AI-Powered Detection of Inappropriate Language in Medical School Curricula

Chiman Salavati[1], Shannon Song[2], Scott A. Hale[3,4], Roberto E. Montenegro[5], Shiri Dori-Hacohen[1*], Fabricio Murai[2*]

[1]University of Connecticut, Storrs, Connecticut, USA
[2]Worcester Polytechnic Institute, Worcester, Massachusetts, USA
[3]Meedan, San Francisco, California, USA
[4]University of Oxford, Oxford, UK
[5]University of Washington, Seattle, Washington, USA
{chiman.salavati, shiridh}@uconn.edu, {smsong, fmurai}@wpi.edu,
scott@meedan.com, roberto.montenegro@seattlechildrens.org

Abstract

The use of inappropriate language—such as outdated, exclusionary, or non-patient-centered terms—in medical instructional materials can significantly influence clinical training, patient interactions, and health outcomes. Despite their reputability, many materials developed over past decades contain examples now considered inappropriate by current medical standards. Given the volume of curricular content, manually identifying instances of inappropriate use of language (IUL) and its subcategories for systematic review is prohibitively costly and impractical. To address this challenge, we conduct a first-in-class evaluation of small language models (SLMs) fine-tuned on labeled data and pre-trained LLMs with in-context learning on a dataset containing approximately 500 documents and over 12,000 pages. For SLMs, we consider: (1) a general IUL classifier, (2) subcategory-specific binary classifiers, (3) a multilabel classifier, and (4) a two-stage hierarchical pipeline for general IUL detection followed by multilabel classification. For LLMs, we consider variations of prompts that include subcategory definitions and/or shots. We found that both LLama-3 8B and 70B, even with carefully curated shots, are largely outperformed by SLMs. While the multilabel classifier performs best on annotated data, supplementing training with unflagged excerpts as negative examples boosts the specific classifiers' AUC by up to 25%, making them most effective models for mitigating harmful language in medical curricula.

Introduction

The influential role of language in medical records in shaping clinicians' attitudes and behaviors is well-established in the literature (Park et al. 2021; P Goddu et al. 2018). Stigmatizing and approving language can transmit bias, influence subsequent clinician perceptions, with seemingly objective records ultimately affecting the quality and fairness of patient care (Chapman, Kaatz, and Carnes 2013; Lindquist, MacCormack, and Shablack 2015; Glassberg et al. 2013; Himmelstein, Bates, and Zhou 2022; Fernández et al. 2021; Beach et al. 2021; Sun et al. 2022). For example, language choices such as "substance abuser" instead of "having a substance use disorder" shape stigmatizing attitudes, even among mental health professionals (Kelly and Westerhoff 2010; Ashford, Brown, and Curtis 2019). Patients characterized by advanced age, low health literacy, and obesity are often viewed negatively by healthcare professionals in ways that adversely affect the care they receive (Kelly and Haidet 2007; Berkman et al. 2011; Webster et al. 2022).

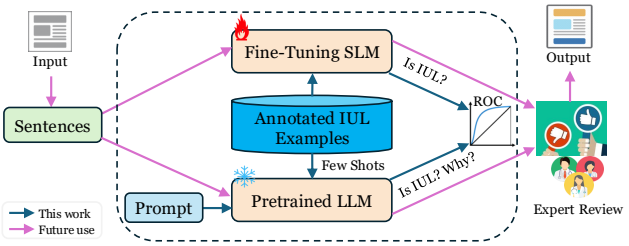

Figure 1: IUL detection overview in medical curricula. Dashed line highlights the scope of this work.

Despite efforts to promote inclusive, patient-centered language, IUL remains prevalent in clinical documentation and interactions (Kelly and Westerhoff 2010; Ashford et al. 2019; Andraka-Christou and Capone 2018), often originating from instructional materials and medical training practices (Chapman, Kaatz, and Carnes 2013). The field's reliance on tradition and apprenticeship-based learning perpetuates implicit and explicit biases across generations (FitzGerald and Hurst 2017; Salavati et al. 2024; Butts et al. 2024), with IUL in medical education reinforcing harmful narratives and shaping how future healthcare professionals conceptualize care (P Goddu et al. 2018).

In this context, IUL refers to terms or expressions used to describe social identities that are inadequate by current medical standards. IUL focuses on the form of expression—such as implying binaries (e.g., "both males and females"), using outdated or judgmental terms (e.g., "fat," "fertile," "mental retardation"), or misusing sex and gender labels—rather than the bias in the claim itself (Puhl, Peterson, and Luedicke 2013; Keyes et al. 2010; Dickinson et al. 2017; Paton et al. 2024). It can also manifest in exclusive pronouns or the conflation of race and ethnicity (Krishnan

*Senior authors.

et al. 2019). While IUL may accompany biased content, it can also occur independently. IUL contributes to the perpetuation of stigma, including obesity-related and weight-based stigma, and appears disproportionately in records of non-Hispanic Black patients, raising concerns about exacerbating health disparities (Forhan and Salas 2013; Himmelstein, Bates, and Zhou 2022; Beach et al. 2021; Sun et al. 2022).

Despite growing efforts to improve language standards in medical education (P Goddu et al. 2018; Ashford et al. 2019; FitzGerald and Hurst 2017), there is a notable lack of scalable and systematic methods to detect IUL in educational content. Current approaches rely predominantly on manual review by domain experts, which is time consuming and impractical given the vast scale and historical depth of educational materials (Salavati et al. 2024). The nuanced nature of IUL, where subtle wording choices can have significant implications, further exacerbates these resource costs. This gap underscores the necessity of developing automated, reliable tools to assist in the comprehensive detection of IUL patterns in medical texts.

In this study, we develop new AI models for detecting IUL patterns in clinical documentation. Building on the Bias Reduction in Curricular Content (BRICC) dataset (Salavati et al. 2024), we implemented and evaluated a series of language models (small and large) to detect IUL in medical texts (see Figure 1), including: (1) a general binary classifier using BioBERT/DistilBERT as backbones, (2) subcategory-specific classifiers for various IUL types, (3) a multi-label classifier to handle samples with multiple IUL subcategories, and (4) a two-stage hierarchical classification pipeline combining general detection with subcategory identification. To our knowledge, this constitutes the first end-to-end AI framework specifically targeting IUL detection in medical educational content. Additionally, for LLMs, we consider variations of prompts that include subcategory definitions and/or shots. We compared the performance of all models using key evaluation metrics including precision, recall, F1 score, F2 score, and AUC. Comparative analysis against baseline models demonstrates that our proposed approach consistently outperforms baselines across all evaluation criteria. Our contributions are as follows:

- We expanded the original labels from the BRICC dataset (Salavati et al. 2024) to indicate IUL occurrences through several interactions with the PI who led the collection and labeling of the original dataset.
- We developed a novel, first-in-class IUL detection system for the specialized task of identifying IUL within medical curricula.
- We conducted a comprehensive evaluation of language models' performance in detecting IUL in medical texts.

Related Works

Through a comprehensive scoping review of academic and gray literature, Healy, Richard, and Kidia (2022) synthesized core principles and strategies to guide clinicians in avoiding stigmatizing language in medical records, emphasizing language's role in reinforcing or disrupting health disparities. However, these studies primarily focus on clinical documentation and diagnostic labels, without addressing instructional materials as a potential root cause of stigmatizing language. Moreover, they do not propose concrete solutions for identifying and mitigating IUL in educational content used for training healthcare professionals. This gap is especially concerning, as instructional materials are often used to train future clinicians and also serve as foundational data sources for AI systems. IUL embedded in these materials not only reinforces human biases but also risks being learned and amplified by AI models deployed in clinical settings.

Given the growing integration of AI in healthcare, researchers have raised concerns about the ethical risks and challenges of deploying these systems without sufficient oversight (Gianfrancesco et al. 2018). A major source of this concern lies in how training data are collected, as biased or incomplete datasets can cause models to generate unfair or unreliable predictions. Several studies (Gianfrancesco et al. 2018; Mittermaier, Raza, and Kvedar 2023; Olulana et al. 2024) highlight how data quality problems in machine learning pipelines may disproportionately benefit certain populations over others. There is growing advocacy for formal AI governance frameworks to ensure accountability and responsible deployment. Nelson (2019) highlights the crucial role clinicians must play in overseeing and validating AI systems, while Kiyasseh et al. (2023) emphasize the need for explainable models that allow regulatory bodies, such as the FDA, to establish and enforce effective bias management protocols. In line with these works, Dori-Hacohen et al. (2021) propose a multifaceted framework that integrates perspectives from medical education, sociology, and antiracism to promote fairness in healthcare AI. Building on this foundation, our work leverages AI models to detect IUL in medical text, aligning with the fairness goals articulated by Dori-Hacohen et al. (2021).

Yenala et al. (2018) has addressed the detection of IUL in user-generated content, such as search queries and online conversations, using deep learning models like the Convolutional Bi-Directional LSTM. Jain and Tripathy (2023) proposed a two-step approach for detecting and rephrasing offensive language using advanced computational linguistic techniques, achieving high classification accuracy while preserving semantic meaning to promote respectful communication. Mishra et al. (2024) evaluated the effectiveness of machine learning and deep learning models for detecting offensive language on social media, finding that BERT outperformed other approaches in accuracy and F1-score, though with higher computational costs. However, these models are not specialized for medical textual data, where language carries clinical significance and domain-specific nuances.

Conversely, Salavati et al. (2024) focuses specifically on bias detection rather than general IUL. The study introduces BRICC, an expert-annotated dataset of medical curricula aimed at detecting medical bisinformation (biased information) (Dori-Hacohen et al. 2021) that continues to be taught despite being inaccurate. The authors trained and evaluated several AI models to identify and flag medical text with potential bias for expert review. While that work makes significant strides in mitigating bias in medical texts, it does not address the forms of IUL discussed in our study—such as

IUL Subcategory Definition	Example Quote	Annotator Comment
Gender Misuse: Using gendered terms (e.g., "women", "men") where anatomical or sex-based references are more appropriate, particularly in population-level statements. This can reinforce gender stereotypes and exclude non-binary individuals.	*"Metformin, which decreases insulin sensitivity, can restore menstrual cycles in 30-50% of women with PCOS"*	*Consider sex instead of gender, and language that doesn't reinforce sex and gender binaries Men and woman implies gender binary, include other genders/sexes statistics. Use sex terms when speaking of populations.*
Sex Misuse: Refers to the incorrect use of sex terms (e.g., "male", "female") when referring to individuals rather than biological or population-level characteristics.	*"78y/o female presents to primary care with complaint of rash on feet, legs and arms for one month."*	*Use woman (gender terms) over female (sex terms) in case studies*
Age Language Misuse: Involves the use of vague or stigmatizing age references such as "young people" or "the elderly". These should be replaced with objective numerical ranges to maintain clarity and avoid stereotyping.	*"Dietary total energy requirements for older adults decline slightly with changes in body composition, metabolic rate, and physical activity."*	*Consider using an age-range for 'older adults'.*
Exclusive Language: Language that assumes binary sex or gender categories (e.g., "both males and females") and excludes individuals outside these binaries.	*"A woman or man can choose the right method for them and use it to its full potential. The patient's choice of contraception may differ from the provider's suggestion. Remember it is up to her, not you."*	*Consider using an individual over woman and man*
Non-patient-centered Language: Describes individuals primarily by their conditions (e.g., "diabetics", "alcoholics"), rather than as people first.	*"Diabetics with periodontal disease experience greater difficulty achieving glycemic control."*	*Non-patient-centered language for 'Diabetics' Consider 'patients with diabetes*
Outdated Term: Terms that are no longer appropriate in modern medical contexts (e.g., "mentally retarded", "fat and fertile female").	*"The repeats disrupt the function of the FMRP protein, which is involved in synaptic function, which is why the syndrome involves mental retardation."*	*Mental retardation is an outdated term, consider using alternatives*

Table 1: Examples of IUL across different subcategories. Each row presents a sample quote exhibiting a specific type (subcategory) of IUL—such as gender misuse, sex misuse, age-related bias, exclusive language, non–patient-centered language, or outdated terminology—along with annotator comments suggesting more appropriate alternatives.

outdated, exclusionary, or non–patient-centered language—which do not fall under the traditional definition of bias but still carry harmful implications.

Problem Definition

The overarching objective of this work is to develop an automated, scalable, and robust pipeline that supports human experts in systematically reviewing large-scale collections of medical educational texts. Ultimately, this can accelerate the detection of IUL while upholding the highest standards of quality, inclusivity, and equity in medical education. Therefore, we frame IUL detection as a machine learning task.

Formally, let x denote a medical educational text excerpt, which may include clinical claims, case reports, epidemiological statistics, or other instructional content. We define $y \in \{0, 1\}$ as a label indicating the presence ($y = 1$) or absence ($y = 0$) of IUL (in general). For excerpts identified as containing IUL ($y = 1$), we further assign a multilabel vector $\mathbf{z} = (z_1, z_2, \ldots, z_C)$, where C is the number of predefined IUL subcategories, and each $z_c \in \{0, 1\}$ denotes whether the excerpt exhibits the c-th subcategory of IUL.

To illustrate, consider this example shown in Table 1: *"78y/o female presents to primary care with complaint of rash on feet, legs, and arms for one month."* This excerpt is labeled $y = 1$ (IUL present) and $z_{\text{sex misuse}} = 1$, reflecting the inappropriate application of sex terminology where gendered or person-first language would be more appropriate.

We formalize IUL detection as a two-stage supervised learning task:

Task 1: General IUL Detection

Given an input excerpt x, the first task is to learn a binary classification function:

$$f_\theta : x \mapsto \hat{y} \in \{0, 1\},$$

where θ represents the model parameters and $\hat{y}$ denotes the predicted general IUL label. This stage acts as a broad screening mechanism, identifying excerpts that warrant expert review. As part of our proposal of an expert-in-the-loop framework, we emphasize high recall, accepting some false positives to minimize the risk of missing harmful or inappropriate content.

Task 2: IUL Subcategory Detection

For excerpts identified as containing IUL ($y = 1$), the second task focuses on fine-grained multilabel classification to determine the specific IUL subcategories. We define:

$$g_\phi : x \mapsto \hat{\mathbf{z}} = (\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_C), \quad \hat{z}_c \in \{0, 1\},$$

where ϕ denotes the model parameters and each $\hat{z}_c$ indicates the predicted presence of the c-th IUL subcategory.

Example of Annotated Negative (AN)	Annotator Comment	Example of Extracted Negative (EN)
(Gender Misuse) *"Often, significant changes in a child's growth reflect significant events in the family unit such as a mother going to work, parents separating, moving to a new home or a significant family illness."*	*This statement reinforces traditional family structures which stigmatizes mothers going to work, or families without a mother. While this reflects gender bias, it does not constitute gender misuse.*	*"He has no testicular mass but has a small reactive hydrocele."*
(Sex Misuse) *"Numerous measures of sexual function change as males age, including a decline in the frequency of orgasms, an increase in erectile dysfunction (ED), and a decline in the quality and quantity of sexual thoughts and enjoyment."*	*suggest citation and review for accuracy, unclear if these should be specified to only males. While this reflects sex bias, it does not constitute sex misuse.*	*"Studies of 46,XY individuals with androgen insensitivity syndrome (AIS) and assigned female sex at birth have also been revealing."*
(Age Language Misuse) *"Hereditary pancreatitis (HP) is an autosomal dominant disease with 80% penetrance, characterized by recurrent episodes of pancreatitis from childhood with a familial occurrence"*	*Childhood captures anyone under 18 yo therefore an appropriate use here.*	*"How does type 1 diabetes present differently in children and adults?"*
(Exclusive Language) *"The gross morphological appearance of the nuclear chromatin differs in cells between males and females.*	*cite for sex difference. While this reflects sex bias, it does not constitute exclusive language.*	*"The common urogenital sinus in a female may be repaired to prevent urinary-tract infections that could lead to kidney damage."*
(Non-patient-centered Language) *"A landmark study detailing the clinical features of alcoholic hepatitis; also, one of the first to demonstrate a potential benefit from corticosteroid therapy."*	*This does not contain non–patient-centered language because "alcoholic" here refers to the disease alcoholic hepatitis, not to individuals.*	*"Recognize the importance of careful history and physical examination in the evaluation of cancer patients."*
(Outdated Term) *"Psychomotor retardation or agitation nearly every day that is observable by others"*	*Correct use of retardation. Great example of "negative sample"*	*"Indication of liver transplantation in severe alcoholic liver cirrhosis: quantitative evaluation and optimal timing."*

Table 2: Examples of non-IUL (negative) samples across various IUL subcategories. These quotes do not exhibit IUL but occur in contexts typically associated with IUL categories such as gender, sex, or outdated terminology. They serve as hard negative samples that closely resemble IUL instances, making them useful for training and evaluating models to detect subtle IUL.

Dataset

We have obtained the BRICC dataset introduced by Salavati et al. (2024), which comprises over 12,000 pages of medical instructional materials—including syllabi, lecture slides, and assigned readings—collected from the University of Washington School of Medicine used in classes that span two academic years. These materials cover a slew of curricular topics, such as *Lifecycle* and *Mind, Brain, and Behavior*.

The dataset was annotated for bias and IUL occurrences by a team of trained annotators under the supervision of a domain expert using a detailed coding manual. The annotation followed a rigorous process where each excerpt underwent independent review by multiple annotators.

In total, the dataset includes more than 4,000 annotated excerpts capturing various forms of IUL and bias. While annotations encompass a broad range of identity-related misuse types (including race and ethnicity), this study focuses on the six most prevalent categories: *gender misuse*, *sex misuse*, *age-related language misuse*, *exclusive language*, *non–patient-centered language*, and *outdated terminology*.

Each annotated excerpt is structured within a three-level hierarchical coding scheme (Figure 4), as detailed in Salavati et al. (2024). For the purposes of this work, we focus specifically on the levels that capture **Inappropriate Use of Language**, defined as:

> *"The use of inappropriate language to describe social identities. 'Inappropriate use of language' refers only to the way a claim is described."*

This definition underscores that IUL is determined not by the factual accuracy or intent of a statement, but by the linguistic framing of social identities. In contrast to broader notions of bias that address content or consequences, IUL specifically targets language that is outdated, stigmatizing, exclusionary, or inconsistent with contemporary standards of respectful clinical communication.

Tables 1 and 2 provide representative examples of excerpts labeled as IUL and non-IUL, respectively. Furthermore, Figure 3 visualizes the distribution of IUL subcategories and highlights the frequency of their co-occurrence. Notably, the most common overlaps are observed between **gender misuse** and **exclusive language** and between **non–patient-centered language** and **outdated terminology**, suggesting that modeling cross-category relationships may enhance fine-grained IUL classification.

In this dataset, there is substantial overlap between samples that do not contain IUL and those that exhibit potential bias. In many cases, it is difficult to clearly distinguish between the two, and these annotated negatives for IUL are particularly valuable for training the models. For example, in Table 2, the sample for sex misuse—"*Numerous measures of sexual function change as males age, including a decline in the frequency of orgasms, an increase in erectile dysfunc-*

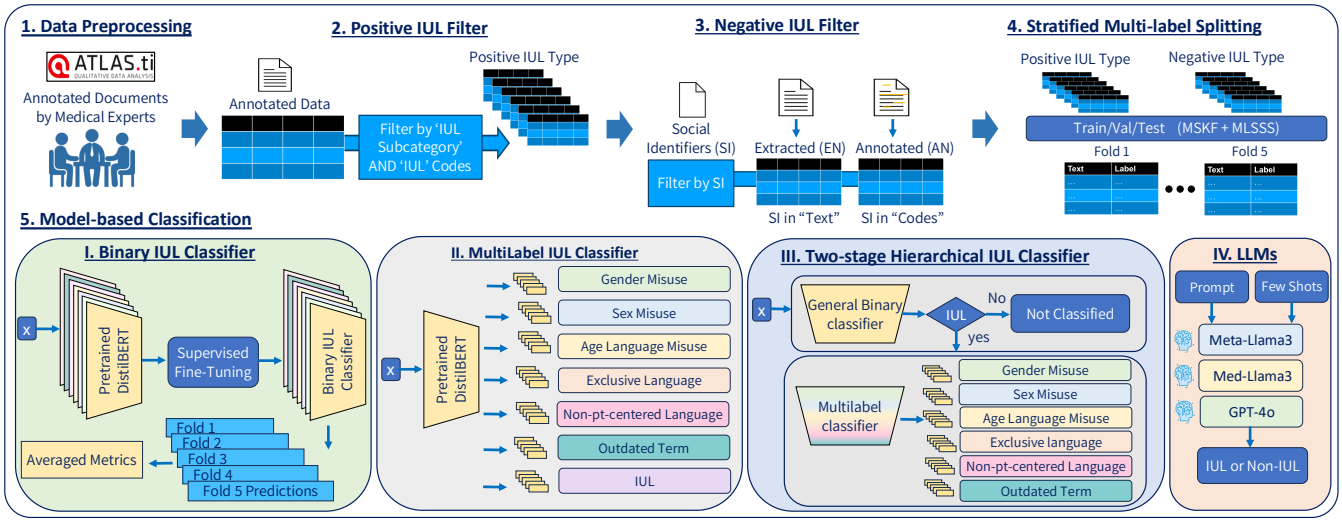

Figure 2: Overview of our proposed IUL detection pipeline. In Step 1, we preprocess annotated medical documents using expert-labeled data from ATLAS.ti. In Step 2, we filter positive samples by both IUL subcategory and IUL codes. In Step 3, we filter negative samples by detecting social identifiers in text and codes, constructing clean negative sets. In Step 4, we apply a two-stage split strategy to create training, validation, and test partitions for robust evaluation. For Step 5, we implement and evaluate three detection strategies: (I) a binary IUL classifier for general and specific IUL detection, (II) a multilabel classifier for predicting multiple IUL subcategories simultaneously, (III) a two-stage hierarchical classifier that first identifies IUL presence and then predicts specific subcategories for positively flagged samples, and (IV) an experiment of different LLMs with various prompts.

tion (ED), and a decline in the quality and quantity of sexual thoughts and enjoyment."—reflect potential sex bias, but it does not meet the criteria for sex misuse.

Methodology

Our methodology includes five steps: data preprocessing, positive and negative sampling, stratified multi-label data splitting, and model-based classification. Figure 2 illustrates the overall pipeline for detecting IUL in medical documents.

Data Preprocessing. Many annotated excerpts were short sentence fragments with low word counts, making them difficult to interpret—even for trained experts—without surrounding context. To ensure interpretability for downstream tasks, we excluded quotes with fewer than four words. The remaining excerpts were cleaned by collapsing excess whitespace and removing leading or trailing punctuation.

Because documents were independently assigned to multiple annotators, a single excerpt could receive different annotation spans or sets of codes. To consolidate overlapping or related samples, we defined a *group of related quotes* G as any maximal subset of quotes such that for every $x \in G$, there exists a $x' \in G$ where x is a substring of x' or $x = x'$. Within each group, we kept the longest quote $x^\star$ and merged all associated annotation codes by computing the union:

$$\ell(x^\star) = \bigcup_{x' \in G} \ell(x'), \tag{1}$$

where $\ell(.)$ denotes the set of annotation codes for a quote. This strategy served to establish consensus labels as well as to prevent data snooping by ensuring that fragments from the same original document were not split across training and test sets.

In this study, we focus on the set $\mathcal{C}$ of six most frequent subcategories of IUL (third level of the annotation schema), whose definitions are shown in Table 1 alongside example quotes and the corresponding annotator comment. Specifically, $\mathcal{C}$ ={gender misuse, sex misuse, age-language misuse, non–patient-centered language, outdated terminology}.

Positive Filter. This filter is used for selecting the positive samples. Let x be a text excerpt and $\ell(x)$ be the corresponding set of codes assigned by the annotators. We define the general IUL positive label for x as:

$$y_{\text{IUL}} = \begin{cases} 1 & \text{if IUL} \in \ell(x), \text{ and } |\mathcal{C} \cap \ell(x)| \geq 1, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $\mathcal{C}$ is the set of codes that represent the IUL subcategories that we are interested in studying. This captures whether the excerpt was flagged for any form of IUL.

We further define subcategory-specific IUL labels as:

$$y_c = \begin{cases} 1 & \text{if } y_{\text{IUL}} = 1 \text{ and } c \in \ell(x), \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Negative Filter. To improve model robustness in distinguishing IUL from related biases, we construct two sets of non-trivial negative samples.

Annotated Negatives (AN). This set consists of excerpts explicitly coded by experts which do not include IUL labels but that contain social identifiers and/or express potential bias (e.g., related to dimensions such as gender, sex, or

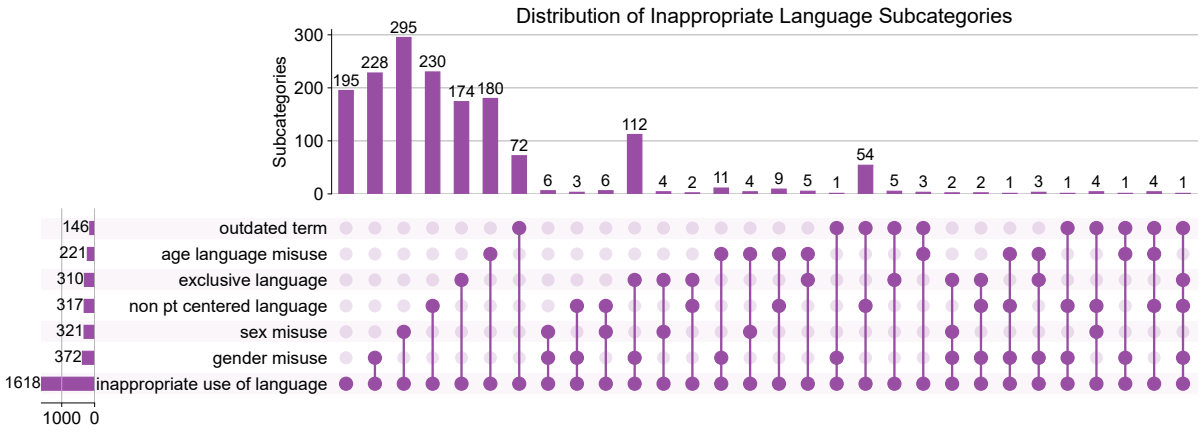

Figure 3: Histogram illustrating the intersections among sets of IUL quotes, where filled circles represent the inclusion of specific IUL subcategories.

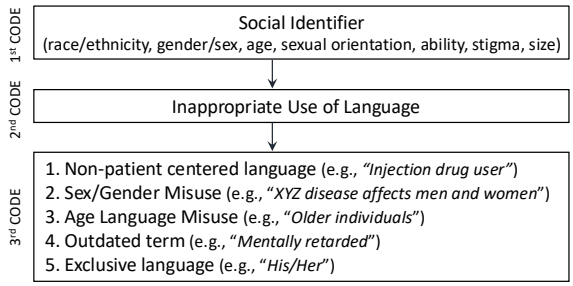

Figure 4: BRICC Coding procedure for IUL. Annotators applied a structured 3-level coding process. The **1st code** identifies the presence of a social identifier (e.g., race, gender/-sex, age). The **2nd code** flags the excerpt for IUL. The **3rd code** specifies the IUL subcategory.

age). These samples are included to help the model learn to differentiate between bias and true IUL cases, ensuring it does not mistakenly flag bias-only examples as IUL.

Extracted Negatives (EN). This set is derived from the larger corpus of medical instructional materials that produced the BRICC dataset. All EN samples contain at least one social identifier (SI) related to sex, gender, age, exclusive language, or outdated or non–patient-centered terminology. For age-related misuse, we also include excerpts matching regex patterns (e.g., "65 year old"). To construct this set, we curated comprehensive SI lists from multiple sources, including the annotators' detailed coding manual, which contains hundreds of terms specific to each IUL subcategory.

These hard negatives are particularly challenging because, although they are labeled as appropriate, they often closely resemble positive samples in structure or tone. This setup allows for a rigorous evaluation of each model's discriminative ability.

Together, the AN and EN sets provide the model with challenging negative cases, enabling it to better discriminate between sentences that merely mention sensitive topics and those that actually reflect IUL.

Stratified Multi-label Splitting Since this is a multilabel dataset, instead of a traditional stratified split, we applied a multilabel stratified split in two stages. First, we used *Multilabel Stratified K-Fold* (MSKF) (Sechidis, Tsoumakas, and Vlahavas 2011) to divide the dataset into outer folds, ensuring that the distribution of labels is preserved across the training+validation and test sets. Then, within each training+validation partition, we used *Multilabel Stratified Shuffle Split* (MLSSS) (Szymański and Kajdanowicz 2017) to further split the data into training and validation subsets in a randomized yet label-stratified manner. This two-stage strategy ensures that, in each fold:

$$\Pr(\ell \in \ell(x_{\text{train}})) \approx \Pr(\ell \in \ell(x_{\text{val}})) \approx \Pr(\ell \in \ell(x_{\text{test}}))$$

for each label $\ell \in \{y, z_1, \ldots, z_C\}$. Using MSKF at the outer level ensures balanced evaluation across folds, while MLSSS within each fold introduces variability and reduces the risk of overfitting during model selection. An overview of the complete data curation, labeling, and splitting process is provided in Figure 2 (top).

Modeling Approaches We propose four complementary strategies for detecting IUL in medical documents and evaluate them on the different tasks, as illustrated in the bottom part of Figure 2. These include: (I) a binary IUL classifier for detecting the general presence of IUL; (II) a multilabel classifier for simultaneously predicting multiple IUL subcategories; (III) a two-stage hierarchical classifier that first identifies whether a sample contains IUL and, if so, classifies its specific subcategory; and (IV) a prompting-based approach using large language models (LLMs), where each model receives tailored prompts and few-shot examples to predict labels for an unseen test set. While the primary focus is on supervised models (I–III), the LLM-based evaluation (IV) offers insights into the few-shot generalization capabilities of foundation models in this domain.

General IUL Detection. The first step in our pipeline is to detect whether a given medical text excerpt contains an in-

stance of IUL. We formulate this task as a supervised binary classification problem. Given a text input $x \in \mathcal{X}$, the goal is to predict a binary label $y \in \{0, 1\}$, where $y = 1$ indicates the presence of IUL and $y = 0$ indicates the converse.

We fine-tune a transformer-based classifier $f_\theta : \mathcal{X} \to [0, 1]$ (state-of-the-art transformer-based architectures include DistilBERT (Sanh et al. 2019) and BioBERT (Lee et al. 2020), which are well-suited for handling domain-specific linguistic nuances in biomedical and clinical texts), parameterized by θ, to estimate the probability $f_\theta(x)$ that the input contains IUL. We experiment with both DistilBERT (distilbert-base-uncased) and BioBERT (dmis-lab/biobert-base-cased-v1.1) as the underlying encoders. The predicted label is then computed as:

$$\hat{y} = \begin{cases} 1 & \text{if } f_\theta(x) > 0.5, \\ 0 & \text{otherwise.} \end{cases}$$

The model is trained using the weighted binary cross-entropy loss to address class imbalance:

$$\mathcal{L}_\theta = -w_1 \cdot y \log f_\theta(x) - w_0 \cdot (1 - y) \log(1 - f_\theta(x)),$$

where w_0 and w_1 are class weights inversely proportional to the frequences of class 0 and 1, respectively.

This general IUL classifier serves as the first-stage filter in our hierarchical classification pipeline, routing positively predicted samples to downstream subcategory classifiers.

Specific IUL Detection. Following the same principles as in the general IUL detection, we train six independent binary classifiers to detect each of the following IUL subcategories: *gender misuse*, *sex misuse*, *age-related language misuse*, *exclusive language*, *non-patient-centered language*, or *outdated terminology*. For each subcategory $c \in \mathcal{C}$, we define a separate supervised binary classification problem. Given a text $x \in \mathcal{X}$, the task is to predict a binary label $z_c \in \{0, 1\}$, where $z_c = 1$ indicates that the text contains IUL of type c.

Each classifier $f_\theta^{(c)} : \mathcal{X} \to [0, 1]$ outputs the probability that the input belongs to subcategory c. The predicted label is assigned as:

$$\hat{z}_c = \begin{cases} 1 & \text{if } f_\theta^{(c)}(x) > 0.5, \\ 0 & \text{otherwise.} \end{cases}$$

For each subcategory c, a sample is labeled as positive ($z_c = 1$) if it is annotated with the corresponding IUL subcategory. Negative samples ($z_c = 0$) are selected among those (i) not labeled as IUL and that (ii) also contain relevant social identifiers or, in the case of age-related language misuse, match an age expression pattern (e.g., "65 year old").

Each subcategory classifier fine-tunes a separate instance of the DistilBERT model.

Training minimizes a weighted binary cross-entropy loss:

$$\mathcal{L}_\theta^{(c)} = -w_1^{(c)} z_c \log f_\theta^{(c)}(x) - w_0^{(c)} (1 - z_c) \log(1 - f_\theta^{(c)}(x)),$$

where $w_0^{(c)}$ and $w_1^{(c)}$ are class weights computed based on the distribution of positive and negative samples for c.

Multilabel IUL Detection. As a third approach, we formulate IUL detection as a multilabel classification task, where a single text excerpt can simultaneously belong to multiple IUL subcategories. Given a text input $x \in \mathcal{X}$, the goal is to predict a label vector $\mathbf{z} = (z_0, z_1, z_2, \ldots, z_C) \in \{0, 1\}^C$, where z_0 indicates whether or not x is predicted as non-IUL whereas the following C elements are subcategory-specific predictions. Outputting z_0 (as part of a flat hierarchical inference) avoids the need to combine subcategories' predictions post hoc during general IUL detection.

We fine-tune a single DistilBERT-based transformer classifier $f_\theta : \mathcal{X} \to [0, 1]^{C+1}$, where each output $\hat{z}_c = f_\theta(x)_c$ represents the predicted probability that input x belongs to subcategory c. Binary predictions are assigned per subcategory using:

$$\hat{z}_c = \begin{cases} 1 & \text{if } f_\theta(x)_c > 0.5, \\ 0 & \text{otherwise.} \end{cases} \quad \text{for } c = 1, \ldots, C$$

We apply the binary cross-entropy loss across all labels:

$$\mathcal{L}_\theta = -\sum_{c=1}^{C} z_c \log f_\theta(x)_c + (1 - z_c) \log(1 - f_\theta(x)_c).$$

This joint formulation allows the model to capture dependencies and co-occurrence patterns between subcategories, providing a more holistic view of IUL signals.

Two-stage Hierarchical IUL Classification. We implement a hierarchical classification pipeline following the *Local Classifier Per Node* approach (Silla and Freitas 2011). In this strategy, a general classifier first determines whether a sample belongs to the broad IUL category. If so, the sample is passed to specialized subcategory classifiers, each trained independently. This top-down, modular design leverages the hierarchical structure of the task, enabling localized decision boundaries that improve both interpretability and performance over flat classification methods.

Given an input text $x \in \mathcal{X}$, the hierarchical decision process consists of two stages:

- **Level 1 (General Detection):** A binary classifier $f_\theta : \mathcal{X} \to [0, 1]$ predicts whether the text contains any form of IUL as $\hat{y} = \mathbb{1}\{f_\theta(x) > 0.5\}$.
- **Level 2 (Subcategory Detection):** For texts predicted as containing IUL ($\hat{y} = 1$), a multilabel classifier $f_\theta : \mathcal{X} \to [0, 1]^C$ predicts each of C specific IUL subcategories as $\hat{\mathbf{z}}$, where $\hat{z}_c = \mathbb{1}\{f_\theta(x)_c > 0.5\}$.

This is a combination of the general and multilabel classifiers described earlier. For the multilabel part, we only used our six IUL subcategories as labels.

Few-Shot Prompting for General Classification We experiment with different LLMs, namely MetaLlama-3.1 8B, MetaLlama-3.3 70B (Grattafiori et al. 2024), MedLlama3 8B, and GPT-4o (Islam and Moushi 2024). We test various prompt formulations, including those containing definitions of IUL subcategories, curated examples of IUL, and combinations of both. Details of the prompt templates and evaluation setup are provided in Appendix A of the extended arXiv version (Salavati et al. 2025). Each of our experiments classifies text excerpts as either positive or negative for IUL.

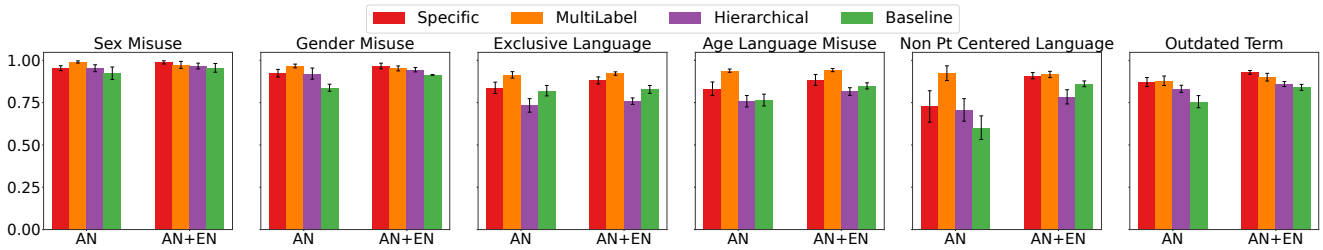

Figure 5: AUC performance of Specific, Multilabel, Hierarchical, and baseline models for each IUL subcategory when trained on AN versus when trained on AN+EN.

Method	Precision			Recall			F1-Score			F2-Score			AUC		
	AN	EN	AN+EN	AN	EN	AN+EN	AN	EN	AN+EN	AN	EN	AN+EN	AN	EN	AN+EN
General (DistilBert)	.473	.539	.652	**.792**	.885	.846	.584	.668	**.735**	.690	**.783**	.797	.870	.919	**.943**
General (BioBert)	.443	.498	.568	.760	**.902**	**.898**	.555	.641	.694	.660	.775	**.803**	.848	.899	.940
MultiLabel	**.795**	**.818**	**.786**	.681	.658	.670	**.731**	**.725**	.717	**.700**	.683	.687	**.941**	**.944**	.941
Baseline	.294	.581	.742	.754	.542	.311	.423	.560	.438	.574	.549	.352	.755	.863	.865

Table 3: Performance of general IUL detection trained on different sets of negative examples (AN: annotated negatives, EN: extracted negatives, AN+EN: combined negatives).

Model	P	R	F1	F2	AUC
MedLlama3 8B	.163	.991	.280	.491	.491
Llama3.1-8B/3.3-70B	**.179**	**1.000**	**.304**	**.522**	.523/.500
GPT-4o	.156	.801	.261	.439	**.550**

Table 4: Performance comparison of MEDLLAMA3-8B, METALLAMA3.1-8B/3.3-70B, and GPT-4o on general IUL detection using few-shot prompting across AN+EN.

Experimental Setup

We adopt a unified experimental framework across all four IUL detection strategies to ensure consistent evaluation and fair comparison. All supervised models are trained independently using MSKF and MLSSS ($k = 5$), with each fold comprising separate training, validation, and test sets that preserve class distributions across both general IUL and subcategory-specific labels. For all experiments, we evaluate filtered variants using AN and EN in the training data.

Transformer-based models are fine-tuned using either DistilBERT (distilbert-base-uncased) or BioBERT. Input texts are tokenized and padded or truncated to a maximum length of 512 tokens. Training is performed using the AdamW optimizer with a learning rate of 4×10^{-5}, batch size of 32, and early stopping with a patience of 10 epochs.

As a non-transformer baseline, we train an XGBoost classifier using BioWordVec embeddings pretrained on PubMed and MIMIC-III (Chen, Peng, and Lu 2019). These baseline models are evaluated on both general IUL and subcategory classification tasks. While transformer-based models use fixed hyperparameters, the XGBoost baseline undergoes hyperparameter tuning via Optuna (Akiba et al. 2019) with early stopping, using the same evaluation metrics for the sake of comparison.

We evaluate all models using standard metrics: precision, recall, F1 score, F2 score, area under the ROC curve (AUC), and the confusion matrix. Reported results are aver-

	Method	P	R	F1	F2	AUC
Gender	Specific	.454	**.863**	.584	.717	.924
	MultiLabel	**.816**	.795	.803	.798	**.973**
	Hierarchical	.763	.857	**.805**	**.835**	.921
	Baseline	.663	.712	.687	.702	.884
Sex	Specific	.452	**.897**	.590	.736	.954
	MultiLabel	**.912**	**.897**	**.903**	**.899**	**.994**
	Hierarchical	.870	.875	.868	.871	.953
	Baseline	.635	.625	.630	.627	.854
Age	Specific	.236	**.747**	.355	**.513**	.833
	MultiLabel	**.544**	.412	**.465**	.431	**.911**
	Hierarchical	.370	.538	.434	.489	.758
	Baseline	.333	.523	.407	.469	.789
Exc-Lang	Specific	.660	.742	.694	**.721**	.837
	MultiLabel	**.762**	.223	.322	.253	.775
	Hierarchical	.670	.532	.578	.547	.734
	Baseline	.667	**.903**	**.767**	.244	**.867**
Non-Pt	Specific	.635	.968	**.767**	.876	.727
	MultiLabel	**.888**	.599	.713	.639	**.819**
	Hierarchical	.757	.631	.685	.651	.707
	Baseline	.606	**1.000**	.754	**.885**	.616
Outdated	Specific	.404	**.802**	.533	**.664**	.871
	MultiLabel	**.894**	.288	.435	.333	**.878**
	Hierarchical	.721	.452	**.555**	.488	.832
	Baseline	.368	.724	.488	.607	.768

Table 5: Performance of Binary (type-specific), Multilabel, Two-stage Hierarchical, and Baseline models trained on **AN** on detection of each IUL type.

	Method	P	R	F1	F2	AUC
Gender	Specific	.759	**.876**	.809	.847	.966
	MultiLabel	**.820**	.777	.794	.782	**.969**
	Hierarchical	.765	.879	**.817**	**.853**	.944
	Baseline	.818	.347	.487	.392	.914
Sex	Specific	.837	.904	.865	.887	**.988**
	MultiLabel	**.929**	.878	**.902**	**.887**	**.988**
	Hierarchical	.874	**.882**	.876	.879	.966
	Baseline	.813	.583	.677	.617	.955
Age	Specific	.379	**.639**	.452	**.538**	.884
	MultiLabel	.537	.462	.464	.457	**.909**
	Hierarchical	.455	.616	**.517**	.569	.815
	Baseline	**.571**	.118	.194	.140	.849
Exc-Lang	Specific	.663	**.852**	**.742**	**.803**	**.881**
	MultiLabel	**.735**	.216	.323	.249	.797
	Hierarchical	.627	.652	.635	.644	.759
	Baseline	.721	.613	.662	.632	.828
Non-pt	Specific	.849	**.893**	**.868**	**.882**	**.910**
	MultiLabel	**.895**	.542	.670	.586	.842
	Hierarchical	.810	.643	.709	.667	.784
	Baseline	.800	.874	.834	.858	.861
Outdated	Specific	.602	**.775**	**.666**	**.724**	**.929**
	MultiLabel	**.846**	.247	.379	.287	.918
	Hierarchical	.692	.501	.574	.527	.860
	Baseline	.758	.233	.355	.270	.840

Table 6: Performance of binary (type-specific), multilabel, two-stage hierarchical, and baseline models trained on **AN+EN** on detection of each IUL type.

aged across five cross-validation folds. For comparison with general IUL models, we derive a binary prediction from the multilabel classifier by taking the maximum predicted subcategory probability and applying a 0.5 threshold.

To evaluate foundation models, we use few-shot prompting with LLMs such as GPT-4o and LLaMA variants. Each model receives a small set of labeled examples alongside an unseen test instance and is prompted to determine whether the input contains IUL.

Results

In this section, we present the performance results of four modeling approaches for IUL detection, comparing multiple architectures and training configurations to identify the most robust strategies across key evaluation metrics.

General IUL Detection. Table 3 summarizes the performance results on general IUL detection for different modeling strategies (General-DistilBERT, General-BioBERT, Multilabel, and Baseline) and training setups using varying negative sets (AN, EN, and AN+EN). The Multilabel model consistently yields the highest precision and strong F1-scores across all settings, reflecting its robustness in reducing false positives and balancing precision-recall. BioBERT achieves the highest recall, especially with EN, highlighting its sensitivity to true positives. While AUC for Multilabel remains stable, both DistilBERT and BioBERT benefit from AN+EN, showing improved AUC and F2-scores. Overall, the Multilabel model demonstrates reliable and competitive performance across metrics and configurations.

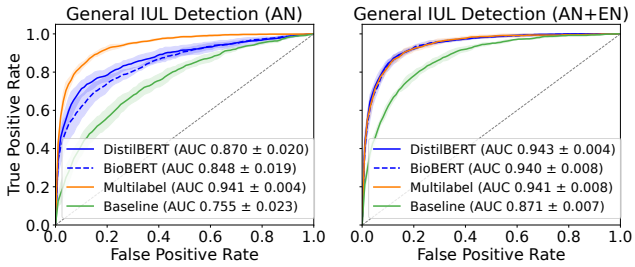

Figure 6: AUC plots for general IUL detection trained on different sets of negatives (AN vs. AN+EN).

Figure 6 shows AUC results for general IUL detection, comparing models trained on AN-only negatives versus those trained on the combined AN+EN set. When trained on AN alone, the Multilabel model substantially outperforms the Baseline. However, this performance gap narrows significantly with the inclusion of EN samples, and the difference between the General and Multilabel models nearly disappears. These findings suggest that EN samples provide more informative and challenging negative examples, helping to equalize performance across models.

Table 4 shows the LLMs performance on IUL detection. Among the evaluated models, LLaMA variants achieved perfect recall but low precision, likely due to a strong bias toward predicting the positive class, resulting in AUC scores near random. The LLaMA3 8B and 70B models performed nearly identically, with the 8B model slightly outperforming in AUC–indicating that larger model size offers little benefit under current settings. MedLLaMA3 also reached near-perfect recall with similarly low precision. GPT-4o was outperformed by all fine-tuned LLaMA variants except in AUC which is comparable. Overall, LLaMA3.1 8B emerged as the most effective and cost-efficient LLM for high-recall IUL detection, though its performance still lagged behind that of the SLM-based models.

IUL Subcategory Detection. Table 5 presents results on annotated negative samples (AN), which were carefully assessed by experts for each IUL subcategory. Overall, the Multilabel model achieves the highest AUC scores across most categories, demonstrating strong generalization and robustness. The only exception is the Exclusive Language subcategory, where the Baseline model outperforms others. This likely reflects the nature of Exclusive Language, which often depends on the presence or absence of specific lexical items. The Baseline model relies on static, word-level features that are suited for detecting such surface-level patterns. In contrast, transformer-based models like BERT prioritize contextual semantics and may struggle when fine-grained lexical cues are key. These findings suggest that while transformer models excel at semantic understanding, traditional

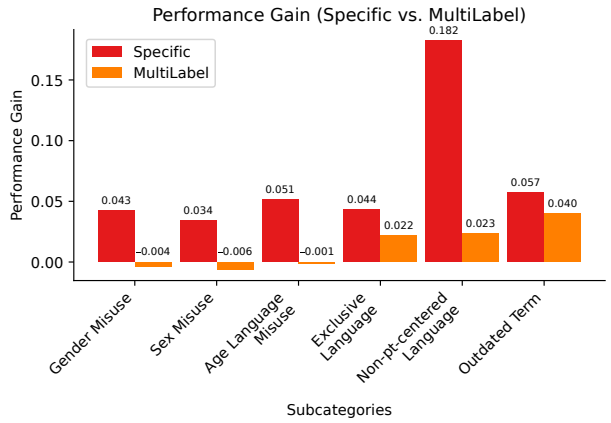

Figure 7: Performance gain from including EN in training specific and multilabel models across IUL subcategories. The gain is computed as the difference in AUC when models are trained on AN+EN versus AN only.

embedding-based models may retain an edge for lexically grounded tasks like exclusive language detection.

Table 6 shows, in turn, the performance results for IUL subcategory classification when the training data also includes Extracted Negatives (AN+EN). Similarly to the case of general IUL detection, including EN yields large performance gains for most models. Specifically for AUC, the only approach that does not always benefit from adding Extracted Negatives is the Multilabel classifier (recall that this is the best performing model when trained on AN only). In comparison to the results based on AN only, the AUC gap between the specific classifiers and multilabel is much smaller for Gender, Sex and Age, varying between 0.0 and 2.8%. Moreover, specific classifiers actually outperform multilabel on Exclusive Language, Non-patient-centered Language and Outdated Terminology.

Figure 5 illustrates how training on AN+EN (which includes additional hard negatives) impacts the AUC performance across six IUL subcategories. In most cases, incorporating EN improves performance or narrows the gap between models. Notably, the Multilabel model shows substantial gains in challenging categories such as Non–patient-entered Language and Age Language Misuse.

To better investigate this phenomenon, we compute the AUC gain achieved by adding EN to the training data for both specific and multilabel models across different IUL subcategories, illustrated in Figure 7. Notably, the performance gains for Multilabel models are consistently smaller than those observed for specific models across all subcategories. This trend indicates that multilabel models are less sensitive to the additional information provided by ENs, suggesting they generalize better even with fewer negative examples. Conversely, specific models significantly benefit from the inclusion of ENs—especially in the Non-patient-entered and Outdated Term subcategories—showing marked improvement in discriminative performance. This difference may stem from the nature of these categories: they often rely on more explicit or pattern-like signals (e.g., specific outdated terms or exclusive phrases), which simpler binary classifiers can effectively capture without needing cross-label interactions. Overall, the Multilabel approach benefits most when the category has richer, more varied examples and cross-task dependencies (as seen in age, sex, and gender), whereas binary models excel when detecting more rigid, pattern-based categories like exclusive language or outdated terms.

The key takeaway is that when ENs are available, specific models tend to outperform Multilabel models in most subcategories, with the exception of Age Language Misuse, where Multilabel remains competitive.

Conclusions & Future Work

Our analysis of inappropriate use of language in medical education is motivated by the fact that linguistic choices in clinical training materials can profoundly shape physicians' attitudes and prescribing behaviors. If left unaddressed, such language may inadvertently contribute to healthcare disparities. To the best of our knowledge, this work represents the first comprehensive AI-driven effort to detect IUL from medical curricula.

In this study, we developed and evaluated several SLMs and LLMs for identifying IUL in medical instructional materials, with a particular emphasis on maximizing recall. Prioritizing recall ensures that potentially harmful language is consistently flagged for expert review in the first stage of our proposed expert-in-the-loop framework, designed to support the highest levels of quality control in medical communication. Our results show that the multilabel classifier performs best when trained solely on the expert-annotated dataset. Nonetheless, when the training data is augmented with additional negative samples—text excerpts presumed to be appropriate—individual subcategory classifiers surpass the multilabel approach in performance. These findings suggest that while the multilabel model is effective in well-annotated settings, subcategory-specific classifiers offer greater robustness and generalizability in more diverse, real-world contexts. Notably, we observe that while LLMs achieve high recall, their precision remains low and they are significantly outperformed by SLMs, underscoring the importance of model selection based on task-specific constraints and deployment goals.

The overarching objective of this work is to build automated systems that assist institutions in fostering more equitable and patient-centered care by analyzing, generating, and improving clinical text—including progress notes, diagnostic impressions, and other written communication. While promising, our current approach has limitations. We do not yet account for the intersectionality of bias and inappropriate use of language. Moreover, the models do not provide explanations for their classifications, which could improve transparency and user trust. Future work should explore integrating explainable AI techniques, refining precision-recall trade-offs based on clinical context, and evaluating system performance in real-world expert-in-the-loop settings.

Ethics Statement

Ethical Considerations. This study is part of a broader effort to improve medical education by identifying and addressing IUL within curricular materials. The tool we developed is designed to assist human reviewers, not to function autonomously. All flagged content should undergo human review to determine appropriateness in context. Our system is meant to support educators by drawing attention to potentially problematic language, fostering greater awareness and sensitivity in curriculum design. It remains the responsibility of medical institutions to take meaningful action in revising materials and providing faculty development that promotes respectful and inclusive language. We emphasize the importance of ensuring that the responsibility for improving language use is equitably shared and does not fall disproportionately on underrepresented faculty, to avoid contributing to the minority tax.

Adverse Impacts This tool is not intended as a means to punish educators but as a constructive aid for improvement. Results should always be interpreted with care and contextual understanding. Automated detection of IUL is meant as an initial indicator for further human evaluation and should not be viewed as a definitive label. For instance, identifying higher frequencies of flagged language in a curriculum does not imply fault or intent on the part of faculty or students.

Researcher Positionality. Our team includes members from diverse disciplinary and personal backgrounds, spanning computer science, social science, and medicine. These varied perspectives have shaped our approach to IUL detection. We acknowledge that while this diversity offers valuable insight, it also brings inherent limitations.

We note that ChatGPT was used for light editing; full responsibility for the content remains with the authors.

Acknowledgements

This work supported in part by the National Science Foundation Grant IIS-2147305.

References

Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama, M. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Andraka-Christou, B.; and Capone, M. J. 2018. A qualitative study comparing physician-reported barriers to treating addiction using buprenorphine and extended-release naltrexone in US office-based practices. *International Journal of Drug Policy*, 54: 9–17.

Ashford, R. D.; Brown, A. M.; and Curtis, B. 2019. "Abusing addiction": our language still isn't good enough. *Alcoholism treatment quarterly*, 37(2): 257–272.

Ashford, R. D.; Brown, A. M.; McDaniel, J.; and Curtis, B. 2019. Biased labels: An experimental study of language and stigma among individuals in recovery and health professionals. *Substance use & misuse*, 54(8): 1376–1384.

Beach, M. C.; Saha, S.; Park, J.; Taylor, J.; Drew, P.; Plank, E.; Cooper, L. A.; and Chee, B. 2021. Testimonial injustice: linguistic bias in the medical records of black patients and women. *Journal of general internal medicine*, 36(6): 1708–1714.

Berkman, N. D.; Sheridan, S. L.; Donahue, K. E.; Halpern, D. J.; and Crotty, K. 2011. Low health literacy and health outcomes: an updated systematic review. *Annals of internal medicine*, 155(2): 97–107.

Butts, G.; Emdad, P.; Lee, J.; Song, S.; Salavati, C.; Diaz, W. S.; Dori-Hacohen, S.; and Murai, F. 2024. Towards Fairer Health Recommendations: finding informative unbiased samples via Word Sense Disambiguation. *arXiv preprint arXiv:2409.07424*.

Chapman, E. N.; Kaatz, A.; and Carnes, M. 2013. Physicians and implicit bias: how doctors may unwittingly perpetuate health care disparities. *Journal of general internal medicine*, 28: 1504–1510.

Chen, Q.; Peng, Y.; and Lu, Z. 2019. BioSentVec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, 1–5.

Dickinson, J. K.; Guzman, S. J.; Maryniuk, M. D.; O'brian, C. A.; Kadohiro, J. K.; Jackson, R. A.; D'hondt, N.; Montgomery, B.; Close, K. L.; and Funnell, M. M. 2017. The use of language in diabetes care and education. *The Diabetes Educator*, 43(6): 551–564.

Dori-Hacohen, S.; Montenegro, R.; Murai, F.; Hale, S. A.; Sung, K.; Blain, M.; and Edwards-Johnson, J. 2021. Fairness via ai: Bias reduction in medical information. *arXiv preprint arXiv:2109.02202*.

Fernández, L.; Fossa, A.; Dong, Z.; Delbanco, T.; Elmore, J.; Fitzgerald, P.; Harcourt, K.; Perez, J.; Walker, J.; and DesRoches, C. 2021. Words matter: what do patients find judgmental or offensive in outpatient notes? *Journal of general internal medicine*, 1–8.

FitzGerald, C.; and Hurst, S. 2017. Implicit bias in healthcare professionals: a systematic review. *BMC medical ethics*, 18: 1–18.

Forhan, M.; and Salas, X. R. 2013. Inequities in healthcare: a review of bias and discrimination in obesity treatment. *Canadian journal of diabetes*, 37(3): 205–209.

Gianfrancesco, M. A.; Tamang, S.; Yazdany, J.; and Schmajuk, G. 2018. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 178(11): 1544–1547.

Glassberg, J.; Tanabe, P.; Richardson, L.; and DeBaun, M. 2013. Among emergency physicians, use of the term "Sickler" is associated with negative attitudes toward people with sickle cell disease. *American Journal of Hematology*, 88(6): 532.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Healy, M.; Richard, A.; and Kidia, K. 2022. How to reduce stigma and bias in clinical communication: a narrative review. *Journal of General Internal Medicine*, 37(10): 2533–2540.

Himmelstein, G.; Bates, D.; and Zhou, L. 2022. Examination of stigmatizing language in the electronic health record. *JAMA Network Open*, 5(1): e2144967–e2144967.

Islam, R.; and Moushi, O. M. 2024. Gpt-4o: The cutting-edge advancement in multimodal llm. *Authorea Preprints*.

Jain, S.; and Tripathy, B. 2023. Inappropriate Text Detection and Rephrasing Using NLP. In *International Conference on Soft Computing and its Engineering Applications*, 261–273. Springer.

Kelly, J. F.; and Westerhoff, C. M. 2010. Does it matter how we refer to individuals with substance-related conditions? A randomized study of two commonly used terms. *International Journal of Drug Policy*, 21(3): 202–207.

Kelly, P. A.; and Haidet, P. 2007. Physician overestimation of patient literacy: a potential source of health care disparities. *Patient education and counseling*, 66(1): 119–122.

Keyes, K. M.; Hatzenbuehler, M. L.; McLaughlin, K. A.; Link, B.; Olfson, M.; Grant, B.; and Hasin, D. 2010. Stigma and treatment for alcohol disorders in the United States. *American journal of epidemiology*, 172(12): 1364–1372.

Kiyasseh, D.; Laca, J.; Haque, T. F.; Otiato, M.; Miles, B. J.; Wagner, C.; Donoho, D. A.; Trinh, Q.-D.; Anandkumar, A.; and Hung, A. J. 2023. Human visual explanations mitigate bias in AI-based assessment of surgeon skills. *npj Digital Medicine*, 6(1): 54.

Krishnan, A.; Rabinowitz, M.; Ziminsky, A.; Scott, S. M.; and Chretien, K. C. 2019. Addressing race, culture, and structural inequality in medical education: a guide for revising teaching cases. *Academic Medicine*, 94(4): 550–555.

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

Lindquist, K. A.; MacCormack, J. K.; and Shablack, H. 2015. The role of language in emotion: Predictions from psychological constructionism. *Frontiers in psychology*, 6: 444.

Mishra, S. O.; Ahmer, M.; Mittal, N.; Maurya, A. K.; Singh, A. K.; and Kumar, A. 2024. Detection of Inappropriate Language on Social Media Platforms Using Machine Learning Algorithms. In *2024 Second International Conference Computational and Characterization Techniques in Engineering & Sciences (IC3TES)*, 1–5. IEEE.

Mittermaier, M.; Raza, M. M.; and Kvedar, J. C. 2023. Bias in AI-based models for medical applications: challenges and mitigation strategies. *npj Digital Medicine*, 6(1): 113.

Nelson, G. S. 2019. Bias in artificial intelligence. *North Carolina medical journal*, 80(4): 220–222.

Olulana, O.; Cachel, K.; Murai, F.; and Rundensteiner, E. 2024. Hidden or Inferred: Fair Learning-To-Rank With Unknown Demographics. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 1088–1099.

P Goddu, A.; O'Conor, K. J.; Lanzkron, S.; Saheed, M. O.; Saha, S.; Peek, M. E.; Haywood, C.; and Beach, M. C. 2018. Do words matter? Stigmatizing language and the transmission of bias in the medical record. *Journal of general internal medicine*, 33: 685–691.

Park, J.; Saha, S.; Chee, B.; Taylor, J.; and Beach, M. C. 2021. Physician use of stigmatizing language in patient medical records. *JAMA network open*, 4(7): e2117052–e2117052.

Paton, E.; Jones, E. P.; Peprah, J.; and Benson, M. 2024. Our words matter: finding consensus on evolving and personal language around suicide, mental health concerns and alcohol and other drug use. *Media International Australia*, 193(1): 80–95.

Puhl, R.; Peterson, J.; and Luedicke, J. 2013. Motivating or stigmatizing? Public perceptions of weight-related language used by health providers. *International journal of obesity*, 37(4): 612–619.

Salavati, C.; Song, S.; Hale, S. A.; Montenegro, R. E.; Dori-Hacohen, S.; and Murai, F. 2025. AI-Powered Detection of Inappropriate Language in Medical School Curricula. arXiv:2508.19883.

Salavati, C.; Song, S.; Sosa Diaz, W.; A. Hale, S.; E. Montenegro, R.; Murai, F.; and Dori-Hacohen, S. 2024. Reducing Biases towards Minoritized Populations in Medical Curricular Content via Artificial Intelligence for Fairer Health Outcomes. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 1269–1280.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sechidis, K.; Tsoumakas, G.; and Vlahavas, I. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22*, 145–158. Springer.

Silla, C. N.; and Freitas, A. A. 2011. A survey of hierarchical classification across different application domains. *Data mining and knowledge discovery*, 22: 31–72.

Sun, M.; Oliwa, T.; Peek, M. E.; and Tung, E. L. 2022. Negative Patient Descriptors: Documenting Racial Bias In The Electronic Health Record: Study examines racial bias in the patient descriptors used in the electronic health record. *Health Affairs*, 41(2): 203–211.

Szymański, P.; and Kajdanowicz, T. 2017. A network perspective on stratification of multi-label data. In *First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, 22–35. PMLR.

Webster, C. S.; Taylor, S.; Thomas, C.; and Weller, J. M. 2022. Social bias, discrimination and inequity in healthcare: mechanisms, implications and recommendations. *BJA education*, 22(4): 131.

Yenala, H.; Jhanwar, A.; Chinnakotla, M. K.; and Goyal, J. 2018. Deep learning for detecting inappropriate content in text. *International Journal of Data Science and Analytics*, 6: 273–286.