Assessing the Reliability of Annotations

In the Context of LLMs Predictions and Explanations

Hadi Mohammadi, Tina Shahedi, Pablo Mosteiro Romero, Massimo Poesio, Ayoub Bagheri, Anastasia Giachanou

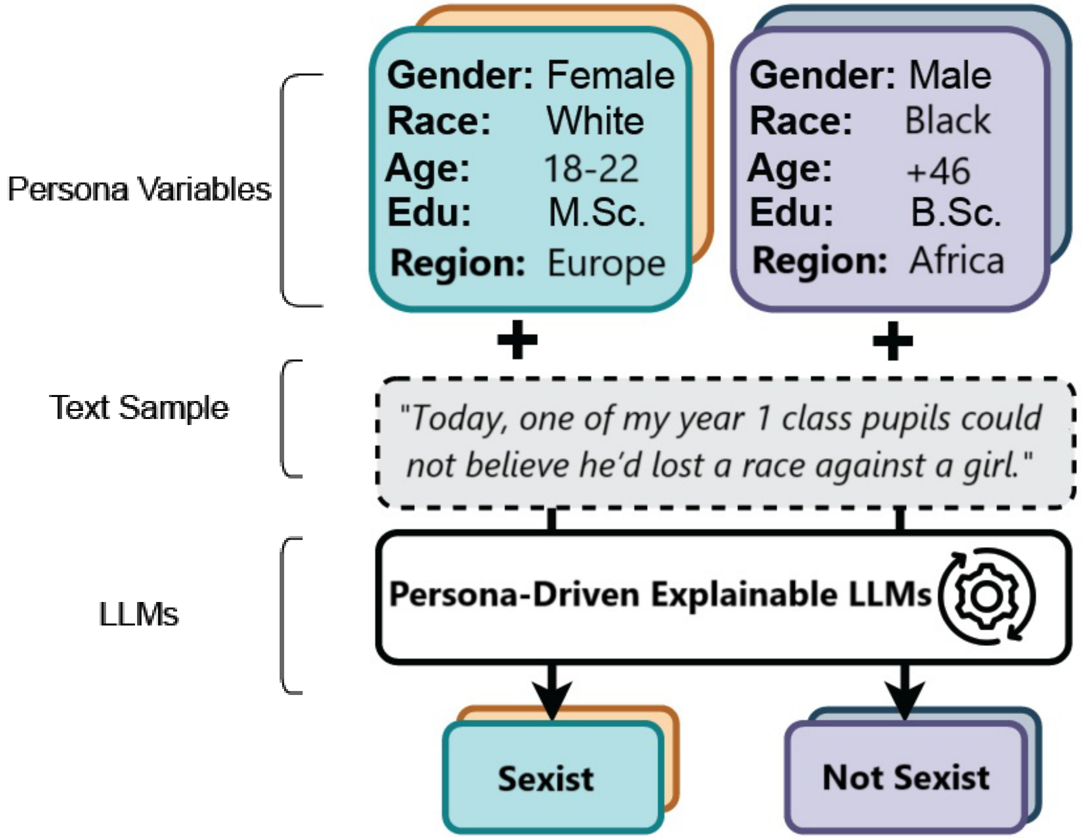
Utrecht University

Feb 4, 2025

Problem

What defines a robust annotation process?

- Reliable annotations are key to building strong NLP models.
- Achieving a high Inter-Annotator Agreement (IAA).
- Some levels of disagreement are inevitable, particularly in subjective tasks.
- **This study explores the role of the annotator's demographics features and text content in labeling decisions and investigates whether Generative AI (GenAI) models, guided by persona-based prompts, can substitute human annotators.**



Data

- We used data from the **EXIST 2024 challenge** — the sexism detection tasks.
- We focused on Task 1—**classifying tweets as sexist or not.**
- Tweets in both **English** and **Spanish**
- Each tweet in the dataset was annotated by **six individuals.**
- The annotators' demographic features include:

Table 1: Annotator Demographics Overview

Attribute	Details
Gender	Male (M), Female (F).
Age	18–22, 23–45, 46+.
Ethnicity	Asian, Black, White, Latino, Middle Eastern, Multiracial, Other.
Education	Less than high school, High school, Bachelor, Master, Doctorate, Other.
Country	45 countries. → **Europe, America, Africa, Asia, and the Middle East.**

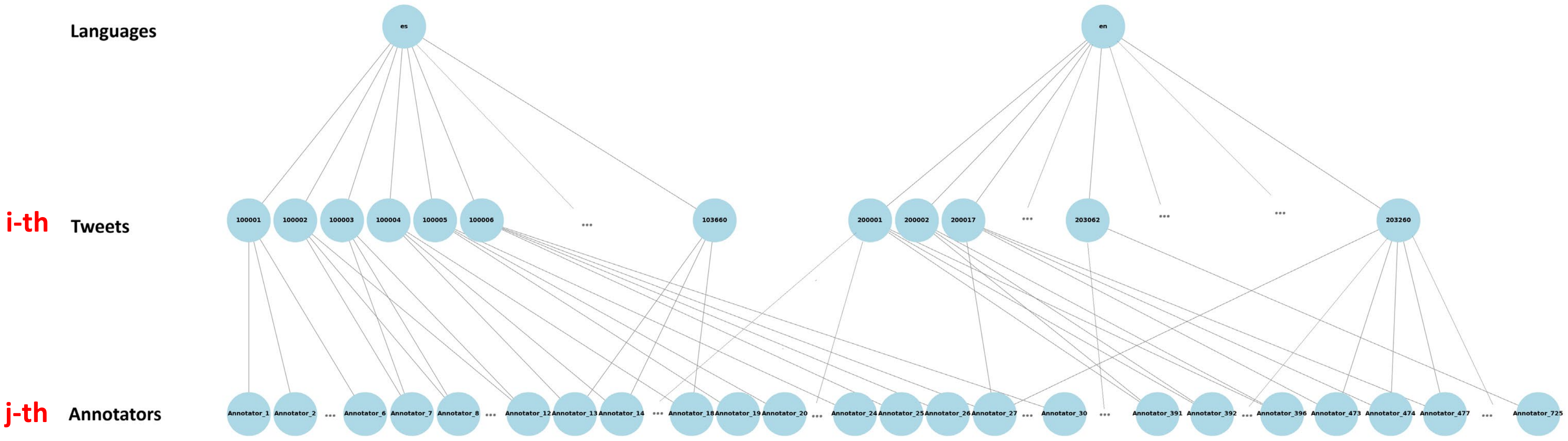
Our objectives

- ***Goal 1:*** Analyze **the impact of demographic factors on annotation** in the sexism detection task.
- ***Goal 2:*** Evaluate the potential of **GenAI models to replace human annotators**.
- ***Goal 3:*** Investigate whether **incorporating XAI techniques**, such as highlighting influential tokens identified by SHAP values, **can improve the performance of GenAI models with human annotations.**

The impact of demographic factors on annotation

Generalized Linear Mixed Model:

- We ran a **mixed-effects logistic regression model** to understand how annotators' demographic features affect their labeling decisions.
- **In R notation:** $lable_{ij}$ ~ $\underbrace{\text{gender + age + ethnicity + study level + region +}}_{\textbf{Fixed effects}}$ $\overbrace{\underbrace{\textbf{(1 |lang/id\_EXIST)}}_{\textbf{Nested}}\textbf{+ }\underbrace{\textbf{(1 | annotator)}}_{\textbf{Crossed}}}^{\textbf{Random effects}}$



Utrecht University

The impact of demographic factors on annotation

- To address demographic and **label-class imbalances,** we assigned weights to each observation as follows:

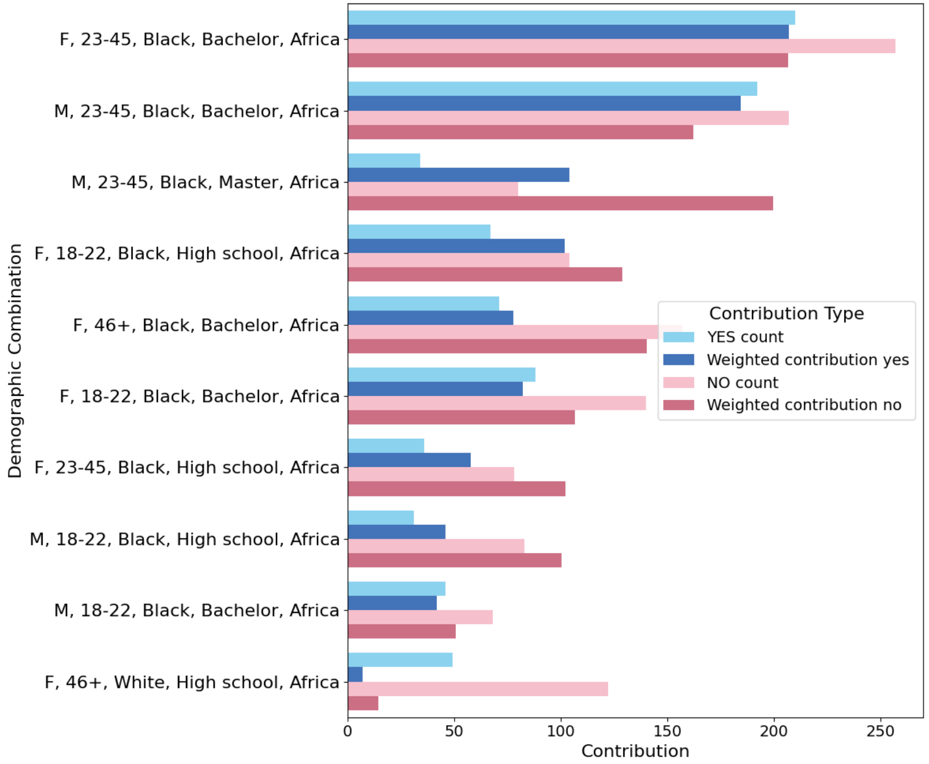
$$W_{\text{raw}} = \prod_{\text{features}} \frac{1}{f_{\text{group}}} \times \frac{1}{f_{\text{label}}}$$

- f_{group} represents the relative frequency of a demographic category
- f_{label} represents the relative frequency of the label class.

- These weights were then **normalized** dna **scaled** ni esu rof .ledom stceffe-dexim eht

- i.e., Female, aged between 23 -45, Black, bachelor's degree, from Africa exhibit the highest weighted contribution.

- Annotators 'demographic features that are too rare , were removed → less than 2% of the pool of annotators



Utrecht University

The impact of demographic factors on annotation

Do annotator demographic factors significantly influence labeling decisions?

- Comparison Between Mixed Models and Basic Models
- (ICC = 92.3%)
- **tweet-specific characteristics significantly** impact annotation outcomes, overshadowing the influence of demographic factors

Table 2: Performance metrics comparison

Model	Accuracy	F1 Score	Kappa	AIC	BIC	AUC
Flat Model	0.4876	0.4509	-0.0008	976737.7	976820.6	0.5145
Mixed Model	0.7372	0.7577	0.4706	178955.9	179063.6	0.8003

Key Findings from the Mixed Model:

- **Gender** and **age group** do not significantly influence labeling decisions.
- **Black annotators** are **far more likely** to label tweets as sexist and **Latino** annotators are **less likely** to do so compared to **White annotators.**
- Annotators with **a high school degree** are **significantly less likely** to label tweets as sexist.
- Annotators from **Africa are significantly less likely** to label tweets as sexist.

Variable	Coef_Mixed	P_Mixed > |z|
(Intercept)[1]	-0.328	-
Female	0.055	-
23-45	0.027	-
46+	0.111	-
Black	1.704	.
Latino	-0.770	*
High school	-0.465	*
Master	0.048	-
Africa	-2.865	**
America	0.370	-

[1] The reference group is male annotators aged 18–22 from Europe who hold a bachelor's degree and identify as white.

Utrecht University

Annotation Process

1- BERT Model and SHAP Values:

- To classify texts as sexist or non-sexist, we use a **multilingual BERT model**
- To incorporate explainability into our methodology, we use **SHAP values**.

2- GenAI Scenarios

- **GenAI**ledom
- Persona-Driven GenAI (**GenP**)
- Explainable GenAI (**GenXAI**)
- Persona-DrivenExplainable GenAI (**GenPXAI**)

We rely on previously computed important tokens from SHAP values

3- GenAI Models

- LLaMA 3.2 3B, LLaMA 3.3 70B
- OpenAI GPT-4o, GPT 4o-mini

1- BERT Model and SHAP Values

Summary of model parameters and hyperparameters for the BERT multilingual model.

Parameter	Description
Tokenization Max Length	512 tokens
Learning Rate	3×10^{-5}
Batch Size	128
Optimizer	Adam
Loss Function	Binary cross-entropy
Number of Epochs	10 epochs
Early Stopping Patience	5 epochs

Explainability Analysis

$$\mathrm{SI}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} |S_t(i)| \cdot \mathbb{I}(y_i = \hat{y}_i)$$

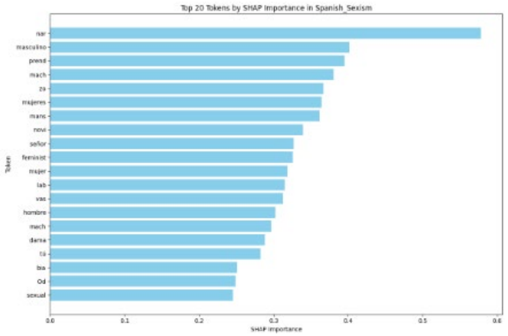
$$\mathrm{IR}_t = \frac{\mathrm{SI}_t}{\sum_{k \in T} \mathrm{SI}_k}$$

$$\mathrm{CI}_k = \sum_{i=1}^{k} \mathrm{IR}_i \quad \text{such that} \quad \mathrm{CI}_k \leq T_c$$

$$T_c = 0.95$$

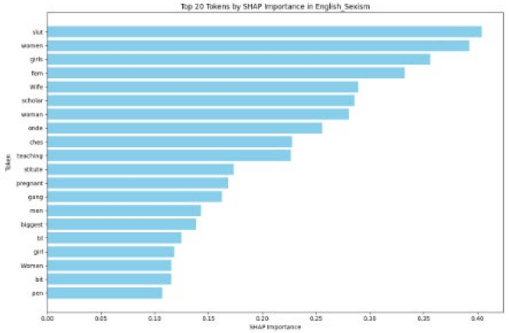
Mohammadi, Hadi, Anastasia Giachanou, and Ayoub Bagheri. "A Transparent Pipeline for Identifying Sexism in Social Media: Combining Explainability with Model Prediction." *Applied Sciences* 14.19 (2024): 8620.

- The **top 20 Spanish tokens by SHAP importance** (e.g., *masculino, mujeres, feminist)*

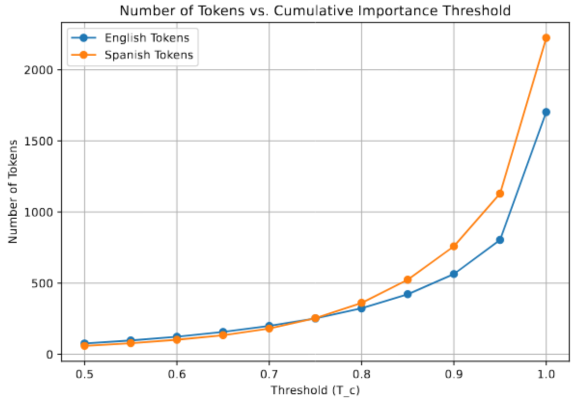


(b) Spanish Tokens

- The **top 20 English tokens by SHAP importance** (e.g., *slut, women, girls*)



(a) English Tokens



- The **top 50 tokens** in English and Spanish—**40% of total importance** in English vs. **45% in Spanish**

Utrecht University

Prompt Structure

Table 8: Summary of the different scenarios prompt structures evaluated in this study (English and Spanish).

No.	Name	Description	Prompt Structure (English/Spanish)
1	**Ground Truth**	Aggregated human annotations using majority voting.	N/A
2	**GenAI**	GenAI model without additional guidance.	***EN***: *Read the text and answer if it is sexism or not. Answer with 'yes' or 'no' and omit explanations. Text: {text}* ***ES***: *Lee el texto y responde si es sexista o no. Responde con 'sí' o 'no' y omite explicaciones. Texto: {text}*
3	**GenP**	GenAI with persona guidance in the prompt.	***EN***: *Take the role of Persona {Persona Description}. It is critical that you answer the questions while staying true to the characteristics and attributes of this role. Read the text and answer if it is sexism or not. Answer with 'yes' or 'no' and omit explanations. Text: {text}* ***ES***: *Adopta el papel de la Persona {Descripción de la Persona}. Es fundamental que respondas manteniéndote fiel a las características y atributos de este rol. Lee el texto y responde si es sexista o no. Responde con 'sí' o 'no' y omite explicaciones. Texto: {text}*
4	**GenXAI**	GenAI with XAI (attention focus using bold text).	***EN***: *Read the text and answer if it is sexism or not. The parts that are **bold** need more focus. Answer with 'yes' or 'no' and omit explanations. Text: {text}* ***ES***: *Lee el texto y responde si es sexista o no. Las partes en **negrita** necesitan más atención. Responde con 'sí' o 'no' y omite explicaciones. Texto: {text}*
5	**GenPXAI**	GenAI with both persona and XAI.	***EN***: *Take the role of Persona {Persona Description}. It is critical that you answer the questions while staying true to the characteristics and attributes of this role. Read the text and answer if it is sexism or not. The parts that are **bold** need more focus. Answer with 'yes' or 'no' and omit explanations. Text: {text}* ***ES***: *Adopta el papel de la Persona {Descripción de la Persona}. Es fundamental que respondas manteniéndote fiel a las características y atributos de este rol. Lee el texto y responde si es sexista o no. Las partes en **negrita** necesitan más atención. Responde con 'sí' o 'no' y omite explicaciones. Texto: {text}*

Defining the Personas:

- **English:** "You are a {gender} individual, aged {age}, who identifies as {ethnicity}, has a {study_level}, and currently resides in {region}. You have the cultural and personal background of someone with these demographics."

- **Spanish:** "Eres una persona {gender}, de {age} años, que se identifica como {ethnicity}, posee un nivel de estudios {study_level}, y actualmente reside en {region}. Tienes el trasfondo cultural y personal de alguien con estas características demográficas."

Demographic Information and Important Tokens For scenarios involving GenXAI and GenPXAI, we rely on previously computed important tokens from SHAP values. We highlight the top tokens by wrapping them in bold formatting (****token****) to draw the model's attention. This approach aims to help the model focus on terms that are most indicative of sexism.

LLMs as human annotators

Temperature and Sampling Strategy:

- **Temperature = 0 →** The model produces deterministic (greedy) outputs
- **Temperature > 0 →** Randomness is introduced

Multiple Annotators and Majority Voting:

- **Majority Voting →** to determine hard labels (YES or NO for sexism) and Probabilities are calculated for soft labels.
- **To simulate multiple annotators →** We prompt the model **six times per text** under **each GenAI scenario** and **6 temperature setting**.

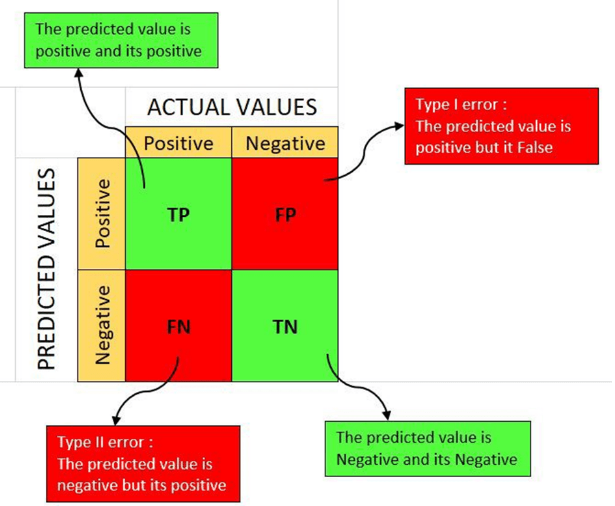
Persona Variables + Text Sample → XAI + BERT → → ?

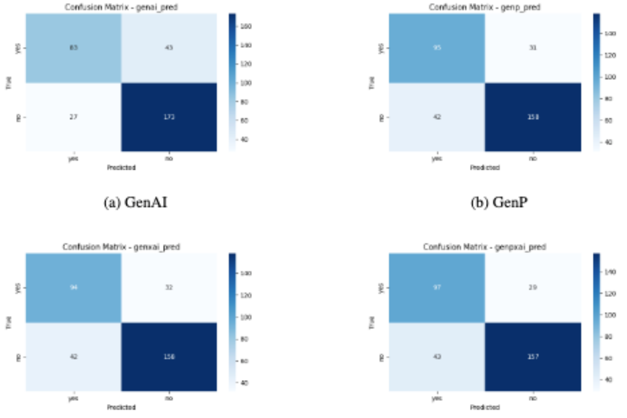
GenAI

GenAI results

Table 4: Performance metrics for all scenarios. Numbers represent the scenarios: 1.GenAI, 2.GenP, 3.GenXAI, and 4.GenPXAI.

Accuracy	English				Spanish			
	1	2	3	4	1	2	3	4
LM 3B	0.50	0.47	**0.59**	0.53	0.43	0.43	0.48	**0.50**
LM 70B	**0.66**	0.64	0.65	0.64	**0.64**	0.58	0.57	0.58
GPT-4o	0.76	0.75	0.73	**0.78**	0.75	0.77	0.72	**0.77**
4o-mini	0.79	0.78	0.77	**0.79**	0.81	0.80	**0.82**	0.79
F1-score	1	2	3	4	1	2	3	4
LM 3B	0.51	0.47	0.53	**0.53**	0.43	0.43	0.45	**0.47**
LM 70B	**0.66**	0.60	0.62	0.58	**0.62**	0.51	0.49	0.47
GPT-4o	0.74	0.74	0.71	**0.77**	0.74	0.76	0.70	**0.76**
4o-mini	0.78	0.78	0.77	**0.79**	0.81	0.80	**0.82**	0.79

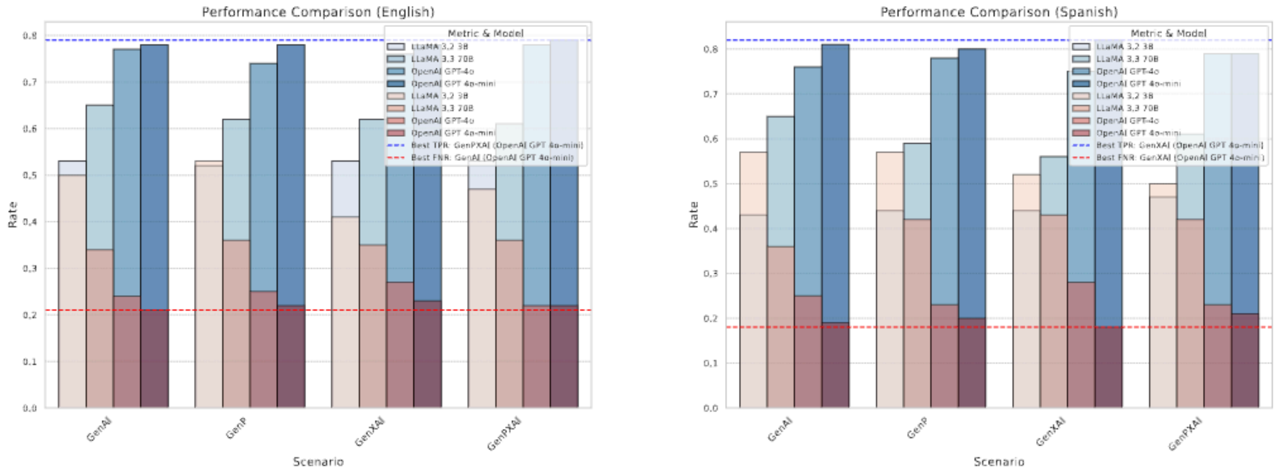




(a) GenAI (b) GenP

(c) GenXAI (d) GenPXAI

Figure 10: Confusion matrices for all scenarios for GPT 4o mini.



(a) English Subset (b) Spanish Subset

Figure 11: Comapring True Positive Rate (TPR) (equivalent to Recall) and False Negative Rate (FNR) all models and senarious.

- **Model Performance: OpenAI GPT-4o** and **GPT-4o-mini** perform best, while **LLaMA 3.2 3B** performs worst, with **LLaMA 3.3 70B** falling in between.

- **Key Takeaways: Smaller models benefit more from XAI (GenXAI)**, while larger models need **persona (GenPXAI)** to offset potential performance drops;

Utrecht University

Future research & Next Steps

- **Refining Persona Design:**

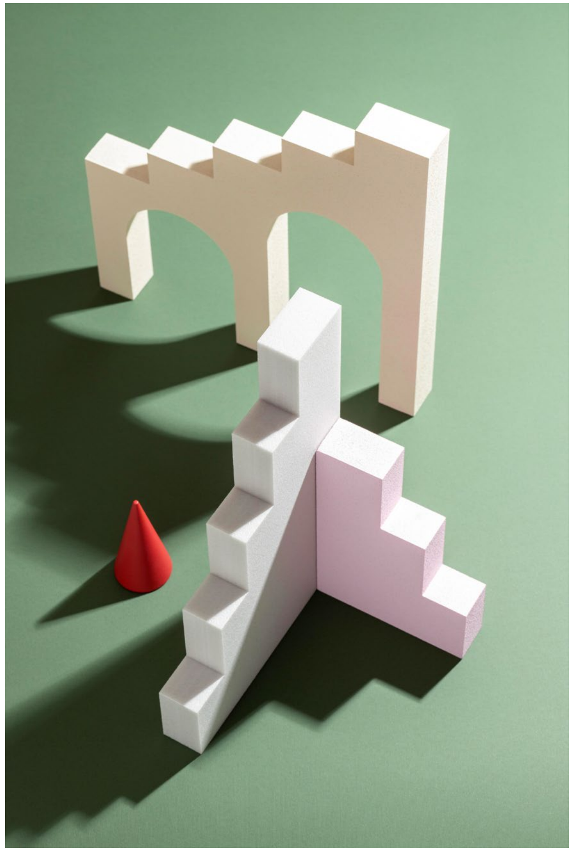
Improve **persona descriptions** to better align with cultural and linguistic contexts, reducing potential biases in GenAI models.

- **More XAI Techniques:**

Exploring **domain-specific explainability (XAI) methods.**

- **Expanding Language Coverage:**

Studying **more languages and dialects.**



Discussion – Questions & Suggestions?

Any Questions?

More Suggestions?

Utrecht University

Thank You!

For further questions or details, please contact:

h.mohammadi@uu.nl