This study looks at a very important but often overlooked problem in online spaces: how to tell the difference between **sexist speech** and **anti-sexist speech**. While many people use social media to attack or insult women—especially female politicians—others use the same platforms to **resist and call out sexism**. The challenge is that anti-sexist responses often use strong or emotional language that can look very similar to the sexist comments they are criticizing. Automated systems, such as those powered by **large language models (LLMs)**, frequently confuse the two. This means that people speaking up against sexism may have their voices wrongly flagged or silenced, while harmful speech still circulates. To study this problem, we collected tweets directed at **female Members of Parliament in the UK** during key political events in 2022. These events were moments when online abuse was especially likely to appear, such as controversies or leadership changes (e.g. Angela Rayner's Basic Instinct reference (month: April); leadership transitions (months: September, October)). We then asked five different LLMs to classify the tweets into three categories: *sexist, anti-sexist, or neither*. We experimented with different prompt styles (like zero-shot and few-shot) and measured not only the models' accuracy but also their **confidence and uncertainty** in their answers. We also compared their outputs with expert human annotations, using a method that preserved disagreements among annotators instead of simply forcing consensus.

Our findings show that **LLMs often misclassify anti-sexist speech as sexist** (as seen in Table S1, Figure 1)**,** especially during heated political moments. This is because the language of resistance often mirrors the tone and phrasing of harmful speech. In addition, the models were usually **overconfident** even when wrong, which is risky for content moderation systems. The novelty of this work lies in treating **anti-sexist speech as a separate category** rather than ignoring it or folding it into existing labels. By doing so, we highlight the risks of current moderation systems, which may unintentionally silence those resisting sexism and reinforce existing inequalities. We argue that moderation tools must move beyond simple harmful/not-harmful categories, include **examples of counter-speech in training data**, and integrate **human review during sensitive political events**.

Labels per month	February	March	April	May	September	October	**Total**
Sexism	10	4	47	1	29	21	112
Anti-sexism	8	1	60	2	22	15	108
Neither	61	19	351	24	302	366	1123
Total	79	24	458	27	353	402	**1343**

Table S1. Data statistics showing the total number of instances for each label over the months.

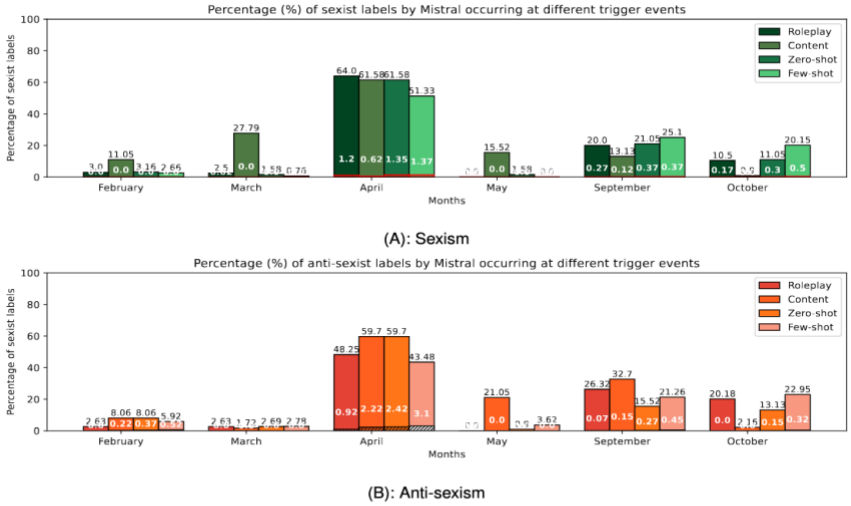

Figure 1. This figure shows the proportion of tweets predicted as sexist (A) and anti-sexist (B) by the Mistral model across six major trigger-event months in 2022. Bars are grouped by prompt type: roleplay, content, zero-shot, and few-shot. Shaded segments indicate the proportion of correctly classified instances. While Mistral frequently overpredicts both categories (particularly around high-salience events like April) its accuracy remains low, especially for anti-sexist speech. These patterns reveal the model's difficulty in distinguishing between harmful and resistant language when both share similar tone and phrasing.

This research contributes both technically and socially: it shows the **limits of current AI moderation systems** and offers insights into how we can design fairer systems that protect democratic participation and ensure that voices challenging sexism are heard rather than suppressed.