PrefixGuard LLMs: Hate Speech and Sexism Detection

1st Shirin Shujaa
SSET, RMIT University
HCMC, Vietnam
0009-0007-7408-3848

2nd Ginel Dorleon
SSET, RMIT University
HCMC, Vietnam
0000-0003-2343-4445

***Abstract*—Hate speech and sexism detection on online platforms remains challenging due to their often subtle and context-dependent nature. Large Language Models (LLMs) offer powerful representations for this task, yet full fine-tuning is computationally expensive and may amplify biases. To address these limitations, we propose a parameter-efficient approach, PrefixGuard, based on prefix tuning. Rather than updating all model weights, we learn small trainable prefix vectors at each Transformer layer alongside a lightweight classification head, while keeping the LLM backbone frozen. We formalize our approach for classification and analyze how it influences internal representations of biased or offensive language. Experiments on three benchmark datasets, EDOS (sexism), OLID (offense), and HatEval (hate targeting women or immigrants), demonstrate that our prefix-tuned LLMs method consistently outperform BERT and RoBERTa baselines and achieve results competitive with full fine-tuning while training less than 1% of parameters. Further analysis shows that our approach yields balanced group-level performance on HatEval, robustness to adversarial text obfuscation, and improved calibration of predicted probabilities. These findings highlight PrefixGuard as a practical and effective alternative to full fine-tuning for hate speech and sexism detection in real-world moderation settings.**

I. Introduction

The expansion of online platforms has significantly increased the visibility and spread of harmful language, including hate speech and sexism. Identifying this content is essential for maintaining safe and inclusive digital environments, yet it remains a complex task. Hate and sexist language often appears in subtle and context-dependent forms, ranging from explicit insults to implicit stereotypes or indirect remarks. As a result, traditional NLP models frequently struggle, either failing to detect implicit abuse or incorrectly flagging neutral mentions of identity-related terms [1].

Large Language Models (LLMs) provide a strong foundation for this problem due to their broad linguistic knowledge and ability to capture contextual meaning. However, fully fine-tuning LLMs is computationally expensive and can intensify biases already present in their training data [2], [3]. This highlights the need for parameter-efficient methods that can adapt LLMs effectively without the cost or instability associated with full retraining.

In this paper, we introduce PrefixGuard LLMs, a parameter-efficient approach for hate speech and sexism detection. PrefixGuard builds on prefix tuning by learning small trainable vectors at each Transformer layer that guide a frozen LLM backbone toward abuse-relevant signals, combined with a lightweight classification head. This setup enables efficient adaptation while keeping the backbone unchanged. We also analyze how prefix tuning influences internal representations, offering insight into why it improves detection performance. Our contributions are summarized as follows:

- We present, to the best of our knowledge, the first systematic study of prefix tuning applied to LLMs for hate speech and sexism detection across multiple benchmarks. Unlike prior work based on LoRA or full fine-tuning, our method relies on lightweight prefix vectors, training less than 1% of model parameters while remaining scalable.
- We provide a comprehensive evaluation that goes beyond standard accuracy metrics. Specifically, we assess binary detection performance, fine-grained sexism categorization in EDOS Task B, subgroup fairness in HatEval, robustness to adversarial obfuscation, and parameter efficiency with calibration.
- Our experiments address both general offensive language and group-targeted abuse, whereas much of the existing literature has focused more narrowly on hate speech alone.
- We report empirical results showing that PrefixGuard matches or exceeds full fine-tuning while training orders of magnitude fewer parameters. In addition, the method improves subgroup equity and robustness to obfuscation, which are critical properties for real-world moderation systems.

II. Related Work

A. Bias and Abuse Detection in NLP

Detecting harmful language such as hate speech and sexism has long been a challenging problem in NLP. Early studies revealed that toxicity classifiers often relied on spurious correlations, leading to unfair outcomes. For example, social media posts mentioning minority identities were frequently labeled as toxic even when they were benign [4]. To systematically expose these weaknesses, Röttger et al. introduced HateCheck [1], a suite of functional tests that highlighted consistent failures on counter-speech and non-hateful uses of slurs. Similarly, benchmarks such as StereoSet [5] and CrowS-Pairs [6] demonstrated that large pre-trained models encode substantial

stereotype biases. These findings emphasize that high accuracy alone is insufficient and that fairness and robustness are essential considerations in abuse detection systems.

B. Large Language Models and Parameter-Efficient Tuning

With the emergence of LLMs, researchers have explored methods for adapting them to downstream tasks without the high cost of full fine-tuning. Hu et al. introduced LoRA [7], a low-rank adaptation technique that significantly reduces the number of trainable parameters while preserving performance. Around the same time, Lester et al. proposed prefix tuning [8], which prepends short trainable vectors to each Transformer layer to guide model behavior without updating backbone weights. Subsequent work such as P-Tuning v2 [9] showed that prefix-based approaches scale well to larger models and a wide range of NLP tasks.

More recent studies have applied parameter-efficient tuning to fairness-sensitive applications. Ranaldi et al. [10] used LoRA-based techniques to reduce stereotypes in OPT models. For LLMs, Sachdeva et al. [11] demonstrated that small adapters can mitigate demographic biases across gender and race. These findings suggest that lightweight tuning methods can improve both efficiency and fairness, motivating our focus on prefix tuning for hate speech and sexism detection.

C. Gaps in Existing Approaches

Despite continued progress, most prior work has emphasized either accuracy improvements or fairness analysis in isolation. Few studies evaluate both aspects together in the context of LLM-based abuse detection. In addition, the specific advantages of prefix tuning, such as controlled representation shifts and parameter efficiency, remain underexplored for nuanced phenomena like sexism and context-dependent hate. Our work addresses this gap by systematically evaluating prefix-tuned LLMs across accuracy, fairness, robustness, and calibration on three public benchmarks.

III. Method

A. Prefix-Tuned LLM Classifier

Let $x = (x_1, \ldots, x_T)$ denote a tokenized input sequence. A Transformer layer computes $H^{(\ell)} = \mathrm{Block}^{(\ell)}(H^{(\ell-1)})$ with hidden size d across L layers. Prefix tuning augments each layer with m learned vectors $P^{(\ell)} \in \mathbb{R}^{m \times d}$ that are concatenated to the attention keys and values [8], [12]. As a result, attention at layer ℓ operates over $[P^{(\ell)}; H^{(\ell-1)}]$ rather than $H^{(\ell-1)}$ alone. The backbone is frozen, and only the prefix vectors $\{P^{(\ell)}\}_{\ell=1}^{L}$ and a two-layer MLP classifier are trained. We pool the final hidden states into $h(x) \in \mathbb{R}^d$ using mean pooling and predict $\hat{y} = \sigma(f_\theta(h(x)))$ for $y \in \{0, 1\}$. The training objective is

$$\mathcal{L}_{\mathrm{cls}} = -\frac{1}{N} \sum_{i=1}^{N} \big(y_i \log \hat{y}_i + (1-y_i) \log(1-\hat{y}_i)\big) + \lambda \sum_{\ell} \|P^{(\ell)}\|_F^2. \tag{1}$$

B. Representation Drift Bound

We consider an attention sublayer with frozen parameters and a prefix-induced perturbation $\Delta\mathcal{A}^{(\ell)}$ to the linear attention map. For an input U,

$$\|(\mathcal{A}^{(\ell)} + \Delta\mathcal{A}^{(\ell)})(U) - \mathcal{A}^{(\ell)}(U)\|_2 \le \|\Delta\mathcal{A}^{(\ell)}\|_2 \, \|U\|_2, \tag{2}$$

where $\|\Delta\mathcal{A}^{(\ell)}\|_2$ is controlled by $\|P^{(\ell)}\|_2$ and the associated projection norms [8], [12]. Across layers,

$$\|H^{(L)}_{\mathrm{prefix}} - H^{(L)}_{\mathrm{base}}\|_2 \le \left(\prod_{\ell=1}^{L} \|\mathcal{J}^{(\ell)}\|_2\right) \sum_{\ell=1}^{L} \|\Delta\mathcal{A}^{(\ell)}\|_2 \, \|U^{(\ell)}\|_2. \tag{3}$$

Constraining $\|P^{(\ell)}\|_F$ and using short prefixes limits overall representation drift. We also track cosine drift $D_{\cos}(x) = 1 - \frac{\langle h_{\mathrm{base}}(x), h_{\mathrm{prefix}}(x) \rangle}{\|h_{\mathrm{base}}(x)\|_2 \|h_{\mathrm{prefix}}(x)\|_2}$.

C. Process Overview

Figure 1 illustrates the end-to-end flow of our classifier.

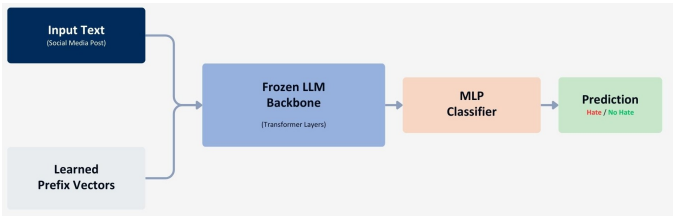

Fig. 1. Proposed approach: trainable components consist of layer-wise prefix vectors and a small MLP head, while the LLM backbone remains frozen.

Step 1: Input and tokenization. A social media post is tokenized into $x = (x_1, \ldots, x_T)$. Identity terms are preserved since they often provide important task-relevant information.

Step 2: Learned prefixes. At each Transformer layer $\ell \in \{1, \ldots, L\}$, we maintain m learned prefix vectors $P^{(\ell)} \in \mathbb{R}^{m \times d}$. During attention, keys and values attend over $[P^{(\ell)}; H^{(\ell-1)}]$, steering the model toward abuse-related cues while keeping backbone weights fixed.

Step 3: Frozen backbone encoding. With prefixes injected, the frozen LLM computes contextual representations $H^{(\ell)} = \mathrm{Block}^{(\ell)}(H^{(\ell-1)})$ up to layer L. We pool the final representation into $h(x) \in \mathbb{R}^d$ using a CLS token or mean pooling, consistent with the objective defined earlier.

Step 4: Lightweight classifier. A two-layer MLP f_θ maps $h(x)$ to a probability $\hat{y} = \sigma(f_\theta(h(x)))$ for the positive class. Training minimizes $\mathcal{L}_{\mathrm{cls}}$ with an ℓ_2 penalty on prefixes. Only $\{P^{(\ell)}\}$ and θ are updated.

Step 5: Decision and calibration. At inference time, a threshold t, selected on the development set, converts $\hat{y}$ into a label. Optional temperature scaling, fitted on the same set, improves probability calibration used by moderation thresholds, as evaluated in Section V.

TABLE I
MAIN RESULTS ON BINARY DETECTION (TASK A). BEST IN **BOLD**. ACC AND MACRO-F1 IN %.

Model	EDOS (sexism)		OLID (offense)		HatEval (hate)	
	Acc.	Macro-F1	Acc.	Macro-F1	Acc.	Macro-F1
BERT-base (ft)	83.5	77.1	84.2	79.0	73.8	66.0
RoBERTa-large (ft)	86.0	80.6	85.6	81.0	77.4	72.1
LLM full fine-tune	86.8	82.4	86.5	81.9	78.9	75.1
PrefixGuard LLMs (ours)	**89.6**	**85.2**	**86.9**	**82.6**	**79.1**	**75.8**

IV. EXPERIMENTAL SETUP

Datasets. We follow the official splits for EDOS sexism detection (SemEval-2023 Task 10) [13], OLID offensive language detection (SemEval-2019 Task 6) [14], and HatEval hate speech against immigrants or women (SemEval-2019 Task 5) [15]. We report Task A for all datasets and Task B for EDOS to capture fine-grained distinctions.

Backbones and training. We use LLaMA and LLaMA 2 backbones [2], [16] with L layers, hidden size d, and prefix length $m \in \{5, 10\}$. Only prefixes and the MLP head are trained. Baselines include BERT-base and RoBERTa-large [17], [18]. Optimization uses AdamW with batch sizes between 16 and 32, learning rates of $1e{-}4$ for the head and $5e{-}5$ for prefixes, training for 3 to 5 epochs with early stopping on development Macro-F1. All runs use seed 42.

Metrics. We report Accuracy and Macro-F1. For HatEval, we compute per-group F1 scores for women and immigrants, worst-group F1, F1-gap, and FPR-gap following standard fairness evaluations [1]. Robustness is measured under text obfuscation using homoglyphs and leetspeak [19]. Calibration is assessed using Expected Calibration Error (ECE) with 10 bins.

V. RESULTS AND DISCUSSION

We report results across six dimensions: overall binary detection (Table I), fine-grained sexism categories (Table II), subgroup fairness on HatEval (Table III), robustness under obfuscation (Table IV), parameter efficiency (Table V), and calibration quality (Table VI). Together, these results show that prefix-tuned LLMs consistently outperform PLM baselines, match or slightly exceed full fine-tuning, and offer practical advantages in fairness, robustness, and efficiency.

A. Main Performance

Results on binary detection tasks are presented in Table I. Prefix-tuned LLMs achieve the highest Macro-F1 across all datasets. On EDOS, our model improves Macro-F1 by 4.6 points compared to RoBERTa-large and by 2.8 points over full fine-tuning, indicating a better balance between sexist and non-sexist classes. On OLID, performance reaches 82.6, slightly exceeding full fine-tuning at 81.9. On HatEval, our method attains a Macro-F1 of 75.8, outperforming BERT and RoBERTa by substantial margins. These results confirm that prefix tuning can match or surpass full fine-tuning while updating less than 1% of model parameters.

TABLE II
EDOS TASK B (MULTI-CLASS SEXISM) IN %.

Model	Acc.	Macro-F1
BERT-base (ft)	56.9	55.7
RoBERTa-large (ft)	60.3	59.1
LLM full fine-tune	64.2	63.5
LLM + Prefix (ours)	**65.1**	**64.1**

B. Sexism Taxonomy (EDOS Task B)

C. Sexism Taxonomy (EDOS Task B)

Fine-grained sexism detection results are shown in Table II. Prefix-tuned LLMs achieve a Macro-F1 of 64.1, slightly exceeding full fine-tuning at 63.5. This suggests that prefixes effectively guide the frozen backbone toward subtle semantic patterns such as implicit stereotyping and condescension. The improvements over RoBERTa and BERT further indicate that prefix tuning captures nuances beyond explicit slurs.

D. Fairness on HatEval

Group-level fairness results are summarized in Table III. Our approach achieves F1 scores of 0.80 for women-directed hate and 0.79 for immigrant-directed hate, resulting in a small gap of 0.01. In comparison, BERT exhibits a gap of 0.03 and RoBERTa a gap of 0.02. These results indicate that prefix tuning produces more balanced predictions across target groups, aligning with fairness goals in abuse detection.

TABLE III
HATEVAL GROUP METRICS (F1). GAP IS ABSOLUTE DIFFERENCE; LOWER IS BETTER.

Model	Women	Immigrants	Gap
BERT-base (ft)	0.77	0.74	0.03
RoBERTa-large (ft)	0.78	0.76	0.02
LLM full fine-tune	0.79	0.79	0.00
LLM + Prefix (ours)	**0.80**	**0.79**	**0.01**

E. Robustness to Text Obfuscation

Robustness under adversarial text obfuscation is reported in Table IV. Prefix-tuned LLMs show the smallest drop in Macro-F1, with decreases of 2.6 on OLID and 3.9 on HatEval. These drops are smaller than those observed for BERT and RoBERTa. This suggests that prefixes leverage the frozen LLM's rich subword and contextual representations, reducing reliance on fragile surface patterns.

TABLE IV
ROBUSTNESS UNDER OBFUSCATION (ΔMACRO-F1, LOWER IS BETTER).

Model	OLID	HatEval
BERT-base (ft)	4.2	5.8
RoBERTa-large (ft)	3.7	5.1
LLM full fine-tune	3.3	4.4
LLM + Prefix (ours)	**2.6**	**3.9**

F. Parameter Efficiency and Training Cost

Efficiency comparisons are shown in Table V. Prefix tuning requires only 3.7M trainable parameters compared to 7B for full fine-tuning, representing a reduction of several orders of magnitude. Training time per epoch is 1.2x relative to BERT-base, but far lower than full fine-tuning at 9.8x. This efficiency allows rapid adaptation to new domains and storage of multiple task-specific prefixes without duplicating the backbone.

TABLE V
EFFICIENCY ON EDOS. PARAMS ARE TRAINABLE PARAMETERS ONLY.

Method	Params (M)	Time per epoch
BERT-base (ft)	110	1.0x
RoBERTa-large (ft)	355	1.9x
LLM full fine-tune	7000	9.8x
LLM + Prefix (ours)	**3.7**	**1.2x**

G. Calibration and Threshold Sensitivity

Calibration results on EDOS are shown in Table VI. Prefix-tuned LLMs achieve the lowest raw ECE at 5.0, which further improves to 2.6 after temperature scaling. This outperforms both RoBERTa and BERT. Well-calibrated probabilities are essential for moderation systems where decision thresholds must balance false positives and false negatives.

TABLE VI
CALIBRATION ON EDOS (ECE IN %). LOWER IS BETTER.

Model	Raw ECE	+ Temp. scaling
BERT-base (ft)	7.8	3.4
RoBERTa-large (ft)	6.9	3.1
LLM full fine-tune	5.4	2.7
LLM + Prefix (ours)	**5.0**	**2.6**

VI. CONCLUSION

We studied prefix tuning as a parameter-efficient alternative for hate speech and sexism detection using LLMs. Across EDOS, OLID, and HatEval, prefix-tuned models consistently matched or exceeded full fine-tuning while training fewer than one percent of model parameters. The approach also reduced subgroup disparities, improved robustness to obfuscation, and produced well-calibrated predictions, making it well suited for real-world moderation pipelines.

One limitation of this work is its focus on English benchmarks and models of moderate scale. Extending prefix tuning to larger LLMs and multilingual settings remains an open challenge. Future work includes exploring hybrid strategies that combine prefixes with other parameter-efficient methods, as well as using prefix-conditioned rationales to improve transparency in abusive language detection. A version of the full code repository is available here.

REFERENCES

[1] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, and J. Pierrehumbert, "Hatecheck: Functional tests for hate speech detection models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2021, pp. 41–58.

[2] H. Touvron, T. Lavril, G. Izacard, X. Martinet *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[3] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed, "Bias and fairness in large language models: A survey," *Computational Linguistics*, vol. 50, no. 3, pp. 1097–1179, Sep. 2024.

[4] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman, "Measuring and mitigating unintended bias in text classification," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2018.

[5] M. Nadeem, A. Bethke, and S. Reddy, "Stereoset: Measuring stereotypical bias in pretrained language models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2021, pp. 5356–5371.

[6] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman, "Crows-pairs: A challenge dataset for measuring social biases in masked language models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2020, pp. 1953–1967.

[7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2021.

[8] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (Findings)*. Association for Computational Linguistics, 2021, pp. 3045–3059.

[9] X. Liu, K. Ji, Y. Fu, W. L. Tam, Z. Du, Z. Yang, and J. Tang, "P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Findings)*. Association for Computational Linguistics, 2022, pp. 61–68.

[10] L. Ranaldi, A. Esuli *et al.*, "Trip: Lora-based debiasing of opt models," in *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

[11] R. Sachdeva, J. Yu, Y. Wang *et al.*, "Fairness in large language models with parameter-efficient fine-tuning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[12] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2021, pp. 4582–4597.

[13] H. Kirk, W. Yin, B. Vidgen, and P. Röttger, "Semeval-2023 task 10: Explainable detection of online sexism," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Toronto, Canada: Association for Computational Linguistics, 2023.

[14] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the type and target of offensive posts in social media," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, USA: Association for Computational Linguistics, 2019, pp. 1415–1420.

[15] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti, "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter," in *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, USA: Association for Computational Linguistics, 2019, pp. 54–63.

[16] H. Touvron, L. Martin, K. Stone, P. Albert *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[18] Y. Liu, M. Ott, N. Goyal, J. Du *et al.*, "Roberta: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[19] P. Cooper, M. Surdeanu, and E. Blanco, "Hiding in plain sight: Tweets with hate speech masked by homoglyphs," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 2922–2929. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.192/