PSK@EEUCA 2026: Fine-Tuning Large Language Models with Synthetic Data Augmentation for Multi-Class Toxicity Detection in Gaming Chat

Srikar Kashyap Pulipaka
Independent Researcher
srikar.kashyap@gmail.com

Abstract

This paper describes our system for the EEUCA 2026 Shared Task on Understanding Toxic Behavior in Gaming Communities. The task involves classifying World of Tanks chat messages into six toxicity categories: Non-toxic, Insults/Flaming, Other Offensive, Hate/Harassment, Threats, and Extremism. We explore multiple approaches including encoder-based models, instruction-tuned LLMs with LoRA fine-tuning, hierarchical classification, one-vs-rest strategies, and various ensemble methods. Our best system combines Llama 3.1 8B with carefully calibrated 5% synthetic data augmentation, achieving an F1-macro score of 0.6234 on the test set, placing 4th out of 35 participating teams. We provide extensive analysis of the dataset's annotation patterns and their impact on model generalization, revealing a critical "validation trap" phenomenon where high validation performance correlates with poor test transfer.

1 Introduction

Online gaming communities face significant challenges with toxic behavior, including harassment, hate speech, and threats. The EEUCA 2026 Shared Task on Understanding Toxic Behavior in Gaming Communities (Thapa et al., 2026) focuses on detecting and classifying toxicity in World of Tanks chat messages, aiming to promote healthier digital spaces through AI-based moderation tools.

The task presents several unique challenges:

- Extreme class imbalance (81% Non-toxic, <1% for rare classes)
- Short, informal text with gaming-specific vocabulary
- Multilingual content requiring cross-lingual understanding
- Subtle distinctions between toxic categories (e.g., skill-based insults vs. identity-based hate)

Our main strategy combines instruction-tuned LLMs (Llama 3.1 8B) with parameter-efficient fine-tuning via LoRA and carefully calibrated synthetic data augmentation. We find that a narrow 5% synthetic data ratio is optimal, with deviations in either direction significantly degrading test performance.

Our key discovery is the "validation trap" phenomenon: models achieving high validation F1 through conservative predictions (matching validation distribution) perform poorly on test data. This affected our larger models most severely, with 12B models showing 0.66 validation F1 but only 0.52 test F1. Our final system achieves 0.6234 F1-macro, placing 4th overall out of 35 teams.

2 Background

2.1 Task Description

The EEUCA 2026 toxicity detection task (Thapa et al., 2026) is part of the 9th Workshop on Event Extraction and Understanding (Hürriyetoğlu et al., 2026). The task requires classifying gaming chat messages into six categories based on the annotation schema from Bhandari et al. (2023):

0. **Non-toxic**: Normal or positive communication
1. **Insults/Flaming**: Personal attacks targeting gaming skill
2. **Other Offensive**: Inappropriate content not directly attacking
3. **Hate/Harassment**: Targeted abuse based on identity
4. **Threats**: Violence or harm threats
5. **Extremism**: Hate ideology and dehumanization

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 192–197

2.2 Dataset

The dataset is derived from the GameTox corpus (Naseem et al., 2025), comprising World of Tanks chat messages. Table 1 shows the severe class imbalance, with Non-toxic messages comprising 81% and rare classes (Threats, Extremism) together representing less than 0.2%.

Class	Count	%
0: Non-toxic	34,797	81.0%
1: Insults/Flaming	5,925	13.8%
2: Other Offensive	1,874	4.4%
3: Hate/Harassment	279	0.6%
4: Threats	60	0.1%
5: Extremism	24	0.1%
Total	42,959	100%

Table 1: Training set class distribution showing severe imbalance.

Our analysis revealed significant data quality patterns: 40.2% of training messages are exact duplicates, and 13.4% have the same text with different labels. Interestingly, training on deduplicated data hurt performance (0.44 vs 0.60 F1), suggesting duplicates provide beneficial implicit oversampling.

2.3 Related Work

Toxicity detection has been extensively studied using transformer-based models (Devlin et al., 2019; Liu et al., 2019). Recent work has shown that instruction-tuned LLMs can achieve strong performance on classification tasks (Wei et al., 2022; Thapa et al., 2025). Parameter-efficient fine-tuning methods like LoRA (Hu et al., 2022) enable adaptation of large models with limited resources.

Gaming-specific toxicity presents unique challenges due to domain vocabulary and skill-based criticism that may or may not constitute toxicity (Kwak et al., 2015). Hate speech detection more broadly has been studied with various approaches (Parihar et al., 2021).

3 System Overview

3.1 Model Architecture

We experimented with multiple architectures:

- **XLM-RoBERTa Large** (560M): Full fine-tuning
- **Gemma 2B** (Gemma Team, 2024): LoRA + 8-bit quantization
- **Gemma 3 12B** (Gemma Team, 2025): LoRA + 4-bit quantization
- **Llama 3.1 8B** (Llama Team, AI @ Meta, 2024): LoRA + 4-bit quantization (best)

Our final system uses Llama 3.1 8B with 4-bit NF4 quantization (Dettmers et al., 2023) and LoRA adapters (rank=16, alpha=64).

3.2 Prompt Engineering

Following insights that class definitions help LLMs discriminate between similar categories, we prepend structured definitions to each input:

```
Classify gaming chat toxicity:
0=Non-toxic: Normal/positive chat
1=Insults: Personal attacks, slurs
2=Other Offensive: Inappropriate but not
direct
3=Hate/Harassment: Targeted abuse
4=Threats: Violence/harm threats
5=Extremism: Hate ideology
Message: [input text]
```

This "short" prompt style achieved optimal balance between context and avoiding truncation.

3.3 Synthetic Data Augmentation

We generate synthetic training data via LLM-based paraphrase augmentation, focusing on minority classes. We used a paraphrase-only strategy after preliminary direct-generation experiments produced generic messages that did not match the short, slang-heavy style of real World of Tanks chat. Each source message was rewritten with the following template:

```
Rewrite this World of Tanks game
chat message using different words but
keeping the same meaning and toxicity
level.
Original: [message]
Requirements: Keep EXACT same meaning
and level of toxicity; use natural
gaming language, abbreviations, slang;
similar length (3-20 words). Output ONLY
the rewritten message.
```

The synthetic pool contained 10,464 filtered paraphrases, all from minority toxicity classes: 8,348 for Class 2 (Other Offensive), 1,633 for Class 3 (Hate/Harassment), 343 for Class 4 (Threats), and 140 for Class 5 (Extremism). We applied basic cleaning, invalid-label and length filtering, label-leakage regex filtering, and embedding-based deduplication within the synthetic set. Since paraphrases are intentionally close to their source messages, we did not remove paraphrases for high similarity to the original training examples. Synthetic

examples were added only to the training partition after splitting real data; validation remained 100% real.

For the final 5% setting, we sampled 1,921 synthetic examples from this pool (1,539 Class 2, 282 Class 3, 64 Class 4, 36 Class 5), yielding an actual synthetic share of 4.998% of the training data. The synthetic ratio proved critical:

- **5% synthetic**: Optimal, with best test transfer
- 0% synthetic: Strong validation F1 but lower test transfer
- 10% synthetic: Lower validation and test performance than 5%

The narrow optimal range suggests synthetic data helps by making predictions more "aggressive" on minority classes, better matching test distribution.

4 Alternative Approaches

We explored several alternative strategies that ultimately underperformed:

Hierarchical Classification: Two-stage approach (binary toxic/non-toxic, then 5-class among toxic) achieved 0.67 validation F1 but only 0.47 test F1, the largest generalization gap observed.

One-vs-Rest: Six binary classifiers with aggressive oversampling (up to 500x) and focal loss (Lin et al., 2017). Too conservative at 0.56 validation F1.

Transfer Learning: Pre-training on DOTA 2 toxicity data before fine-tuning resulted in validation trap (0.68 val $\rightarrow$ 0.55 test).

Ensemble Methods: Probability averaging, voting, and confidence routing generally hurt performance because our best single model dominated all classes.

Post-hoc Calibration: Platt scaling, isotonic regression, and temperature scaling provided no improvement.

5 Experimental Setup

5.1 Training Configuration

- Model: Llama 3.1 8B
- Quantization: 4-bit NF4
- LoRA: rank=16, alpha=64, dropout=0.0
- Learning rate: 5e-5 (cosine schedule)
- Epochs: 4
- Batch size: 4 (gradient accumulation: 4)
- Loss: class-weighted cross-entropy
- Synthetic ratio: 5%
- Max sequence length: 384

5.2 Evaluation

The official metric is macro-averaged F1 score across all six classes. We used the provided validation split for development and hyperparameter tuning.

6 Results

6.1 Main Results

Table 2 compares our approaches. Llama 3.1 8B with 5% synthetic data achieves the best test performance. The unboosted 5% synthetic model scored 0.6232; a small post-hoc Class 2 boost increased the official submitted score to 0.6234.

System	Val F1	Test F1
XLM-RoBERTa Large	0.30	–
Gemma 2B	0.63	0.52
Gemma 12B	0.66	0.52
Two-stage	0.67	0.47
Llama 8B (no synth)	0.6554	0.5971
Llama 8B + 5% synth	**0.6271**	**0.6234**

Table 2: System comparison. Best test result in bold.

6.2 Synthetic Data Ablation

Table 3 shows the critical sensitivity to synthetic ratio.

Synth Ratio	Val F1	Test F1
0%	0.6554	0.5971
5%	**0.6271**	**0.6232**
10%	0.5499	0.5851

Table 3: Effect of synthetic data ratio on Llama 8B.

To understand why 5% transferred best, we compared test prediction distributions for the Llama 8B models in Table 4. The 5% model reduced Non-toxic predictions and increased predictions for Classes 2 and 3, the confusable minority categories most affected by the train/test annotation shift. Higher synthetic ratios did not preserve this balance in class-level decisions and reduced test F1.

Prediction	0% synth	5% synth	10% synth
Class 0: Non-toxic	79.6%	79.0%	78.7%
Class 1: Insults	14.8%	14.3%	13.9%
Class 2: Other	4.9%	5.7%	6.6%
Class 3: Hate	0.6%	0.7%	0.6%
Test F1	0.5971	**0.6232**	0.5851

Table 4: Test prediction distribution for Llama 8B synthetic-data variants.

6.3 Per-class Performance

Table 5 shows per-class F1 for the final submitted system on the provided development split and the released test labels. Performance correlates roughly with class frequency, with Class 2 (Other Offensive) and Class 3 (Hate/Harassment) being particularly challenging.

Class	Dev F1	Test F1
0: Non-toxic	0.95	0.94
1: Insults/Flaming	0.75	0.74
2: Other Offensive	0.47	0.44
3: Hate/Harassment	0.45	0.43
4: Threats	0.57	0.33
5: Extremism	0.57	0.86

Table 5: Per-class F1 for the final submitted system.

7 Analysis

7.1 The Validation Trap

Our most significant finding is the "validation trap": models achieving high validation F1 through conservative predictions (matching the 81% Non-toxic distribution) performed poorly on test. Evidence includes:

- Gemma 12B: 0.66 val $\rightarrow$ 0.52 test
- Transfer learning: 0.68 val $\rightarrow$ 0.55 test
- Two-stage: 0.67 val $\rightarrow$ 0.47 test

Models predicting more minority classes (2, 3) performed better on test, suggesting different annotation patterns between splits.

7.2 Why 5% Synthetic Works

The 5% ratio appears to increase minority class predictions without overwhelming original patterns. The distribution analysis in Table 4 supports this interpretation: relative to the no-synthetic Llama 8B model, the 5% model predicts fewer Non-toxic messages and more Class 2/3 messages, which improves test transfer. Higher synthetic ratios did not yield the same class-level accuracy: the 10% model shifted predictions further toward Class 2 but lost roughly 0.038 test F1, suggesting that excessive synthetic data can reinforce artifacts or shift the model away from the test annotation pattern.

7.3 Error Analysis

Common error patterns include:

- Confusion between Class 1 (Insults) and Class 2 (Other Offensive)
- Multilingual messages misclassified as Non-toxic
- Gaming slang incorrectly flagged as toxic

8 Conclusion

We presented a comprehensive exploration of approaches for gaming toxicity detection. Key findings:

1. Llama 3.1 8B outperformed both smaller and larger models
2. Synthetic data has a narrow sweet spot (5%)
3. Validation metrics can be misleading due to distribution shift
4. Ensembles don't help when one model dominates

Our system achieves 0.6234 F1-macro, placing 4th out of 35 teams. Future work could explore better handling of distribution shift and external gaming-specific data.

Limitations

Our analysis is limited to this specific dataset. The "validation trap" phenomenon may be dataset-specific and not generalize. Computational constraints limited exploration of larger models and longer training. The synthetic data approach requires access to commercial LLM APIs.

Ethics Statement

This work involves detecting toxic content in gaming chat. Models could potentially be misused to generate toxic content or for surveillance. We advocate for responsible deployment in content moderation systems with human oversight, transparency

about automated decisions, and appeal mechanisms for users.

References

Aashish Bhandari, Siddhant B Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. CrisisHateMM: Multimodal analysis of directed and undirected hate speech in text-embedded images from Russia–Ukraine conflict. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1994–2003.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 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*, pages 4171–4186.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

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*.

Ali Hürriyetoğlu, Surendrabikram Thapa, Hristo Tanev, Laxmi Thapa, and Surabhi Adhikari. 2026. Overview of the workshop on event extraction and understanding: Challenges and applications. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Haewoon Kwak, Jeremy Blackburn, and Seungyeop Han. 2015. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3739–3748.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.

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. *arXiv preprint arXiv:1907.11692*.

Llama Team, AI @ Meta. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Usman Naseem, Shuvam Shiwakoti, Siddhant Bikram Shah, Surendrabikram Thapa, and Qi Zhang. 2025. GameTox: A comprehensive dataset and analysis for enhanced toxicity detection in online gaming communities. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 440–447.

Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. 2021. Hate speech detection using natural language processing: Applications and challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1302–1308. IEEE.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025. Large language models (llm) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*, 15(1):4.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Kritesh Rauniyar, Laxmi Thapa, Surabhi Adhikari, Kristina T. Johnson, Ali Hürriyetoğlu, Hristo Tanev, and Usman Naseem. 2026. Understanding toxic behavior in gaming communities using ai to promote healthier digital spaces. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

A Full Test Performance

Table 6 reports the full test-set classification report for the final submitted system. These scores were computed after the official test labels were released, using the submitted predictions that achieved 0.6234 macro-F1.

Class	Precision	Recall	F1	Support
0: Non-toxic	0.9620	0.9242	0.9427	4351
1: Insults/Flaming	0.7563	0.7318	0.7438	742
2: Other Offensive	0.3396	0.6128	0.4370	235
3: Hate/Harassment	0.4103	0.4444	0.4267	36
4: Threats	0.3000	0.3750	0.3333	8
5: Extremism	0.7500	1.0000	0.8571	3
Macro average	0.5864	0.6814	0.6234	5375
Weighted average	0.9016	0.8800	0.8887	5375

Table 6: Full test-set classification report for the final submitted system.

B Additional Experimental Results

Table 7 summarizes additional systems and ablations explored during development. The pattern reinforces the main paper's validation-trap finding: several systems improved validation F1 but transferred poorly to the test set, while the final Llama 8B system with a small amount of synthetic data gave the best test performance.

System	Val F1	Test F1	Notes
Zero-shot GPT-4o-mini	0.4630	0.4126	Direct prompting; over-predicted minority classes
Two-stage Gemma 2B	0.6749	∼0.47	Binary toxic detector plus toxic-only classifier
Gemma 2B	0.63	∼0.52	Single-stage LoRA baseline
Gemma 12B	0.662	∼0.52	Higher validation F1 but conservative test predictions
Prompted ensemble	0.6201	0.5762	Average of prompted 2B models
Multi-step ensemble	0.6280	0.5810	Confidence-based routing
Gemma 2B train-all	–	0.5898	Trained on combined train and validation data
Llama 8B, no synthetic	0.6554	0.5971	Best single model before augmentation
Llama 8B + 10% synthetic	∼0.65	0.5851	Higher synthetic ratio hurt transfer
Transfer DOTA2 → GameTox	0.6815	∼0.55	Gaming-domain pretraining caused validation trap
Llama 8B + 5% synthetic	0.6271	0.6232	Best unboosted model
Final Class 2 boost	–	**0.6234**	Official submitted system

Table 7: Additional systems and ablations evaluated during development.