Large Language Models with Reinforcement Learning from Human Feedback Approach for Enhancing Explainable Sexism Detection

Ali Riahi Samani and Tianhao Wang and Kangshuo Li and Feng Chen*
Department of Computer Science
The University of Texas at Dallas, Texas, USA

Abstract

Recent advancements in natural language processing, driven by Large Language Models (LLMs), have significantly improved text comprehension, enabling these models to handle complex tasks with greater efficiency. A key feature of LLMs is their ability to engage in contextual learning, which allows them to understand and apply instructions given in natural language to new scenarios without requiring additional training. This capability is particularly valuable in social media, where LLMs can be crucial in addressing challenges in explainable sexism detection. We hypothesize that by leveraging contextual learning capabilities, LLMs can provide clear, explainable insights into why certain content is flagged as problematic, thus enhancing transparency in the sexism detection process. To this end, we propose a Reinforcement Learning from Human Feedback (RLHF) based fine-tuning framework for sexism detection. We studied two well-known LLMs, Mistral-7B and LLaMA-3-8B, in zero-shot, supervised fine-tuning, and RLHF scenarios to conclude the superior ability of LLMs in sexism detection. The experimental results reported in this work, based on three tasks of Explainable Detection of Online Sexism (EDOS), highlight the importance of RLHF for building explainable systems in online discourse. Furthermore, we found that the LLaMA-3-8B model achieves the best results using the RLHF approach, scoring 0.8681 on Task A (binary sexism detection), 0.6829 on Task B (category classification of sexism), and 0.4722 on Task C (fine-grained sexism vectors) test sets.

1 Introduction

Online platforms have become essential in our daily lives, enabling communication and information sharing. Social media networks allow global interaction (Boyd and Ellison, 2007). While beneficial, these platforms also introduce challenges, such as sexism (Fox et al., 2015) and gender-based violence (RUSSO and PIRLOTT, 2006). Sexism, a form of discrimination (Jun, 2024), has become increasingly problematic, especially with the abuse women face online (Maeve Duggan, 2017). Addressing this requires identifying sexist content to improve interaction quality.

Detecting sexism involves recognizing offensive language and gender biases. Transformer models have excelled in natural language processing (NLP) tasks like sarcasm detection (Mishra et al., 2020; Yin and Zubiaga, 2021), but classifying sexism remains complex (Magnossao de Paula et al., 2021). Compared to traditional models like SVMs (Walawalkar et al., 2002), CNNs (Gambäck and Sikdar, 2017), and LSTMs (Mut Altın et al., 2020), language models (Mohammadi et al., 2023) show better results in understanding context and semantics, crucial for detecting sexism.

However, automated tools often lack transparency in explaining why content is flagged (Kirk et al., 2023). Explaining why flagged content is sexist is essential for creating fair, inclusive, and transparent online spaces. It helps platforms comply with legal standards, build user trust, and reduce biases. By explaining why content is flagged as sexist, sexism detection systems facilitate efficient content moderation, educate users, and inform data-driven policy-making, ultimately fostering a more respectful online community. Consequently, the Explainable Detection of Online Sexism (EDOS) task, as introduced by (Kirk et al., 2023), advances the creation of accurate and interpretable methods for identifying and classifying sexist content in English. It includes fine-grained classifications for content collected from two large social media platforms like Gab and Reddit. The EDOS task is divided into three hierarchical subtasks: (1) Task A, focusing on the binary detection of sexist content; (2) Task B, categorizing the detected sexism into four specific types; and (3) Task C, providing

*Corresponding author: feng.chen@utdallas.edu

Proceedings of the 31st International Conference on Computational Linguistics, pages 6230–6243
January 19–24, 2025.

a detailed analysis by identifying 11 distinct forms of sexist content.

Large Language Models (LLMs) have shown significant potential in various NLP tasks. In sexism detection, LLMs are capable of analyzing large volumes of text data from various platforms, and they are capable of identifying subtle and overt forms of sexist language through fine-tuning using specialized datasets (Rhue et al., 2024). Their advanced language understanding allows them to not only detect problematic content but also provide explanations for their decisions, which is crucial for transparency and trust. LLMs can be particularly useful for the EDOS tasks, as they can handle hierarchical classification structures (Zhang et al., 2024). By leveraging LLMs, they can continuously learn and stay updated with evolving language trends (Wu et al., 2024), maintaining accuracy and relevance in detecting and explaining online sexism. This makes LLMs a powerful asset in developing more ethical and transparent automated content moderation systems.

Despite numerous studies utilizing transformer models, such as BERT variants (Mohammadi et al., 2023) for sexism detection, the exploration of LLMs for explainable applications in sexism detection remains underexplored, especially considering the recent advancements and trends in NLP brought about by the emergence of various LLMs. To address this gap, in this study, our objective aims to explore the application of LLMs for an explainable framework for sexism detection using the Reinforcement Learning with Human Feedback (RLHF) (Bai et al., 2022) framework. This study focuses on two prominent LLMs, namely Mistral (Jiang et al., 2023) and LLaMA-3 (Touvron et al., 2023), to assess their effectiveness in this context. We developed a comprehensive sexism detection framework through a two-step process: (1) supervised fine-tuning (SFT), which involves training the models on a labeled dataset to recognize sexist content, and (2) implementing RLHF, which enables the LLMs to refine their understanding and responses based on human feedback. This iterative process allows the models to better align with human preferences (Ouyang et al., 2022) and ethical considerations, ultimately enhancing their accuracy and explainability in identifying and categorizing sexist language. By integrating these methodologies, we aim to create a robust system capable of effectively detecting and explaining sexist content online.

In summary, our contributions are as follows:

- First, we implemented a traditional model to assess the progress made in sexism detection and its potential for providing explainable systems in this domain.
- Second, we evaluated the zero-shot performance of LLMs across the three tasks of EDOS to determine their capability in identifying fine-grained categories of sexism.
- Third, we propose a Parameter-Efficient Fine-Tuning (PEFT) (Han et al., 2024) mechanism for LLMs using Parameter-efficient finetuning to trigger the language understanding capability of LLMs for sexism detection.
- Lastly, we introduced RLHF on top of the supervised fine-tuned LLM to improve the model's ability to distinguish between overlapping categories of sexism, thereby enhancing the quality of explanations provided.

We have implemented our work publicly available for the research community. The code and resources can be accessed at our GitHub repository at https://github.com/aliriahi90/RL_LLM_EDOS.

The rest of this paper is organized as follows: In the section 2, we briefly discuss related works. In section 3 we describe the methodology used in our study. Results and analysis are presented in section 4. Finally, discussions and future works are in section 5, and conclusions are given in section 6.

2 Related Work

Sexism detection on social media has been explored through deep learning methods like LSTMs, CNNs, and transformer models such as BERT and DistilBERT (Devlin et al., 2019; Sanh et al., 2020). These models, used in datasets like EXIST (Rodríguez-Sánchez, Francisco et al., 2021), have shown strong performance, though challenges persist in classifying subjective social media content (Kalra and Zubiaga, 2021). Multilingual sexism detection has also gained attention, with BERT applied to Spanish and English tweets (Mozafari et al., 2020), and the creation of a Chinese sexism dataset expanding research beyond English (Jiang et al., 2022). Ensemble models and fine-tuning techniques, such as Majority Voting and transformer ensembles like RoBERTa (Liu et al., 2019), have further improved classification performance (de Paula and da Silva, 2022).

A recent study of (Kibriya et al., 2024) presents a deep learning model for detecting hate speech by employing advanced NLP techniques and explainable AI tools like SHAP and LIME to enhance interpretability. Another study by (Pan et al., 2024) explores LLMs with different learning strategies, finding that fine-tuning with the Zephyr model (Tunstall et al., 2024) significantly outperforms other methods in detecting sexist content, although it struggles with false positives. In the realm of social media, particularly football discourse, (Santos et al., 2024) demonstrates how fine-tuned BERT-based models effectively classify racist content, emphasizing the importance of context-sensitive training and transparency through explainable AI methods. (Azadi et al., 2024) improves sexism classification in bilingual contexts by fine-tuning XLM-RoBERTa and leveraging GPT-3.5's (OpenAI, 2024) few-shot learning, illustrating the models' ability to handle complex linguistic variations. Work of (Sultana and Begum Kali, 2024) on ChatGPT highlights its potential in identifying sexist remarks in software development communications, showing LLMs promising results in detecting specific sexist behaviors but suggesting that further refinement is needed to capture contextual clarity effectively.

3 Methodology

In this study, we developed a Reinforcement Learning (RL) framework using the Direct Policy Optimization (DPO) (Rafailov et al., 2024) policy optimizer for explainable sexism detection. Our methodology consists of five key modules (as visualized in Figure 1). In the following sections, we will delve into the details of each step, providing an overview of the processes.

3.1 Explainable Detection of Online Sexism - Taxonomy and Dataset

The taxonomy for explainable detection of online sexism (EDOS) developed by (Kirk et al., 2023) was created using a grounded theory approach (Glaser and Strauss, 1999), incorporating empirical entries to refine the schema. Our study is based on the three subtasks designed by EDOS, aimed at training and evaluating our framework for the fine-grained and explainable detection of online sexism. The EDOS dataset's subtasks categorize sexist content into three hierarchical tasks, as outlined in Appendix A, Table **??**.

The train, validation, and test dataset statistics for EDOS are presented in Appendix A, Table **??**. The EDOS dataset is inherently imbalanced, reflecting real-world scenarios where certain types of sexist content are more prevalent than others. Despite this imbalance, the portions for each class in the train, test, and validation datasets remain consistent. This consistency is crucial for developing robust models, as it is crucial to handle skewed class distributions effectively. Addressing this imbalance is essential for ensuring the model's performance is reliable and generalizable across different types of sexist content.

3.2 Data Initialization

In the first step of our methodology, we focused on data initialization using the EDOS task datasets. This dataset consists of input data x and corresponding ground truth labels y, derived from label sets C. To structure the data for effective model training, we utilized prompt templates P for EDOS tasks. For Tasks A, B, and C, each prompt template P was designed to align with the unique requirements of each task, ensuring that the input format effectively guided the model toward learning task-specific patterns for better identification. For tasks, we designed several prompt templates, which are presented in Appendix B. Each prompt consists of `instruction` on tasks by referring to the task definitions, `context`, which lists possible classes to identify for a given post, and `{POST}` as an input x to find a response $y_{generated}$ for using LLM generator function $LLM(x') \rightarrow y_{generated}$, where $x' = P(x)$.

In general, the data initialization can be defined as follows: $Init(\mathcal{D}(x, y), P) \rightarrow \{(x'_i, y_i)\}_{i=1}^{N}$, where, N is size of dataset, $x' = P_A(x)$ for task A (1), $x' = P_B(x)$ for task B (2), and $x' = P_C(x)$ for task C (3). Later the obtained, $\mathcal{D} = \{(x'_i, y_i)\}_{i=1}^{N}$ used for fine-tuning LLMs.

3.3 Supervised Fine-Tuning

Supervised fine-tuning adapts pre-trained LLMs to a specific task by training it on a labeled dataset. Let $\mathcal{D} = \{(x'_i, y_i)\}_{i=1}^{N}$ denote the labeled dataset, where x'_i represents an input example and y_i denotes its corresponding label. The goal is to adjust the model parameters θ to minimize the loss function $\mathcal{L}$ on this dataset:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(f(x'_i; \theta), y_i)$$

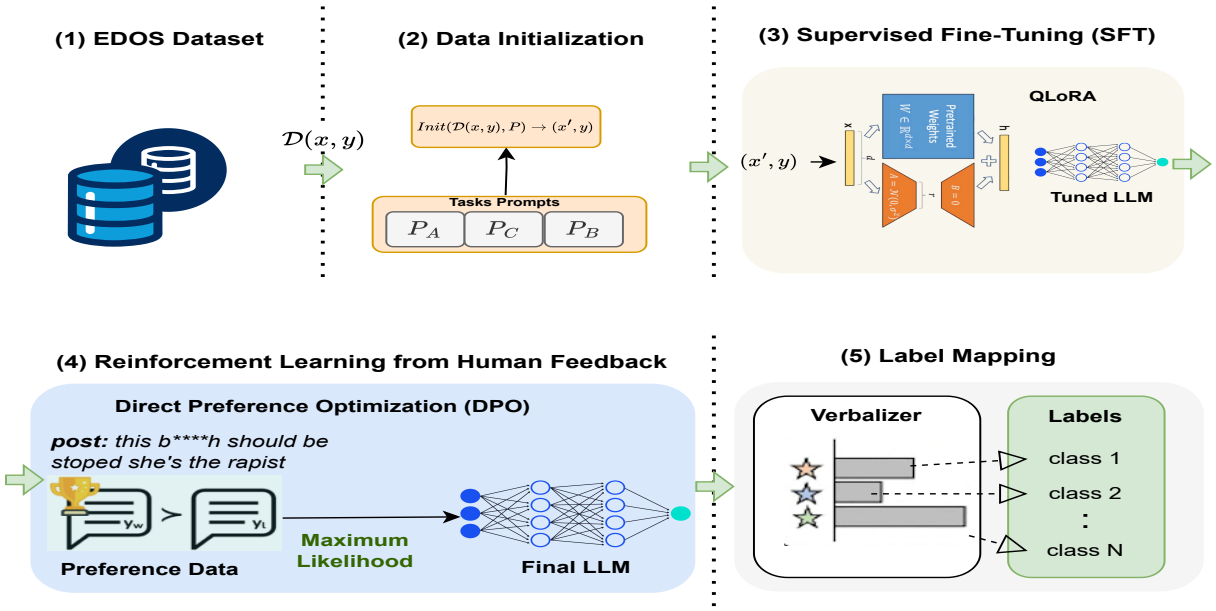

Figure 1: RLHF framework for fine-tuning LLMs for explainable sexism detection.

where $f(x_i'; \theta)$ is the model's prediction for input x_i' given parameters θ, and $\mathcal{L}$ is the loss function measuring the difference between the predicted and actual labels. To enhance the efficiency of fine-tuning, we employ Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2024), which optimizes the process by combining two techniques: (1) Quantization, and (2) Low-Rank Adaptation (Hu et al., 2022).

Quantization reduces the precision of the model's weights from floating-point numbers to lower-bit representations, which in our case we set to `4-bit`. If W represents the weight matrix of the model, quantization maps W to a quantized weight matrix $\tilde{W}$ such that: $\tilde{W} = \text{Quantize}(W)$, where Quantize(.) denotes the quantization function that reduces the bit-width of W. This step reduces memory usage and computational cost.

Low-rank adaptation approximates the updates needed during fine-tuning using a low-rank decomposition. By focusing on a low-rank approximation, the computational cost is significantly reduced compared to full-rank updates. Let ΔW be the update matrix for the weights. Instead of updating the entire weight matrix, low-rank adaptation represents ΔW as a product of two smaller matrices A and B, where $\Delta W \approx AB^T$. The updated weights are: $W_{\text{new}} = W + \Delta W = W + AB^T$.

This combination of quantization and low-rank adaptation allows us to leverage the full power of LLMs while keeping the fine-tuning process computationally feasible making it suitable for sexism detection, by the reduction in computational cost and memory usage. It also allows for more efficient processing of large datasets that are often required for training robust sexism detection models, whereas in our case Task A training dataset may require many computations resources without quantization. Moreover, QLoRA allows for rapid iteration and improvement of the model. This is especially important in dynamic environments where the nature of online discourse evolves, and models need to be updated frequently to maintain their accuracy in detecting sexist content.

3.4 Reinforcement Learning From Human Feedback

We used DPO (Rafailov et al., 2024) in RLHF, which is a method for training LLMs that focuses on directly optimizing for human preferences. We considered $y_{generated}$ as outputs generated by LLMs, and humans provide feedback as the y_{truth} as a human preference. This feedback is used to train a preference model, which learns to predict human choices. Instead of using a complex reward system, DPO uses this preference model to guide the LLM's learning process, ensuring that the LLMs produce outputs that align with human preferences.

For $y_{truth} \in C$, let's consider C_i as a ground truth label, we considered the rest of C_k labels, where $k \neq i$ is a false label to the DPO. It allows DPO to optimize the policy to differentiate between different classes of sexist content. While it is sim-

ple, the novelty is there we have multiple classes that have overlaps in their definitions, so by using this optimization objective, the policy provides LLMs with further learning's to make differences within the different levels of sexism taxonomy. Let's consider two outputs, $y^i_{generated} \in C_i$ and $y^k_{generated} \in C_{k \neq i}$. If a human prefers $y^i_{generated} \in C_i$ (ground truth) over $y^k_{generated} \in C_{k \neq i}$ (another category of sexism as a false generation), we can represent this preference as a binary outcome, such as $Probability(y^i_{generated} > y^k_{generated}) = 1$. The preference model is trained to predict the correct category of sexism, and during optimization, the LLM aims to generate outputs that maximize these generated preferences.

In DPO, the objective is to maximize the sum of generated preferences across all output pairs, allowing LLMs to better distinguish the different categories of sexisms in different levels of taxonomy, aligning the LLM's behavior with human expectations without needing an intermediate reward function.

3.5 Label Mapping

Label mapping or answer set mapping (Liu et al., 2023) is an important step in transforming the human-readable responses generated by LLMs into discrete, actionable labels for determining the category of sexism in a post. This process is designed to ensure that the generated text responses align accurately with predefined label sets C per task. Initially, responses are generated for the training, validation, and test sets for the tasks. These responses are typically explanatory in nature, providing insights into why a particular post might be categorized under a specific type of sexism.

To systematically extract predicted labels C from generated texts $y_{generated}$, we employ the label mapping technique $L : y_{generated} \rightarrow y_{predict}$ where $y_{predict} \in C$. The process begins by using Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) to obtain embeddings of the generated responses for train $E_{train} = SBERT(y^{train}_{generated})$, validation $E_{validation} = SBERT(y^{validation}_{generated})$, and test sets $E_{test} = SBERT(y^{test}_{generated})$. Next, using obtained embeddings, we train a logistic regression LR model using the training set embeddings E_{train}. The LR model is tasked with learning the relationship between the embeddings and the corresponding predefined labels $y_{truth} \in C$. Once the model is trained, it can predict labels for the embeddings of the generated responses in the validation and test sets using $L(y_{generated}) = LR(SBERT(y_{generated}))$, thereby converting the model's generations into the preferred labels for the tasks.

SBERT is selected for its ability to produce high-quality sentence embeddings that capture semantic similarities between texts. These embeddings serve as feature vectors that represent the generated texts in a numerical form.

4 Results

In this section, we will provide a detailed discussion of the experimental setups employed, conduct an in-depth analysis of the training processes, and present the results obtained from our experiments.

4.1 Experimental Setups

4.1.1 Models

Baselines. We employed four models as baseline comparisons for our methodologies. These baseline models utilize traditional text representation techniques combined with classical machine learning classifiers, providing a foundational performance benchmark against which we can measure the effectiveness of our advanced approaches. The models are TFIDF + AdaBoost, RoBERTa + SVM, DeBERTa + LDA, SetFit (Tunstall et al., 2022). Moreover, we used fine-tuned DeBERTa-v3-baseline (He et al., 2023), which is the best performer baseline of EDOS (Kirk et al., 2023) for comparison.

EDOS Systems. The final EDOS tasks (Kirk et al., 2023) at SemEval-2023 encompass a comprehensive range of approaches, with 84 methods addressing Task A, 69 tackling Task B, and 63 dedicated to Task C. For our study on sexism detection using LLMs, we consider the statistical summary of tasks and the top two approaches namely PingAn-LifeInsurance (Zhou, 2023) and FiRC-NLP (Hassan et al., 2023). These scores provide a general point of comparison, allowing us to evaluate the performance of our LLM-based methods against established standards within the field.

Proposed Methods. Our experiments focus on evaluating two well-known large language models: Mistral-7B (Jiang et al., 2023) and LLaMA-3-8B (Touvron et al., 2023), specifically the instruct variants (`LLaMA-3-8B-Instruct` and `Mistral-7B-Instruct`). For each of these models, we conducted three distinct experimental se-

tups:

- **Zero-Shot**: This involves using the models without any fine-tuning.
- **SFT**: supervised fine-tuning, where the models are fine-tuned on task-specific data.
- **SFT + RLHF**: SFT combined with RLHF, further refines the models based on additional human-guided learning.

4.1.2 Fine-tuning Details

For experimentation, we used a set of default parameters and manually adapted a few parameters across different models and tasks to optimize performance and efficiency in fine-tunings. Parameters for SetFit, SFT, and RLHF models are presented as follows:

SetFit. The SetFit method utilizes the `AdamW` optimizer and processes 40 randomly chosen shots to balance computational cost and performance, with a batch size of 16. The transformer model is trained for 5 epochs, reflecting the large sample size during training, while the classifier undergoes 15 epochs to ensure effective fine-tuning given its lower resource requirements. The maximum token length is set to 512 tokens, a standard for transformer models like BERT, accommodating most input sequences without truncation.

SFT. In the SFT setups for Mistral-7B, Task A uses QLoRA parameters ($r = 8$, `4-bit` quantization), the `AdamW` optimizer with a learning rate of 2×10^{-3} (determined through manual experimentation), 5 epochs, a batch size of 5, and gradient accumulation steps of 8, chosen to handle the large dataset and available computational resources. For Tasks B and C, the learning rate is reduced to 2×10^{-4} to prevent memorization. Task B trains for 30 epochs with a batch size of 4, minimizing gradient drift, while Task C uses the same settings but includes a QLoRA dropout rate of 0.05 to enhance precision on smaller sample sizes. LLaMA-3-8B follows similar configurations with adjustments: Task A uses a learning rate of 2×10^{-4}, 10 epochs, a batch size of 6, and gradient accumulation of 4. Task B employs 10 epochs with a batch size of 4, and Task C uses a longer text length (1024 tokens), 30 epochs, and a batch size of 2 to process task-specific requirements.

RLHF. With RLHF, both fine-tuned Mistral-7B and LLaMA-3-8B models adopt QLoRA parameters with $r = 8$ and quantization at `4-bit`, adjusting the alpha to 16 and a dropout rate of 0.1. This setup involves a `AdamW` optimizer with a learning rate of $5 * 10^{-4}$, 16 epochs, a batch size of 5, and gradient accumulation steps of 8, maintaining a maximum text length of 512 tokens.

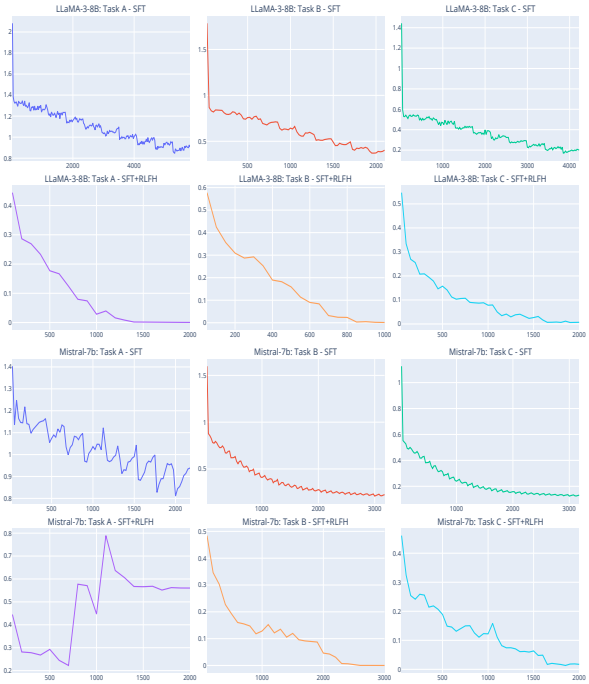

Figure 2: Training analysis of LLaMA-3-8B and Mistral-7B on SFT and RLHF for tasks.

4.2 Training Analysis

In this section, we present the loss behavior observed across SFT and SFT+RLHF fine-tuning for both Mistral and LLaMA-3 LLMs across three tasks of sexism detection. Different loss behaviors are presented in Figure 2. As we can see, during SFT, the LLaMA-3 model exhibited a general trend of decreasing loss across all three tasks, though the loss trajectory showed fluctuations. Notably, Task A demonstrated significant variability in loss reduction, highlighting the complexity and challenge this task poses for LLMs. These fluctuations were even more evident in the Mistral model for Task A, emphasizing the difficulty that LLMs face in distinguishing the simplest category of sexism, even during fine-tuning.

In contrast to SFT, RLHF fine-tuning showed a more stable and consistent reduction in loss for the LLaMA-3 across all tasks. This indicates an improvement in the model's ability to distinguish various categories of sexist content effectively. While the Mistral model did not perform as well on Task A during RLHF, it displayed good behavior by maintaining a steady decrease in loss, suggesting

that continued training could enhance its performance. Overall, these observations highlight the challenges associated with fine-tuning LLMs for sexism detection, particularly for tasks that require a slight understanding of the domain. However, they also highlight the potential of RLHF to improve model performance in distinguishing complex content categories during the learning process for both LLMs that we studied in this work.

4.3 Results Analysis

Our study aimed to evaluate the effectiveness of LLMs in explainable sexism detection by comparing traditional models and two state-of-the-art LLMs, Mistral-7B and LLaMA-3-8B, across zero-shot, SFT, and SFT + RLHF models. The results from our experiments on validation and test sets, measured using F1-Macro scores are presented in Table **??**. It reveals insightful patterns in the performance of various models and fine-tuning techniques. By analyzing the performances, we can draw conclusions about how different approaches handle the challenges posed by EDOS tasks.

4.3.1 Task A – Binary Sexism Detection

This task focuses on distinguishing between sexist and non-sexist posts. Traditional models like TFIDF + AdaBoost, RoBERTa + SVM, DeBERTa + LDA, and DeBERTa-v3-base (this baseline introduced in EDOS by Kirk et al. (2023)) offer reasonable baseline performances. Among these, DeBERTa-v3-base stands out as the best, showing its strength in binary classification—even advanced techniques like few-shot learning with SetFit struggle to outperform the base transformer model. However, the zero-shot performance of both Mistral-7B and LLaMA-3-8B is notably low, which suggests that these models require substantial task-specific fine-tuning to handle sexism detection effectively. SFT brings significant improvements, where Mistral-7B achieves an F1-score of 0.8220, slightly surpassing LLaMA-3-8B at 0.8156. These scores outperform most systems from SemEval-2023 (Kirk et al., 2023) and defeat the 50% of approaches developed for the task, nonetheless still there is room for improvement. Moreover, both LLMs significantly boosted their performance after seeing examples related to sexist content during fine-tuning. These findings, recommend that LLMs without fine-tuning are not helpful for sexism detection, and task-specific fine-tuning is required to be useful in building explainable systems. RLHF further refines performance. LLaMA-3-8B reaches an F1-score of 0.8603, marking the highest result. RLHF not only improves performance but stabilizes model behavior, making it more robust, even boosting Mistral-7B by 1%. This improvement recommends that preference-based learning within RLHF can lead us to obtain more robust systems for explainable systems.

4.3.2 Task B – Category of Sexism

This task involves classifying sexist posts into one of four categories, making it more complex than Task A. Traditional models underperform compared to advanced LLMs, with lower F1-Macro scores indicating their limited ability to differentiate between sexism categories. Zero-shot models like Mistral-7B and LLaMA-3-8B also struggle without fine-tuning. SFT greatly improves both models' classification abilities, with LLaMA-3-8B outperforming Mistral-7B. RLHF further boosts F1-Macro scores significantly, improving Mistral-7B by 9% and LLaMA-3-8B by 7%. This suggests that RLHF can be more useful and helps the models to better understand and differentiate between the higher levels of sexism taxonomy.

4.3.3 Task C – Fine-Grained Vectors of Sexism

This task involves classifying posts into 11 specific categories, making it the most complex. Traditional models show lower F1-Macro scores, revealing their limitations in handling such detailed classifications. However, the only exception here is RoBERTa + SVM which surprisingly is better than our SFT+RLHF-based fine-tuned LLMs by approximately 1%. Zero-shot results are notably poor, reflecting the models' insufficient ability to handle fine-grained distinctions without fine-tuning. Both Mistral-7B and LLaMA-3-8B demonstrate considerable improvement with SFT, with LLaMA-3-8B achieving the highest scores. RLHF further boosts F1-Macro scores, particularly for LLaMA-3-8B, highlighting its ability to handle complex classifications and make precise distinctions across multiple categories.

4.3.4 Comparison with EDOS Systems

The results presented in Table **??** demonstrate a clear gap between the performance of the proposed methods and the top-performing systems from the EDOS competition, namely PingAnLifeInsurance (Zhou, 2023) and FiRC-NLP (Hassan et al.,

Model	Validation			Test		
	Task A	**Task B**	**Task C**	**Task A**	**Task B**	**Task C**
Baselines						
TFIDF + AdaBoost	0.8067	0.4608	0.2346	0.8216	0.4253	0.2492
RoBERTa + SVM	**0.8220**	**0.6364**	**0.4564**	0.8207	**0.6236**	**0.4862**
DeBERTa + LDA	0.7697	0.5286	0.3617	0.7796	0.5106	0.3375
SetFit	0.5819	0.4415	0.2498	0.6165	0.4176	0.2797
DeBERTa-v3-base (He et al., 2023)	-	-	-	**0.8235**	0.5926	0.3171
EDOS Systems						
Top Performers						
PingAnLifeInsurance (Zhou, 2023)	-	-	-	**0.8746**	**0.7212**	**0.5605**
FiRC-NLP (Hassan et al., 2023)	0.8684	0.7591	0.6129	0.8740	0.7058	0.5404
Statistical Summary of EDOS Systems (Kirk et al., 2023)						
System count	-	-	-	84	69	63
Q1	-	-	-	0.7994	0.5730	0.3153
Mean	-	-	-	0.8095	0.5899	0.3829
Median	-	-	-	0.8322	0.6191	0.4230
Q3	-	-	-	0.8537	0.6501	0.4758
Proposed Methods						
Mistral-7B LLM Performance						
Zero-Shot	0.4978	0.3020	0.1452	0.4894	0.3367	0.1430
SFT	0.8049	0.6345	0.3589	0.8220	0.5849	0.3783
SFT + RLHF	0.8479	0.6565	0.3549	0.8368	0.6719	0.3940
LLaMA-3-8B LLM Performance						
Zero-Shot	0.5189	0.3339	0.1107	0.5085	0.2879	0.1406
SFT	0.8089	0.6224	0.4431	0.8156	0.6148	0.4449
SFT + RLHF	**0.8592**	**0.6998**	**0.5392**	**0.8603**	**0.6829**	**0.4722**

Table 1: Results on validation and test sets using F1-Macro evaluation metric. The best results for *Baselines*, *EDOS Systems*, and *Proposed Methods* are in bold. The summary count presents the number of participants in the EDOS and the Q1, mean, median, and Q3 statistical summary are based on participant results for tasks.

2023). Both top-performing systems leverage sophisticated techniques that significantly enhance their effectiveness in fine-grained classification tasks. The PingAnLifeInsurance (Zhou, 2023) approach employed a multitask learning framework in combination with pretraining on two million unlabeled samples, using large transformer models like RoBERTa-large and DeBERTa-v3-large. This pretraining, combined with task-specific fine-tuning and multitask learning, yielded state-of-the-art results, achieving an F1-score of 0.8746 on sub-task A and competitive rankings on sub-tasks B and C. Similarly, the FiRC-NLP (Hassan et al., 2023) system employed an ensemble of fine-tuned DeBERTa variants, leveraging k-fold cross-validation for robust training.

In comparison, the proposed methods based on Mistral-7B and LLaMA-3-8B achieve reasonable performance but fall short, particularly in tasks requiring fine-grained classification. For instance, while LLaMA-3-8B SFT+RLHF shows competitive performance, achieving 0.8603 for Task A, 0.6829 for Task B, and 0.4722 for Task C on the test set, these results are notably below the performance achieved by PingAnLifeInsurance and FiRC-NLP. The significant difference in performance can be attributed to several factors: 1) *Pretraining and Domain Adaptation*: The top systems employed extensive pretraining on domain-specific unlabeled data, a step that is absent in the proposed methods. 2) *Multitask Learning and Ensembles*: The multitask-learning framework and ensemble strategies used by top performers likely contributed to their models' robustness and ability to generalize across subtasks. 3) *Fine-Tuning Strategies*: The fine-tuning strategies, including k-fold cross-validation and careful model selection, allowed top performers but this requires a high computation.

The proposed methods, while utilizing RLHF and SFT, currently lack the advanced techniques like multitask learning or domain-specific pretraining seen in the EDOS systems. However, the results demonstrate the promise of LLM-based methods, particularly in zero-shot and SFT+RLHF scenarios. To bridge the performance gap, future iterations of the proposed methods could benefit from incorporating multitask learning objectives, domain-adapted pretraining, and ensemble techniques to enhance their competitiveness.

4.3.5 Impact of SFT and RLHF

The improvements with SFT and RLHF highlight the models' ability to better understand and perform tasks through fine-tuning and alignment with human preferences. The sharp performance gains from zero-shot to SFT and further with RLHF indicate that task-specific training and iterative feedback significantly enhance their capacity to recognize sexism categories. Fluctuations in loss behavior, particularly in Mistral-7B (as observed in Figure 2), reflect the challenges of distinguishing fine-grained forms of sexism, but both SFT and RLHF help stabilize and improve accuracy. LLaMA-3-8B consistently outperforms across tasks, demonstrating its robustness in handling complex classifications. These results underscore the effectiveness of combining SFT with RLHF to achieve superior performance in explainable sexism detection.

Moreover, the study of Pan et al. (2024) demonstrated the effectiveness of fine-tuning backbone LLMs with a classifier head for hate speech detection tasks, achieving state-of-the-art performance on benchmarks like EDOS and HatEval. However, our work takes a different approach by employing instruction tuning using QLoRA-based adapters, which, as highlighted by Biderman et al. (2024), theoretically suffer less from catastrophic forgetting compared to traditional fine-tuning methods. This distinction enables our models to retain their generalization capabilities while being fine-tuned for specific tasks, such as EDOS. Furthermore, the RLHF approach incorporated in our framework leverages human feedback to enhance transparency and performance, resulting in superior explainability and competitive accuracy across all EDOS tasks, particularly in fine-grained sexism vector classification. This approach demonstrates the advantages of adapter-based instruction tuning over conventional fine-tuning strategies in building robust and interpretable hate speech detection systems.

5 Limitations

While the presented models demonstrate significant improvements in sexism detection and classification tasks, there are several limitations that warrant further investigation.

Task Complexity. For more complex tasks, like fine-grained classifications (Task C), even the state-of-the-art models exhibit reduced accuracy. This suggests that the models struggle with capturing subtle nuances between different forms of sexism, highlighting a limitation in their ability to generalize across a broader spectrum of categories.

Computational Costs. Fine-tuning large language models, especially when combined with techniques like RLHF, requires substantial computational resources. This restricts the accessibility of these methods for wider use, particularly in resource-constrained environments.

Real-Time Online Use. The current models are not optimized for real-time or online detection of sexist content from social media or websites. Implementing the models for live content moderation would require addressing challenges such as real-time processing, scalability, and privacy concerns. Future research could focus on adapting the models for online, real-time applications to enhance their practical utility.

6 Conclusions

In conclusion, our study demonstrates the significant impact of fine-tuning and reinforcement learning from human feedback on the performance of LLMs in detecting various forms of sexism leading toward more effective explainable systems. Our results reveal that zero-shot LLMs initially struggle with sexism detection, underscoring the necessity for task-specific fine-tuning. Moreover, supervised fine-tuning substantially enhances model performance. However, the application of RLHF further refines these improvements, demonstrating its effectiveness in stabilizing the interpretation of sexist content. Notably, QLoRA-based fine-tuned LLaMA-3-8B, when combined with RLHF, achieves the best results across all tasks of explainable sexism detection, highlighting the model's robustness and adaptability.

References

AmirMohammad Azadi, Baktash Ansari, and Sina Zamani. 2024. Bilingual sexism classification: Fine-

tuned xlm-roberta and gpt-3.5 few-shot learning. *Preprint*, arXiv:2406.07287.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. 2024. Emergent and predictable memorization in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

danah m. Boyd and Nicole B. Ellison. 2007. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230.

Angel Felipe Magnossão de Paula and Roberto Fray da Silva. 2022. Detection and classification of sexism on social media using multiple languages, transformers, and ensemble models. In *IberLEF@ SEPLN*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Fox, Carlos Cruz, and Ji Young Lee. 2015. Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. *Computers in Human Behavior*, 52:436–442.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, Vancouver, BC, Canada. Association for Computational Linguistics.

Barney Glaser and Anselm Strauss. 1999. *Discovery of Grounded Theory: Strategies for Qualitative Research*, 1st edition. Routledge, New York. EBook published 15 July 2017.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Preprint*, arXiv:2403.14608.

Fadi Hassan, Abdessalam Bouchekif, and Walid Aransa. 2023. FiRC at SemEval-2023 task 10: Fine-grained classification of online sexism content using DeBERTa. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1824–1832, Toronto, Canada. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. Swsr: A chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27:100182.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Heesoon Jun. 2024. *Sexism*, pages 139–170. Springer Nature Switzerland, Cham.

Amikul Kalra and Arkaitz Zubiaga. 2021. Sexism identification in tweets and gabs using deep neural networks. *Preprint*, arXiv:2111.03612.

Hareem Kibriya, Ayesha Siddiqa, Wazir Zada Khan, and Muhammad Khurram Khan. 2024. Towards safer online communities: Deep learning and explainable ai for hate speech detection and classification. *Computers and Electrical Engineering*, 116:109153.

Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 task 10: Explainable detection of online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Maeve Duggan. 2017. Online Harassment 2017.

Angel Felipe Magnossao de Paula, Roberto Fray da Silva, and Ipek Baris Schlicht. 2021. Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, pages 356–373. CEUR Workshop.

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Tackling online abuse: A survey of automated abuse detection methods. *Preprint*, arXiv:1908.06024.

Hadi Mohammadi, Anastasia Giachanou, and Ayoub Bagheri. 2023. Towards robust online sexism detection: A multi-model approach with bert, xlm-roberta, and distilbert for exist 2023 tasks. In *CLEF (Working Notes)*, pages 1000–1011.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. A bert-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII*, pages 928–940, Cham. Springer International Publishing.

Lütfiye Seda Mut Altın, Alex Bravo, and Horacio Saggion. 2020. LaSTUS/TALN at TRAC - 2020 trolling, aggression and cyberbullying. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 83–86, Marseille, France. European Language Resources Association (ELRA).

OpenAI. 2024. Gpt-3.5. `https://www.openai.com/`.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Ronghao Pan, José Antonio García-Díaz, and Rafael Valencia-García. 2024. Comparing fine-tuning, zero and few-shot strategies with large language models in hate speech detection in english. *CMES - Computer Modeling in Engineering and Sciences*, 140(3):2849–2868.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Lauren Rhue, Sofie Goethals, and Arun Sundararajan. 2024. Evaluating llms for gender disparities in notable persons. *Preprint*, arXiv:2403.09148.

Rodríguez-Sánchez, Francisco, Carrillo-de-Albornoz, Jorge, Plaza, Laura, Gonzalo, Julio, Rosso, Paolo, Comet, Miriam, and Donoso Vázquez, Trinidad. 2021. Overview of EXIST 2021: sexism Identification in Social neTworks.

NANCY FELIPE RUSSO and ANGELA PIRLOTT. 2006. Gender-based violence. *Annals of the New York Academy of Sciences*, 1087(1):178–205.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Guto Leoni Santos, Vitor Gaboardi dos Santos, Colm Kearns, Gary Sinclair, Jack Black, Mark Doidge, Thomas Fletcher, Dan Kilvington, Patricia Takako Endo, Katie Liston, and Theo Lynn. 2024. Kicking prejudice: Large language models for racism classification in soccer discourse on social media. In *Advanced Information Systems Engineering*, pages 547–562, Cham. Springer Nature Switzerland.

Sayma Sultana and Maimuna Begum Kali. 2024. Exploring chatgpt for identifying sexism in the communication of software developers. In *Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '24, page 400–403, New York, NY, USA. Association for Computing Machinery.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Lewis Tunstall, Edward Emanuel Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M Rush, and Thomas Wolf. 2024. Zephyr: Direct distillation of LM alignment. In *First Conference on Language Modeling*.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *Preprint*, arXiv:2209.11055.

L. Walawalkar, Mohammad Yeasin, Anand M. Narasimhamurthy, and Rajeev Sharma. 2002. Support vector learning for gender classification using audio and visual cues: A comparison. In *Pattern Recognition with Support Vector Machines*, pages 144–159, Berlin, Heidelberg. Springer Berlin Heidelberg.

Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. 2024. Continual learning for large language models: A survey. *Preprint*, arXiv:2402.01364.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Yunyi Zhang, Ruozhen Yang, Xueqiang Xu, Rui Li, Jinfeng Xiao, Jiaming Shen, and Jiawei Han. 2024. Teleclass: Taxonomy enrichment and llm-enhanced hierarchical text classification with minimal supervision. *Preprint*, arXiv:2403.00165.

Mengyuan Zhou. 2023. PingAnLifeInsurance at SemEval-2023 task 10: Using multi-task learning to better detect online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2188–2192, Toronto, Canada. Association for Computational Linguistics.

A EDOS Dataset Characteristics

Binary Task A	Category (Task B)	Vector (Task C)	Example
Sexist	Threats, plans to harm and incitement	Threats of harm	Maybe it's just me, but %women-name% sure looks white. Can we kill her, now?
		Incitement and encouragement of harm	Hi %women-name%, your site is f***g garbage. Please kill yourself.
	Derogation	Descriptive attacks	Women are awful women's advice is awful.
		Aggressive and emotive attacks	What a fucking c***t. Acting like she cares about her husband.
		Dehumanising attacks and overt sexual objectification	You just need to find a good old fashioned h***.
	Animosity	Causal use of gendered slurs, profanities, and insults	Okay well you have fun being a lonely miserable little b***h for the rest of your life.
		Immutable gender differences and gender stereotypes	Yes. Low SMV men are invisible to women.
		Backhanded gendered compliments	It certainly does not hurt, even if she is a 'butterface' it's a step closer to being more attractive.
		Condescending explanations or unwelcome advice	Ladies... never get tattoos.
	Prejudiced Discussion	Supporting mistreatment of individual women	If she ever reports a man for sexual harassment, it should be immediately disregarded.
		Supporting systemic discrimination against women as a group	Take women's rights away.
Not Sexist			The next person to play James Bond will be female. Watch it happen.

Table 2: EDOS taxonomy of sexism categories (Task B) and fine-grained vectors (Task C).

Task	Train	Validation	Test	$\|C\|$
TASK A - Binary Sexism Detection	14,000	2,000	2,000	2
TASK B - Category of Sexism	3,398	486	970	4
TASK C - Fine-Grained Vectors of Sexism	3,398	486	970	11

Table 3: EDOS dataset train, validation, and test set statistics. The $|C|$ represents the number of categories per task.

B Prompt Templates

```
Binary Sexism Detection: A two-class (or binary) classification where
systems have to predict whether a post is sexist or not sexist.

Given a post, determine whether the post belongs to a sexist or not
sexist class.
### Post: {POST}
### Class:
```

Listing 1: Prompt template P_A for Task A - Binary Sexism Detection

```
Category of Sexism: for posts which are sexist, classify them into one of
four categories:
1. threats, plans to harm and incitement
2. derogation
3. animosity
4. prejudiced discussion

Given a post, determine which class it belongs to.
### Post: {POST}
### Class:
```

Listing 2: Prompt template P_B for Task B - Category of Sexism Detection

```
Fine-grained Vector of Sexism: for posts which are sexist, classify them
into one of 11 categories:
1.1 threats of harm
1.2 incitement and encouragement of harm
2.1 descriptive attacks
2.2 aggressive and emotive attacks
2.3 dehumanising attacks and overt sexual objectification
3.1 casual use of gendered slurs, profanities, and insults
3.2 immutable gender differences and gender stereotypes
3.3 backhanded gendered compliments
3.4 condescending explanations or unwelcome advice
4.1 supporting mistreatment of individual women
4.2 supporting systemic discrimination against women as a group

Given a post, determine which class it belongs to.
### Post: {POST}
### Class:
```

Listing 3: Prompt template P_C for Task C - Fine-Grained Vector of Sexism Detection