PrefixGuard: Lightweight LLMs for Hate Speech and Sexism Detection

Ginel Dorleon
SogetiLabs, Capgemini
Toulouse, France
0000-0003-2343-4445

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

***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 proliferation of online platforms has amplified the spread of harmful language, including hate speech and sexism. Detecting such content is critical for maintaining safe and inclusive digital spaces, yet it remains a challenging task. Hate and sexist language often appear in subtle or context-dependent forms, ranging from overt slurs to implicit stereotypes or backhanded remarks. Traditional NLP models frequently misclassify these cases, either overlooking implicit abuse or over-flagging benign mentions of identity terms [1].

Large Language Models (LLMs) offer a promising foundation for this task due to their broad linguistic knowledge and contextual understanding. However, fully fine-tuning LLMs is computationally prohibitive and risks amplifying biases present in the training data [2]. This motivates the need for parameter-efficient methods that adapt LLMs effectively without incurring the cost or instability of 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 steer the frozen LLM backbone toward abuse-relevant cues, combined with a lightweight classification head. This approach enables efficient adaptation while leaving the backbone unchanged. We further analyze how prefix tuning shapes internal representations, providing insights into why it improves detection.
Our contributions are as follows:

- We present a systematic study of prefix tuning of LLMs for hate speech and sexism detection across multiple benchmarks. Unlike prior work relying on LoRA or full fine-tuning, our approach adapts LLMs through lightweight prefix vectors, requiring less than 1% of parameters and enabling efficient scalability.
- We provide a comprehensive evaluation that goes beyond standard accuracy. Specifically, we assess five complementary dimensions: binary detection performance, nuanced sexism categorization (EDOS Task B), subgroup fairness in hate speech (HatEval), robustness to adversarial obfuscation, and parameter efficiency with calibration.
- Our experiments address both general offensive language and group-targeted abuse, whereas prior studies have focused more narrowly on hate speech alone.
- We report empirical insights showing that PrefixGuard match or exceed full fine-tuning while training orders of fewer parameters. Moreover, PrefixGuard improves subgroup equity and robustness to obfuscation, which are essential for deployment in real-world moderation systems.

II. Related Work

A. Bias and Abuse Detection in NLP

Detecting harmful language such as hate speech and sexism has been a long-standing challenge in NLP. Early work showed that toxic language classifiers often relied on spurious correlations, leading to unfair outcomes. For example, tweets mentioning minority identities were disproportionately flagged as toxic, even when benign [3]. To systematically test these weaknesses, Röttger et al. introduced HateCheck [1], a suite of functional tests that revealed systematic failures of hate speech classifiers on counter-speech and non-hateful slur uses. Similarly, benchmarks such as StereoSet [4] and CrowS-Pairs [5] highlighted that large pre-trained models encode significant stereotype biases. These findings underscore the importance of not only achieving high accuracy, but also ensuring fairness and robustness in abuse detection systems.

B. Large Language Models and Parameter-Efficient Tuning

With the rise of LLMs, researchers have explored adapting them for classification tasks without the high cost of full fine-tuning. Hu et al. introduced LoRA [6], a low-rank adaptation method that significantly reduces trainable parameters while preserving performance. Around the same time, Lester et al. proposed prefix tuning [7], which prepends short trainable vectors to each Transformer layer to steer model behavior without updating backbone weights. Follow-up work, such as P-Tuning v2 [8], showed that prefix-like methods scale well to larger models and diverse NLP tasks.

Recent studies have applied parameter-efficient tuning to fairness-sensitive domains. Ranaldi et al. [9] used LoRA-based debiasing to reduce stereotypes in OPT models. For LLMs, authors showed that fine-tuning with small adapters can reduce demographic biases across gender and race. These findings suggest that lightweight tuning can improve both efficiency and fairness, motivating our use of prefix tuning for hate speech and sexism detection.

C. Gaps in Existing Approaches

Despite progress, most prior work has focused either on accuracy improvements or on fairness evaluation in isolation. Few studies integrate both perspectives in the context of LLMs for abuse detection. Moreover, the specific benefits of prefix tuning, such as representation steering 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 $x = (x_1, \ldots, x_T)$ be a tokenized input. A transformer layer computes $H^{(\ell)} = \mathrm{Block}^{(\ell)}(H^{(\ell-1)})$ for hidden size d and 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 key and value sequences [7], [10]. Attention at layer ℓ attends over $[P^{(\ell)}; H^{(\ell-1)}]$ rather than $H^{(\ell-1)}$ only. The backbone is frozen; only $\{P^{(\ell)}\}_{\ell=1}^{L}$ and a 2-layer MLP head are trained. We pool the final hidden states into $h(x) \in \mathbb{R}^d$ (mean) and predict $\hat{y} = \sigma(f_\theta(h(x)))$ for $y \in \{0, 1\}$. The 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

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

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

with $\|\Delta\mathcal{A}^{(\ell)}\|_2$ controlled by $\|P^{(\ell)}\|_2$ and projection norms [7], [10]. Over layers,

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

so constraining $\|P^{(\ell)}\|_F$ and using small m limits global drift. We also monitor 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.

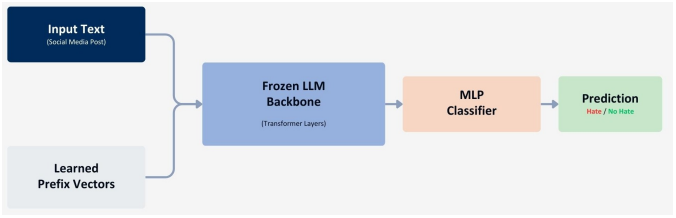

Fig. 1. Proposed approach: trainable components are the layer-wise prefix vectors and the small MLP head. The LLM backbone is frozen.

Step 1: Input and tokenization. A social media post is tokenized into $x = (x_1, \ldots, x_T)$. No identity terms are removed since they carry task signal.

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 $[P^{(\ell)}; H^{(\ell-1)}]$ rather than $H^{(\ell-1)}$ alone, steering the model toward abuse-relevant cues while keeping all backbone weights frozen.

Step 3: Frozen backbone encoding. With prefixes injected, the frozen LLM computes contextual states $H^{(\ell)} = \mathrm{Block}^{(\ell)}(H^{(\ell-1)})$ 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 (hate or sexist). Training minimizes $\mathcal{L}_{\mathrm{cls}}$ with an ℓ_2 penalty on prefixes $\lambda \sum_\ell \|P^{(\ell)}\|_F^2$. Only $\{P^{(\ell)}\}$ and θ are updated.

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

IV. Experimental Setup

Datasets. We follow official splits for EDOS sexism detection (SemEval-2023 Task 10) [11], OLID offensive language (SemEval-2019 Task 6) [12], and HatEval hate against immigrants or women (SemEval-2019 Task 5) [13]. We report Task A (binary) for all datasets and Task B (EDOS taxonomy) for nuance.

Backbones and training. We use LLaMA and Llama 2 backbones [2], [14] with L layers, hidden size d, and per-layer

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**

prefix length $m \in \{5, 10\}$. Only prefixes and the MLP head are trained. Baselines are BERT-base and RoBERTa-large [15], [16]. Optimization uses AdamW, batch 16–32, learning rates $1e{-}4$ (head) and $5e{-}5$ (prefix), 3–5 epochs, early stopping on dev Macro-F1, seed=42.

Metrics. We report Accuracy and Macro-F1. For HatEval, we compute per-group F1 (women vs. immigrants), worst-group F1, F1-gap, and FPR-gap following fairness evaluations for abusive language [1]. Robustness: performance under text obfuscation (homoglyphs, leetspeak) [17]. Calibration: Expected Calibration Error (ECE, 10-bin).

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). Taken together, these results show that prefix-tuned LLMs consistently improve accuracy over PLM baselines, match or slightly surpass full fine-tuning, and offer practical advantages in fairness, robustness, and efficiency. Each subsection below discusses one of these aspects in detail.

A. Main Performance

Results on binary detection tasks are presented in Table I. Prefix-tuned LLMs achieve the best Macro-F1 across all datasets. On EDO, our model improves Macro-F1 by +4.6 points compared to RoBERTa-large and by +2.8 points over full fine-tuning, showing superior balance between sexist and non-sexist classes. On OLID, performance is also highest (82.6), slightly ahead of full fine-tuning (81.9). On HatEval, our method achieves Macro-F1 of 75.8, outperforming BERT and RoBERTa baselines by 9.8 and 3.7 points respectively. These results confirm that prefix tuning matches or surpasses full fine-tuning while updating less than 1% of parameters.

B. Sexism Taxonomy (EDOS Task B)

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**

Fine-grained sexism detection results are shown in Table II. Prefix-tuned LLMs achieve 64.1 Macro-F1, slightly surpassing full fine-tuning (63.5). This indicates that prefixes effectively steer the frozen backbone toward semantic nuances like implicit stereotyping and condescension. The improvements over RoBERTa (59.1) and BERT (55.7) further demonstrate that prefix tuning captures subtleties beyond explicit slurs.

C. Fairness on HatEval

Group-level fairness results are summarized in Table III. Our method reaches F1=0.80 for women-directed hate and 0.79 for immigrant-directed hate, with a minimal gap of 0.01. In comparison, BERT exhibits a 0.03 gap and RoBERTa a 0.02 gap. This suggests prefix tuning delivers more balanced predictions across groups, aligning with fairness objectives 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**

D. Robustness to Text Obfuscation

Robustness evaluation under adversarial text obfuscation is reported in Table IV. Prefix-tuned LLMs show the smallest drop in Macro-F1 (2.6 on OLID, 3.9 on HatEval), compared to larger declines for BERT (4.2/5.8) and RoBERTa (3.7/5.1). This robustness indicates that prefixes exploit the frozen LLM's rich subword and contextual knowledge, avoiding brittle reliance on surface forms.

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**

E. Parameter Efficiency and Training Cost

Efficiency comparisons are presented in Table V. Prefix tuning requires only 3.7M trainable parameters versus 7B for full fine-tuning, a reduction of several orders of magnitude. Training time per epoch is 1.2x relative to BERT-base, but vastly lower than full fine-tuning (9.8x). This parameter efficiency enables quick adaptation to new domains and storage of multiple 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**

F. Calibration and Threshold Sensitivity

Calibration results on EDOS are given in Table VI. Prefix-tuned LLMs obtain the lowest raw ECE (5.0) among all models, further reduced to 2.6 after temperature scaling. This surpasses RoBERTa (6.9→3.1) and BERT (7.8→3.4). Well-calibrated probabilities are crucial for moderation systems where thresholds 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 investigated prefix tuning as a parameter-efficient alternative for hate speech and sexism detection with LLMs. Across EDOS, OLID, and HatEval, prefix-tuned models consistently matched or exceeded full fine-tuning while training less than one percent of parameters. The approach also reduced subgroup disparities, improved robustness to obfuscation, and yielded well-calibrated predictions, making it practical for real-world moderation pipelines.

A limitation of this work is the focus on English-only benchmarks and models of moderate scale. Extending prefix tuning to larger LLMs and multilingual settings remains an open challenge. Future directions include exploring hybrid strategies that combine prefixes with other parameter-efficient methods, and leveraging prefix-conditioned rationales to improve transparency in abusive language detection. A version of the complete code repository is accessible 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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] L. Ranaldi, E. S. Ruzzetti, D. Venditti, D. Onorati, and F. M. Zanzotto, "A trip towards fairness: Bias and de-biasing in large language models," in *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, D. Bollegala and V. Shwartz, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024. [Online]. Available: https://aclanthology.org/2024.starsem-1.30/

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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*, Dec. 2023.