Cyberbullying Detection via Aggression-Enhanced Prompting

Aisha Saeid, Anu Sabu, Girish A. Koushik, Ferrante Neri & Diptesh Kanojia
Nature Inspired Computing and Engineering (NICE) Research Group,
School of Computer Science & Electronic Engineering, University of Surrey, UK
a.saeid@surrey.ac.uk

1 Introduction

- **Online platforms** foster connection but also create spaces for harmful behaviour such as **cyberbullying**.
- **Detecting cyberbullying is challenging** [1] due to its subtle, implicit, and **context-dependent** language.
- **LLMs** offer potential but **face limits with domain-specific vocabulary** [3] and reliance on implicit cues.

2 Research Question

Does integrating aggression detection as an auxiliary task **improve the generalisation and detection via LLMs** for **cyberbullying** detection?

3 Methodology

- Proposed Enriched Prompt Pipeline (EPP) to add aggression context for Cyberbullying Detection.
- Applied LoRA fine-tuning for efficient LLM adaptation.
- **Pipeline: predict aggression label → enrich prompt → detect cyberbullying.**

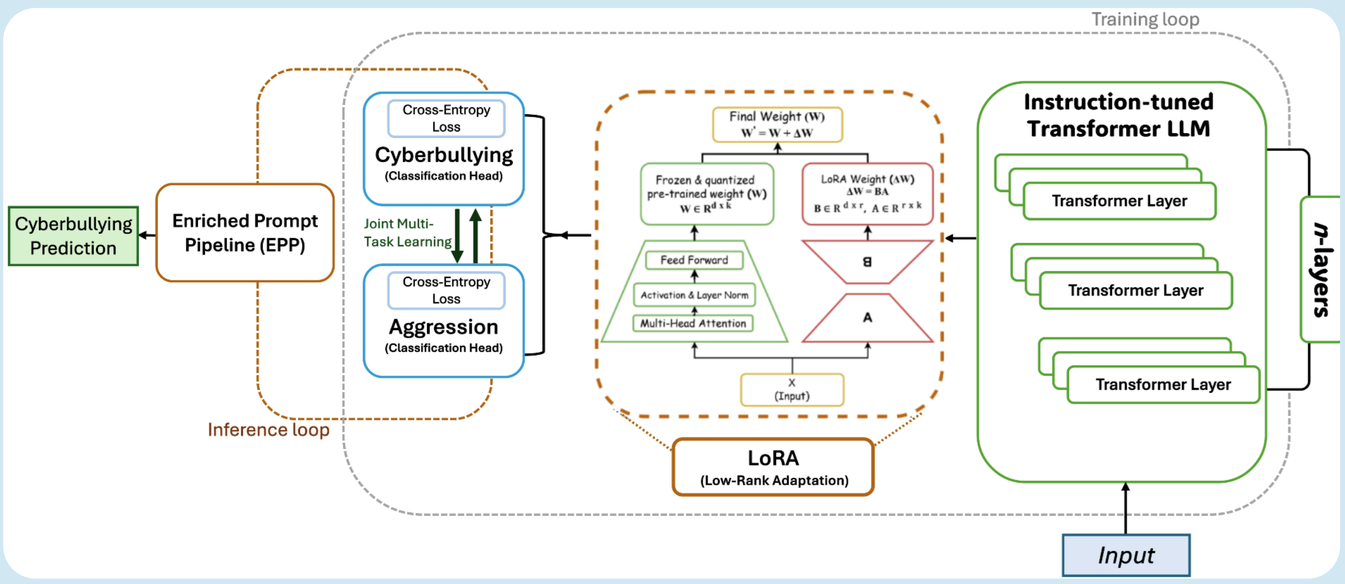

Figure 1: An overview of the proposed system architecture for cyberbullying detection. The diagram illustrates the training loop, which incorporates a LoRA adapter and a joint multi-task learning framework for aggression and cyberbullying, as well as the inference loop that utilises the Enriched Prompt Pipeline (EPP).

4 Results & Discussion

- **EPP consistently outperforms LoRA** and **MTL-LoRA** for cyberbullying detection across model variants.
- Gemma-2-2B **improves 16 %points**, and Gemma-2-9B **improves 6% points**, when leveraging EPP.
- EPP leverages LoRA-trained aggression labels to enhance cyberbullying prompts rather than retraining.

Model	Aggression Detection					Cyberbullying Detection				
	Zero-shot	Few-shot	LoRA	MTL	EPP	Zero-shot	Few-shot	LoRA	MTL	EPP
Gemma-2-2B	0.54	0.56	**0.67**	0.51	**0.67**	0.63	0.83	**0.84**	0.90	**0.99**
Gemma-2-9B	0.57	0.60	0.65	0.53	0.65	0.79	0.83	**0.93**	0.89	**0.99**
Gemma-3-4B	0.53	**0.63**	0.50	0.49	0.50	0.34	0.57	0.84	0.76	0.86

Table 1: Macro-F1 score comparison of models across zero-shot, few-shot, LoRA, MTL, and EPP evaluations for aggression and cyberbullying detection.
Note: **Bold** indicates the best-performing method.

5 Conclusion & Future Work

- Enriched Prompt Pipeline (EPP) with aggression cues substantially improves cyberbullying detection and model generalisation.
- Shows that lightweight, context-aware prompt augmentation is effective for socially sensitive NLP tasks.
- **Future work:** expand evaluation to more LLMs and address deployment challenges.
- **Integration of Aggression clues → stronger LLM performance on Cyberbullying**

References

[1] Fati et al. (2025). Enhancing multiclass cyberbullying classification with transformers. CMES.
[2] Khan et al. (2022). Aggression detection in social media using deep learning. Applied Sciences.
[3] Muminovic (2025). Benchmarking LLMs for cyberbullying detection. arXiv:2505.18927.

Scan to read the full paper on arXiv