VITECH@DravidianLangTech2026: Prompting and LoRA Adaptation for Tamil Abusive Language Detection - A Comparative Study of Open LLMs

Triambiga Krubhakaran **Senthil Kumar B**

Kaviya Nagarajan **Balaji N**

Velammal Institute of Technology
Chennai, Tamil Nadu, INDIA
{it24079, senthilkumar.it, it24036, balajin}@velammalitech.edu.in

Abstract

The detection of abusive Tamil text using large language models (LLMs) has received relatively little attention compared to BERT variations. We empirically evaluated four families of open-weight LLMs —Gemma, LLaMA, Qwen, and DeepSeek-Distilled— on the Tamil dataset provided by the shared task. The models are assessed under two in-context learning settings (zero-shot and few-shot) and a parameter-efficient fine-tuning approach using LoRA, with model sizes of approximately 2B and 8B parameters. Experimental results show that 8B models consistently outperform their 2B counterparts, indicating the benefit of increased model capacity. Among the adaptation techniques, LoRA fine-tuning significantly outperforms both zero-shot and few-shot prompting. Across all evaluated settings, Google's Gemma-2-9B model with LoRA fine-tuning achieved the best performance compared to the other model families and our test result was ranked 12th among all 22 submissions with the 0.7959 f1-score.

1 Introduction

The proliferation of user-generated content on social media has intensified the spread of abusive and offensive language online. Automatic abusive language detection has therefore become a critical research problem in natural language processing (NLP), particularly for content moderation and online safety applications (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). While substantial progress has been made for high-resource languages such as English (Zampieri et al., 2019), comparatively limited work has focused on low-resource and morphologically rich languages like Tamil.

Tamil presents several linguistic and computational challenges for abusive language detection. Social media text frequently contains spelling variations, dialectal forms, informal expressions, transliteration into Latin script, and Tamil–English code-mixing. These characteristics complicate lexical matching and contextual modeling. Furthermore, the scarcity of large-scale annotated corpora restricts the effectiveness of fully supervised approaches (Jauhiainen et al., 2021). Recent advances in large language models (LLMs) have demonstrated strong zero-shot and few-shot generalization across diverse NLP tasks (Brown et al., 2020; Touvron et al., 2023). Open-weight models have further enabled task adaptation through parameter-efficient fine-tuning methods such as Low-Rank Adaptation (LoRA) (Hu et al., 2022). However, their effectiveness in detecting abusive text in low-resource areas such as Tamil text remains underexplored.

We systematically evaluate four open-weight LLM families: Gemma, Llama, DeepSeek Distilled Qwen, and Qwen. For each family, we experiment with 2B and 8B parameter variants to analyze the impact of model scale. We assess three adaptation paradigms: 1.Zero-shot prompting, 2.Few-shot in-context learning and 3.Parameter-efficient fine-tuning using LoRA. This experimental design enables a controlled comparison between prompt-based inference and supervised adaptation strategies.

Our study aims to address the following research questions (**RQ**):

- **RQ1:** How effective are open-weight LLMs in zero-shot and few-shot settings for Tamil abusive text detection?
- **RQ2:** What is the impact of model scale (2B vs 8B) on classification performance?
- **RQ3:** To what extent does LoRA-based fine-tuning improve performance over prompt-only approaches?

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 436–441

2 Related Work

A recent UN Women-commissioned survey [1] of women in the public sphere reveals that around 70% of women participants who work in the fields of human rights, activism, and/or journalism said that they had experienced online violence in the course of their work. This emphasizes the need to detect abusive text on social networks.

Early shared tasks on abusive Tamil content were organized at the DravidianLangTech workshop series. In DravidianLangTech-ACL 2022 (Priyadharshini et al., 2022), the task focused on classifying abusive comments in Tamil and Tamil–English code-mixed text, the participants employing machine learning, deep learning, and transformer methods such as mBERT and MuRIL BERT (Rajalakshmi et al., 2022). Subsequent shared tasks expanded the scope to Tamil and Telugu abusive comment detection (Priyadharshini et al., 2023), with competitive systems leveraging transformer-based architectures such as mBERT and XLM-RoBERTa achieving the highest performance (Hegde et al., 2023).

More recently, the shared task at NAACL 2025 under the DravidianLangTech workshop (Rajiakodi et al., 2025) focused on abusive Tamil and Malayalam text targeting women on social media. The findings emphasized the effectiveness of the multilingual pre-trained model IndicBERT-v2 over the other BERT variants (Hanif and Rahman, 2025). Despite these advances, prior work has largely focused on fully supervised fine-tuning of encoder-based transformer models. Systematic investigations of large language models (LLMs) based on decoder-only for Tamil abusive text detection remain limited.

Our work addresses this gap by conducting a controlled evaluation of multiple open-weight LLM families on different parameter scales (2B and 8B) under zero-shot, few-shot, and LoRA-based adaptation paradigms within a shared task setting. This analysis offers new insights into the trade-offs between model scale, adaptation strategy, and performance in Dravidian NLP contexts.

3 Dataset Statistics

The dataset used in this study was released as part of the shared task by Sivagnanam et al. (2026). It consists of manually annotated Tamil sentences labeled as either Abusive or Non-Abusive. The training split contains 3,652 sentences, while the test split comprises 913 sentences. The training data is relatively balanced across classes, with 1,769 abusive instances and 1,883 non-abusive instances. This near-balanced class distribution reduces bias toward majority-class predictions and supports reliable macro-level evaluation.

In terms of sentence length, the dataset exhibits notable variability. In the training set, sentence length ranges from a minimum of 3 words to a maximum of 383 words, with an average length of 14.48 words per sentence. Similarly, the test set ranges from 6 to 158 words per sentence, with an average of 13.95 words per sentence. Although the average sentence length is relatively short, the presence of long outlier instances introduces modeling challenges, particularly for prompt-based methods with context length constraints.

4 Experiment

We evaluated two model scales - approximately 2B and 8B parameters - across four open-weight large language model (LLM) families: Gemma, Llama, DeepSeek-R1-Distill-Qwen, and Qwen3. The given training dataset of 3652 sentences is split (80:20) into training and validation of 2921 and 731 sentences respectively. The best model across all settings was used to submit our result for the given test data.

4.1 Model Description

For our experiments, we utilized eight decoder-only transformer models on two parameter scales (1–2B and 7–9B), all obtained from Hugging Face and executed in the Google Colab Pro environment. The smaller models (1–2B), hereafter called 2B models, were used for lightweight evaluation and adaptation, while the higher-capacity models (7–9B), hereafter called 8B models, enabled more expressive contextual modeling in identical experimental settings.

In both 2B and 8B model sizes, we used the Gemma series (Gemma-2B[2], Gemma-2-9B[3]) developed by Google DeepMind; Qwen series (Qwen3-1.7B [4] & Qwen2.5-7B [5]) from Alibaba Cloud, a multilingual foundation model optimized

[1] http:https://www.unesco.org/en/articles/global-survey-reveals-rising-violence-against\-women-journalists

[2] https://huggingface.co/google/gemma-2b

[3] https://huggingface.co/google/gemma-2-9b

[4] https://huggingface.co/Qwen/Qwen3-1.7B

[5] https://huggingface.co/Qwen/Qwen2.5-7B

for reasoning; Llama-Tamil (Llama-Tamil-1B [6] & Tamil-Llama-7b[7]), a Tamil-adapted variant of Llama originally developed by Meta AI; and DeepSeek-R1-Distill-Qwen series (Distill-Qwen-1.5B [8] & Distill-Qwen-7B[9]) from DeepSeek, a distilled model designed to retain reasoning capabilities in a smaller architecture.

These models were evaluated under zero-shot, few-shot, and LoRA settings for the detection of abusive Tamil text.

4.2 Adaptation Strategies

Each model was evaluated under three adaptation paradigms:

4.2.1 Zero-Shot Prompting

The models were prompted with a structured instruction-based template as shown in the Appendix.A, specifying the classification labels (abusive or non-abusive). Instead of updating model parameters, the task is described entirely through natural language instructions within the prompt. The model leverages the knowledge acquired during pretraining to map the input text to one of the predefined class labels.

4.2.2 Few-Shot Prompting

The large language model (LLM) learns the task pattern from in-context examples and applies it to a new, unseen instance. This technique leverages the model's in-context learning capability. In the few-shot setting, 5 labeled examples from each class were included in the prompt prior to the test instance. Examples were selected to ensure a balanced representation of both classes. The prompt structure used in few-shot learning is shown in the Appendix.A

4.2.3 LoRA-based Fine-tuning

For supervised adaptation, we used Low-Rank Adaptation (LoRA) (Hu et al., 2022). LoRA layers were injected into the attention projection matrices while freezing the base model parameters. This approach enables parameter-efficient fine-tuning, which is particularly suitable for moderate dataset sizes in low-resource languages.

[6] https://huggingface.co/Jesmma/LLaMA-Tamil_1B

[7] https://huggingface.co/abhinand/tamil-llama-7b-base-v0.1

[8] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B

[9] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B

2B LLM	zero-shot	few-shot
Gemma-2b	**56.05**	33.86
Llama-Tamil-1B	**49.27**	48.50
DeepSeek-R1-Distill-Qwen-1.5B	49.72	**51.15**
Qwen3-1.7B	12.07	**58.59**

Table 1: Effectiveness of 2B LLMs in zero-shot & few-shot settings. **Bold** indicates the better f1-score among the prompting

8B LLM	zero-shot	few-shot
Gemma-2-9B	**64.06**	58.17
Tamil-Llama-7B	**65.28**	65.19
DeepSeek-R1-Distill-Qwen-7B	**61.77**	15.10
Qwen2.5-7B	58.71	**65.51**

Table 2: Effectiveness of 8B LLMs in zero-shot & few-shot settings. **Bold** represents the better f1-score among the prompting

4.3 Implementation

All experiments were implemented using the HuggingFace Transformers library, with the PEFT framework for LoRA-based parameter-efficient fine-tuning. The NVIDIA T4 GPU (16GB VRAM) available in the Google Colab free tier serves as an effective computational tool for 2B parameter models. The 8B parameter models was trained on a NVIDIA A100 GPU with 40GB VRAM.

Implemented LoRA for both 2B and 8B parameter models to compare with the performance of the prompting techniques. For both the 2B and 8B models, LoRA was configured as given in the Appendix.B. A higher rank in 2B allows the adapter to capture more complex features that the model might otherwise miss. The 8B models use a lower rank to maintain memory efficiency without overparameterizing already large representations. For 2B models, LoRA is applied only to attention projections (q, k, v, o), focusing on contextual representation learning. For 8B models, adaptation extends to MLP layers, enabling a deeper task-specific transformation. Both 2B and 8B models were adapted by fine-tuning the hyper parameters in LoRA as given in the Appendix.B.

5 Results

The four types of open-weight LLMs at two sizes -2B & 8B- are empirically evaluated using the validation dataset and compared for their effectiveness

2B/8B LLM	Prompt	LoRA	Diff. (△)	% increase
Gemma-2B	56.05	78.75	22.7	40%
Llama-Tamil-1B	49.27	58.04	8.77	18%
Distill-Qwen-1.5B	51.15	73.04	21.89	43%
Qwen3-1.7B	58.59	77.48	18.89	43%
Gemma-2-9B	64.06	81.49	17.43	27%
Tamil-Llama-7B	65.28	80.32	15.04	23%
Distill-Qwen-7B	61.77	69.77	8	13%
Qwen2.5-7B	65.51	76.1	10.59	16%

Table 3: Improvement in the performance of LoRA fine-tuning over prompt-only approach. The best f1-score among zero/few-shot in each model is considered as the prompt score.

LoRA Adaptation		
LLM	**2B**	**8B**
Gemma	78.75	**81.49**
Llama	58.04	**80.32**
DeepSeek-R1-Distill-Qwen	**73.04**	69.77
Qwen	**77.48**	76.10

Table 4: Effectiveness of LoRA adaptation in 2B & 8B LLMs. Bold indicates the better f1-score among the model sizes.

over the three adaptation techniques: zero-shot, few-shot and LoRA. Their effectiveness can be analyzed by answering the research questions (**ARQ**):

- **ARQ1:** From Table.1 & 2, it is evident that Zero-shot works better when the model already has strong multilingual semantic grounding, whereas few-shot helps when the model needs explicit label anchoring or format conditioning.

- **ARQ2:** By comparing the f1-scores in Table.1, 2 across the 2B and 8B LLMs for the prompting techniques, generally 8B size LLMs perform better in zero-shot and few-shot techniques, except for DeepSeek-Distill model. For LoRA fine-tuning, the behavior differs between model families (Table.4). In the case of Gemma and Llama-based models, the 8B variants benefit more from LoRA adaptation. In contrast, DeepSeek-R1-Distill-Qwen and Qwen models perform better with 2B size LLMs. This behavior may be attributed to the relatively small size of the training dataset, where smaller models generalize more effectively and are less prone to overfitting during LoRA adaptation. Furthermore, abusive Tamil text classification is a domain-specific task that may not require the additional representational capacity of larger models. The distilled architecture of DeepSeek-Distill-Qwen-2B may also contribute to improved task efficiency and discriminative learning for classification-oriented objectives.

- **ARQ3:** The percentage increase in performance by LoRA fine-tuning over prompt-only approach implies that LoRA fine-tuning is better than the prompt techniques as shown in Table.3. The LoRA adaptation improve the performance by 36% and 20% approximately for 2B and 8B LLMs respectively, over prompt-only approaches.

6 Conclusion

The open-weight LLMs of two sizes - 2B & 8B - was empirically evaluated through three adaptation techniques. Among the three, LoRA fine-tuning performs better than the two prompting techniques. Although the percentage improvement by LoRA is greater in 2B size LLMs due to poor performance in prompting techniques, 8B LLMs such as Gemma-2-9B and Tamil-Llama-7B achieved the highest f1-score with 81.49% and 80.32% respectively, during the validation. The predictions for the test data are inferred from the top three LoRA fine-tuned LLM models, Gemma-2-9B, Tamil-Llama-7B and Qwen3-1.7B. The system scored a 12th rank among the 22 submissions.

Overall, the above study on open-weight LLM in 2B and 8B sizes infers that the LLM which is multilingual, larger in model size, and fine-tuning performs better in low-resource language.

7 Limitations

The study uses a relatively small dataset of approximately 3,600 samples. Large language models, particularly 8B-scale models, generally require substantially larger datasets for stable adaptation and generalization. Although four LLM families were evaluated, the study focuses primarily on models in the 2B and 8B parameter range. Additionally, the evaluation is restricted to four LLM families and a single shared-task setting focused on abusive Tamil text classification. The study also does not investigate tokenizer behavior, dialectal variation, or advanced prompting strategies, all of which can influence performance in low-resource multilingual contexts.

8 Ethical Considerations

This work uses the publicly released dataset from the shared task organizers solely for research purposes. Since the dataset contains offensive and harmful content targeting women, systems can produce false positives or false negatives, especially in low-resource languages such as Tamil, human oversight remains important in real-world deployment. Although this work evaluates model performance on a downstream classification task, it does not provide a comprehensive fairness analysis.

Limited AI-assisted writing tools were used to improve the grammatical clarity and presentation of the manuscript. All experimental design, implementation, analysis, and interpretation of results were conducted and verified by the authors. The authors carefully reviewed and validated all AI-assisted content to ensure accuracy, originality, and compliance with academic integrity standards.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Tareque Md Hanif and Md Rashadur Rahman. 2025. CUET_Agile@DravidianLangTech 2025: Fine-tuning transformers for detecting abusive text targeting women from Tamil and Malayalam texts. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 315–319, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Asha Hegde, Kavya G, Sharal Coelho, and Hosahalli Lakshmaiah Shashirekha. 2023. MUCS@DravidianLangTech2023: Leveraging learning models to identify abusive comments in code-mixed Dravidian languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 266–274, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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*.

Tommi Jauhiainen, Tharindu Ranasinghe, and Marcos Zampieri. 2021. Comparing approaches to Dravidian language identification. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 120–127, Kiyv, Ukraine. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumar Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Malliga Subramanian, Subalalitha Chinnaudayar Navaneethakrishnan, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Prasanna Kumar Kumaresan, Karnati Sai Prashanth, Mangamuru Sai Rishith Reddy, and Janakiram Chandu. 2023. Overview of shared-task on abusive comment detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 80–87, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Ratnavel Rajalakshmi, Ankita Duraphe, and Antonette Shibani. 2022. DLRG@DravidianLangTech-ACL2022: Abusive comment detection in Tamil using multilingual transformer models. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 207–213, Dublin, Ireland. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyad-

Prompt:
prompt = f"""Task: Classify the following text as either "abusive" or "non-abusive". Text: {text} Classification:"""

Table 5: Zero-shot

harshini, Rajameenakshi J, Kathiravan P, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the shared task on abusive Tamil and Malayalam text targeting women on social media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 671–681, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Bhuvaneswari Sivagnanam, Kathiravan Pannerselvam, Jananayagan V, Charmathi Rajkumar, Ramesh Kannan R, Ratnavel Rajalakshmi, Shunmuga Priya Muthusamy Chinnan, Saranya Rajiakodi, and Bharathi Raja Chakravarthi. 2026. From Comments to Harm: A Findings Report on Abusive Tamil Text Targeting Women on Social Media. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

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.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

9 Appendix

A Prompts

The prompts used in zero-shot and few-shot learning for 2B and 8B size LLMs are shown in Table.5, 6.

Prompt:
prompt=f"""Classify Tamil text as "abusive" or "non-abusive". Respond with ONLY one word. {FEW_SHOT_EXAMPLES} Now classify this text. Respond with only "abusive" or "non-abusive": Text:{text} Classification:"""

Table 6: Few-shot

Config	**2B**	**8B**
Rank - r	16	8
Alpha - α	32	16
Scaling factor	2*r	2*r
Target	attention	attention+MLP

Table 7: LoRA Configuration

B LoRA Configuration & Hyper parameters

The LoRA configuration during LLM adaptation and the corresponding fine-tuning hyperparameters in Table.7, 8 respectively.

Parameter	**2B**	**8B**
Batch size	32	32
Learning rate	2e-4	1e-4
Weight decay	0.01	0.01
Warm-up ratio	0.1	0.1
Optimizer	adamw_torch	paged_adamw

Table 8: LoRA Hyper parameters