0.782±0.00 8	0.001	0.260±0.00 3	0.019	0.771±0.00 5	-0.001

	Asian	0.772±0.006	-0.009	0.224±0.001	-0.017	0.783±0.003	0.011
	Other/Unknown	0.783±0.005	0.002	0.240±0.004	0	0.772±0.004	0
GEMMA-2	Male	0.814±0.008	-0.006	0.207±0.002	0.001	0.810±0.005	-0.003
	Female	0.820±0.006	0.006	0.206±0.003	-0.001	0.813±0.004	0.003
	0-25	0.813±0.004	-0.002	0.207±0.004	0.001	0.815±0.009	-0.001
	26-50	0.819±0.009	0.004	0.205±0.003	-0.001	0.805±0.006	-0.011
	51-75	0.807±0.005	-0.007	0.192±0.002	-0.014	0.826±0.003	0.01
	75 plus	0.816±0.003	0.001	0.209±0.001	0.003	0.818±0.008	0.001
	White	0.821±0.007	0.008	0.193±0.002	-0.013	0.801±0.009	-0.015
	Black	0.817±0.005	0.004	0.204±0.003	-0.002	0.817±0.008	0.001
	Hispanic	0.813±0.009	0	0.213±0.004	0.007	0.823±0.006	0.007
	Asian	0.803±0.006	-0.01	0.219±0.005	0.014	0.817±0.007	0
	Other/Unknown	0.805±0.008	-0.008	0.206±0.003	0	0.812±0.009	-0.004
MISTRAL-0.2	Male	0.775±0.005	-0.004	0.254±0.001	-0.004	0.766±0.009	-0.008
	Female	0.779±0.008	0.004	0.258±0.002	0.004	0.775±0.003	0.008

	0-25	0.763±0.00 4	-0.014	0.273±0.00 3	0.032	0.758±0.00 8	-0.017
	26-50	0.774±0.00 7	-0.003	0.245±0.00 4	0.004	0.774±0.00 6	-0.001
	51-75	0.783±0.00 9	0.006	0.219±0.00 3	-0.023	0.776±0.00 5	0.001
	75 plus	0.780±0.00 3	0.003	0.238±0.00 2	-0.004	0.777±0.00 4	0.001
	White	0.783±0.00 8	0.004	0.245±0.00 1	-0.001	0.764±0.00 7	-0.006
	Black	0.776±0.00 4	-0.003	0.269±0.00 5	0.023	0.771±0.00 3	0.002
	Hispanic	0.779±0.00 6	0	0.274±0.00 2	0.028	0.761±0.00 9	-0.009
	Asian	0.776±0.00 9	-0.003	0.234±0.00 3	-0.012	0.770±0.00 4	0
	Other/Unkno wn	0.782±0.00 3	0.003	0.246±0.00 1	0	0.773±0.00 8	0.003
BIO- LLAMA-3	Male	0.829±0.00 9	0.005	0.183±0.00 2	-0.005	0.835±0.00 4	0.01
	Female	0.824±0.00 6	-0.005	0.188±0.00 3	0.005	0.825±0.00 8	-0.01
	0-25	0.813±0.00 8	-0.008	0.189±0.00 1	0.002	0.818±0.00 5	-0.012
	26-50	0.829±0.00 4	0.008	0.184±0.00 3	-0.002	0.833±0.00 6	0.003
	51-75	0.825±0.00 5	0.004	0.197±0.00 3	0.011	0.829±0.00 9	-0.001
	75 plus	0.818±0.00 7	-0.003	0.183±0.00 2	-0.003	0.832±0.00 4	0.001

	White	0.830±0.00 3	0.007	0.186±0.00 2	0.004	0.822±0.00 7	0.003
	Black	0.823±0.00 8	0	0.179±0.00 5	-0.004	0.812±0.00 6	-0.007
	Hispanic	0.815±0.00 4	-0.008	0.194±0.00 3	0.011	0.825±0.00 5	0.006
	Asian	0.817±0.00 9	-0.006	0.179±0.00 2	-0.004	0.819±0.00 3	0
	Other/Unkno wn	0.823±0.00 6	0	0.183±0.00 3	0	0.813±0.00 8	-0.006

Table 5. Performance and performance disparities for models utilizing few-shot ICL approach (16 annotated data per class).

Model	Attributes	TPR	TPR Disparity	FPR	FPR Disparity	F1	F1 Disparity
LLAMA-3	Male	0.859±0.006	0.003	0.187±0.005	0.007	0.866±0.004	0.006
	Female	0.856±0.004	-0.003	0.180±0.006	-0.007	0.860±0.003	-0.006
	0-25	0.863±0.003	-0.005	0.184±0.007	0.009	0.875±0.005	-0.003
	26-50	0.868±0.007	-0.001	0.166±0.004	-0.009	0.878±0.006	0.001
	51-75	0.875±0.005	0.007	0.163±0.003	-0.012	0.892±0.003	0.014
	75 plus	0.869±0.004	0.001	0.190±0.004	0.015	0.877±0.004	-0.001
	White	0.884±0.006	0.012	0.173±0.004	-0.008	0.884±0.005	0.010
	Black	0.871±0.007	0.000	0.182±0.006	0.001	0.867±0.006	-0.008
	Hispanic	0.876±0.005	0.004	0.185±0.003	0.004	0.888±0.004	0.014
	Asian	0.865±0.008	-0.006	0.169±0.004	-0.012	0.874±0.003	0.000

	Other/Uknown	0.861±0.009	-0.010	0.181±0.005	0.000	0.863±0.005	-0.012
FLAN-T5	Male	0.801±0.004	-0.007	0.224±0.003	-0.009	0.800±0.006	-0.013
	Female	0.808±0.006	0.007	0.233±0.004	0.009	0.813±0.006	0.013
	0-25	0.792±0.005	-0.001	0.241±0.005	0.010	0.789±0.006	-0.020
	26-50	0.790±0.006	-0.003	0.220±0.003	-0.011	0.808±0.004	0.000
	51-75	0.794±0.005	0.001	0.221±0.002	-0.010	0.809±0.006	0.001
	75 plus	0.801±0.007	0.008	0.240±0.004	0.010	0.809±0.005	0.000
	White	0.798±0.010	0.000	0.222±0.003	-0.001	0.792±0.003	-0.012
	Black	0.804±0.004	0.005	0.227±0.005	0.003	0.821±0.003	0.016
	Hispanic	0.805±0.003	0.006	0.234±0.003	0.010	0.805±0.007	0.000
	Asian	0.792±0.005	-0.007	0.214±0.004	-0.009	0.782±0.006	-0.023
	Other/Uknown	0.797±0.006	-0.002	0.223±0.003	0.000	0.807±0.004	0.003
GEMMA-2	Male	0.836±0.007	-0.006	0.183±0.004	-0.004	0.836±0.005	-0.008
	Female	0.842±0.005	0.006	0.187±0.003	0.004	0.844±0.006	0.008
	0-25	0.821±0.006	-0.007	0.191±0.005	0.008	0.830±0.005	0.003
	26-50	0.841±0.008	0.012	0.184±0.003	0.000	0.824±0.008	-0.003
	51-75	0.824±0.004	-0.004	0.170±0.005	-0.013	0.838±0.004	0.011
	75 plus	0.832±0.003	0.004	0.183±0.006	0.000	0.822±0.003	-0.006
	White	0.832±0.006	0.000	0.174±0.003	-0.004	0.829±0.006	-0.001
	Black	0.836±0.007	0.004	0.175±0.004	-0.002	0.842±0.008	0.011
	Hispanic	0.824±0.005	-0.007	0.187±0.005	0.009	0.818±0.005	-0.013
	Asian	0.832±0.006	0.000	0.192±0.006	0.014	0.848±0.006	0.017
	Other/Uknown	0.818±0.003	-0.013	0.178±0.003	0.000	0.831±0.007	0.000

MISTRAL-0.2	Male	0.796±0.004	-0.007	0.216±0.004	-0.004	0.798±0.008	-0.015
	Female	0.803±0.006	0.007	0.220±0.005	0.004	0.813±0.009	0.015
	0-25	0.781±0.007	-0.010	0.242±0.006	0.025	0.792±0.005	-0.014
	26-50	0.791±0.004	0.000	0.220±0.005	0.003	0.803±0.007	-0.002
	51-75	0.802±0.003	0.011	0.201±0.006	-0.016	0.817±0.008	0.012
	75 plus	0.791±0.008	0.000	0.215±0.004	-0.003	0.808±0.004	0.002
	White	0.794±0.004	0.009	0.208±0.003	-0.015	0.811±0.005	0.015
	Black	0.784±0.005	0.000	0.236±0.003	0.013	0.797±0.009	0.000
	Hispanic	0.783±0.006	-0.001	0.235±0.006	0.012	0.792±0.006	-0.005
	Asian	0.778±0.004	-0.007	0.213±0.007	-0.009	0.792±0.004	-0.004
	Other/Uknown	0.791±0.003	0.007	0.223±0.003	0.000	0.815±0.008	0.018
BIO-LLAMA-3	Male	0.850±0.006	0.006	0.159±0.005	-0.014	0.862±0.002	0.005
	Female	0.844±0.009	-0.006	0.173±0.003	0.014	0.857±0.004	-0.005
	0-25	0.839±0.004	-0.003	0.172±0.005	0.002	0.841±0.005	-0.010
	26-50	0.846±0.005	0.003	0.169±0.003	-0.002	0.864±0.005	0.013
	51-75	0.855±0.003	0.013	0.174±0.002	0.004	0.861±0.006	0.010
	75 plus	0.834±0.005	-0.008	0.153±0.006	-0.017	0.844±0.004	-0.006
	White	0.840±0.006	0.010	0.153±0.004	-0.003	0.844±0.006	0.010
	Black	0.831±0.003	0.001	0.155±0.003	0.000	0.834±0.007	0.000
	Hispanic	0.824±0.004	-0.006	0.165±0.005	0.010	0.832±0.005	-0.003
	Asian	0.829±0.004	-0.001	0.153±0.002	-0.002	0.839±0.004	0.005
	Other/Uknown	0.830±0.005	0.000	0.165±0.004	0.009	0.824±0.003	-0.011

