syuhhh@EEUCA 2026: A Three-Stage Progressive Training Framework for Fine-Grained Toxicity Detection in Online Gaming Communities

Yuhao Shi
School of Computer Science and Technology
Tongji University
syhhh@tongji.edu.cn

Yu Wang*
School of Computer Science and Technology
Tongji University
csyuwang@tongji.edu.cn

Shengjie Zhao*
School of Computer Science and Technology
Tongji University
shengjiezhao@tongji.edu.cn

Abstract

This paper presents our 1st-place system for the Shared Task on Fine-Grained Toxicity Detection in Online Gaming (GameTox) at the 9th EEUCA Workshop, co-located with ACL 2026. The task targets 6-class fine-grained toxic intent classification on the official GameTox dataset, comprising 53,000 real-world *World of Tanks* chat utterances. We propose a three-stage progressive training framework built on XLM-RoBERTa-large: (1) gaming domain adaptive MLM pre-training, (2) multilingual toxicity transfer fine-tuning, and (3) supervised contrastive learning (SCL)-enhanced target task tuning. We further incorporate LLM-driven data augmentation and long-tailed class synthesis. Our system achieves a Macro F1 of **0.7041**, ranking 1st among 35 teams. Ablation studies validate each module's contribution, and we release our code to facilitate follow-up research.

1 Introduction

Online gaming has become a dominant form of global digital social interaction, with billions of users engaging in real-time chat daily. However, the anonymity of in-game environments enables the proliferation of toxic behaviors—insults, harassment, threats, and extremist speech—causing significant harm to user well-being and platform governance (Parihar et al., 2021). Unlike toxicity on mainstream social media, in-game chat utterances are ultra-short, filled with domain-specific slang, abbreviations, and highly informal expressions, rendering general toxicity detection models ineffective at capturing implicit fine-grained toxic intent (Naseem et al., 2025).

This work addresses the GameTox Shared Task (Thapa et al., 2026) at the 9th EEUCA Workshop (Hürriyetoğlu et al., 2026). The task is a 6-class single-label classification problem on the GameTox dataset (Naseem et al., 2025): Non-toxic (0), Insults and Flaming (1), Other Offensive Texts (2), Hate and Harassment (3), Threats (4), and Extremism (5), following the annotation schema from Bhandari et al. (2023). Systems are ranked by Macro F1-score, which assigns equal weight to all categories, emphasizing low-resource high-risk classes.

Three core technical challenges motivate our work: (i) **Domain shift**—standard PLMs are pre-trained on formal long-form corpora, producing insufficient feature extraction for game chat's semantic sparsity; (ii) **Extreme long-tailed distribution**—over 70% of samples are Non-toxic while three high-risk minority classes account for less than 5% combined, causing models to favor majority classes; (iii) **Blurred category boundaries**—distinguishing semantically adjacent categories (e.g., Insults vs. Other Offensive) requires intent-level feature discrimination beyond surface keywords.

To address these challenges, we propose a three-stage progressive training framework on XLM-RoBERTa-large, incorporating domain-adaptive MLM pre-training, multilingual toxicity transfer, and SCL-enhanced fine-tuning with a dual-head architecture. We further introduce LLM-driven data augmentation and long-tailed class synthesis. Our final system achieves Macro F1 = 0.7041, outperforming the best competitor by 3.16 absolute points and the vanilla XLM-RoBERTa-large baseline by 8.92 points. Our code is publicly available.[1]

2 Task, Dataset & Related Work

2.1 Task Definition

GameTox (Thapa et al., 2026) is a 6-class single-label classification task. Given a game chat utterance, the model predicts its toxic intent label (0–5).

*Co-corresponding authors.

[1] https://github.com/oosyh/syuhhh-EEUCA2026

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 161–168

The official evaluation metric is Macro F1-score, which neutralizes the majority-class bias of standard accuracy metrics and places equal emphasis on rare but high-risk toxic categories.

2.2 Dataset Overview

The GameTox dataset (Naseem et al., 2025) contains 53,000 human-annotated utterances from *World of Tanks*—the largest and most fine-grained gaming toxicity benchmark to date. Two characteristics pose critical challenges: (1) extreme long-tailed distribution (70%+ Non-toxic; minority toxic classes <5% combined); and (2) severe semantic sparsity (average utterance length: 12 tokens, dense with in-game slang and abbreviations).

2.3 Additional Data Resources

We use two types of publicly available resources beyond the official training set:

Gaming Domain Corpus (Stage 1). We construct a combined corpus from: (1) the public Dota 2 in-game chat dataset (Raman et al., 2021); (2) a self-constructed multi-game balanced chat corpus; and (3) Twitter toxic comment datasets from gaming communities (Davidson et al., 2017). This corpus aligns the model's vocabulary and syntax representations with game chat scenarios.

Jigsaw Multilingual Dataset (Stage 2). We use the Jigsaw 2018 Toxic Comment dataset and its multilingual translations (Jigsaw/Conversation AI, 2018, 2020), covering 5 languages with 6-dimensional toxicity labels—highly consistent with our target task. For English, we retain all toxic samples and downsample non-toxic to 100,000; for non-English, we retain all toxic and sample 10% of non-toxic to maintain cross-lingual context without diluting toxic signal.

2.4 Related Work

Toxicity detection has evolved from manual feature engineering to PLM fine-tuning, with XLM-RoBERTa establishing strong baselines across multilingual toxic benchmarks (Parihar et al., 2021). For gaming toxicity, prior work has highlighted domain shift as the primary limiting factor, with GameTox being the most comprehensive benchmark (Naseem et al., 2025).

Intent-aware modeling has demonstrated effectiveness in related structured prediction tasks. Wang and Zhao (2026) show that modeling behavioral intention via anomaly-connected components substantially improves fine-grained detection under weak supervision—a finding that motivates our intent-level feature discrimination approach. Building on this line of work, event completeness modeling and local semantic signal extraction have further advanced weakly-supervised video understanding (Wang and Chen, 2026; Wang et al., 2026). Semantic query-based and action-semantic consistent approaches to temporal localization (Wang et al., 2025, 2024) similarly demonstrate that aligning feature representations with intent-level semantics is critical for fine-grained categorization under annotation constraints, a principle we adopt in our contrastive learning design.

Supervised contrastive learning (SCL) has proven effective for imbalanced classification by directly optimizing feature space structure; few-shot recognition approaches (Liu et al., 2026) further demonstrate the value of semantic-temporal representation for low-resource scenarios analogous to our minority toxic classes. LLM-based semantic augmentation has also addressed short-text sparsity in low-resource scenarios (Thapa et al., 2025). Cross-domain collaborative modeling with spatio-temporal fusion (Wang et al., 2022) additionally inspires our multi-stage training pipeline design.

3 System Methodology

Our framework (Figure 1) systematically aligns XLM-RoBERTa-large from general language understanding to game chat domain, then to fine-grained toxic intent classification.

3.1 Backbone: XLM-RoBERTa-large

We select XLM-RoBERTa-large for three reasons: (1) pre-training on 2.5T tokens across 100+ languages naturally supports multilingual game chat and cross-lingual transfer; (2) it is the state-of-the-art backbone for toxicity detection, with superior informal-text feature extraction; (3) its transformer architecture is fully compatible with MLM pre-training, classification fine-tuning, and SCL.

3.2 Stage 1: Gaming Domain Adaptive MLM Pre-training

Motivation. PLMs pre-trained on formal corpora suffer significant domain shift on game chat texts—ultra-short, slang-rich, and abbreviation-dense. Stage 1 adapts the model to game chat's unique linguistic distribution.

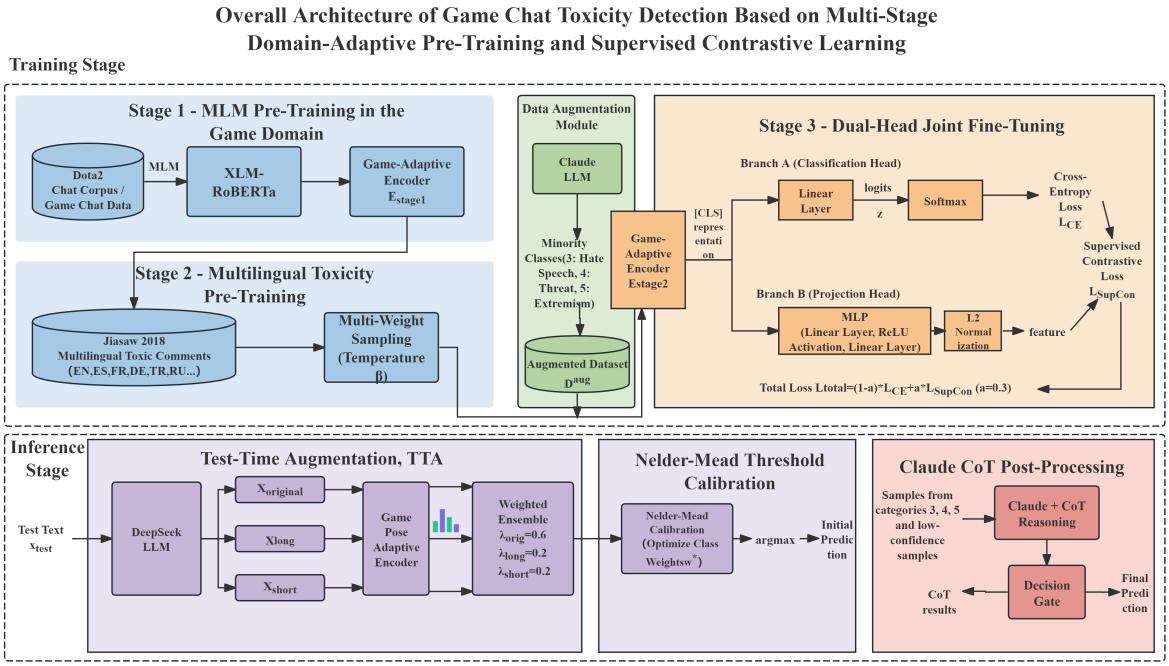

Figure 1: Overall architecture of our three-stage progressive training framework. The pipeline covers domain adaptive pre-training (Stage 1), multilingual toxicity transfer fine-tuning (Stage 2), SCL-enhanced end-to-end fine-tuning (Stage 3), and targeted optimization strategies.

Implementation. We perform standard MLM (masking probability 0.15) on our combined gaming corpus. Key configuration is in Table 1. We set the max sequence length to 128, well-suited to short game utterances, enabling the model to capture domain-specific slang and syntax without excessive padding.

Table 1: Gaming domain MLM pre-training configuration.

Hyperparameter	Value
Backbone	XLM-RoBERTa-large
Max Seq Length	128
MLM Masking Prob	0.15
Global Batch Size (4 GPUs)	128
Learning Rate	2e-5
Training Epochs	3
Optimizer	AdamW (λ=0.01)
Mixed Precision	FP16

3.3 Stage 2: Multilingual Toxicity Transfer Fine-tuning

Motivation. The limited and imbalanced GameTox training set risks over-fitting on majority classes. Stage 2 injects generalizable toxic semantic knowledge via large-scale multilingual supervision before target task adaptation.

Implementation. We fine-tune on the mixed Jigsaw multilingual dataset (Table 2) using multi-label binary cross-entropy loss, with max sequence length 224 (compatible with longer Jigsaw samples), learning rate 2e-5, 2 training epochs, and warmup ratio 0.1. Model selection is based on validation Macro AUC.

3.4 Stage 3: SCL-Enhanced End-to-End Fine-tuning

This stage directly targets the 6-class GameTox classification. A dual-head model (Figure 1) processes each utterance through: (1) the XLM-RoBERTa **encoder** producing a [CLS] pooled embedding; (2) a **Classification Head** (linear layer, 6-class logits); and (3) a **Projection Head** (2-layer MLP with ReLU, 128-dim normalized embedding for SCL).

Joint Loss Function. We combine class-balanced cross-entropy and supervised contrastive loss:

$$\mathcal{L}_{\text{total}} = (1-\alpha)\,\mathcal{L}_{\text{CE}} + \alpha\,\mathcal{L}_{\text{SCL}}, \quad \alpha = 0.3 \quad (1)$$

$\mathcal{L}_{\text{CE}}$ uses class weights inversely proportional to label frequency, counteracting long-tailed bias. $\mathcal{L}_{\text{SCL}}$

Table 2: Data distribution of the multilingual transfer fine-tuning dataset.

Language	Original Samples	Toxic Samples	Final Retained
English (Anchor)	159,571	16,225	116,225
Russian	159,571	16,225	30,560
Turkish	159,571	16,225	30,560
Spanish	159,571	16,225	30,560
French	159,571	16,225	30,560
Total	797,855	81,125	238,465

with temperature τ=0.07 is:

$$\mathcal{L}_{\text{SCL}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\frac{e^{\langle z_i, z_p\rangle/\tau}}{\sum_{a\in A(i)} e^{\langle z_i, z_a\rangle/\tau}} \tag{2}$$

where z_i is the normalized projection embedding of sample i, $P(i)$ the set of same-label samples, and $A(i)$ all other samples in the batch.

Training Configuration. Max sequence length: 96 (suited to ultra-short utterances); batch size: 32 per device; epochs: 5; layer-wise LR: 1e-5 (encoder), 5e-5 (heads); AdamW with weight decay 0.01; linear warmup (10%) with linear decay; gradient clipping at 1.0. Validation set: 10% of training data via stratified split.

3.5 Targeted Optimization Strategies

LLM-Driven Short Text Augmentation. We use an LLM to enrich ultra-short game chat samples by adding plausible game scenario context while preserving the original toxic intent, bridging the gap between 12-token utterances and the longer inputs expected by Stage 2's pre-trained representations. The exact prompt template is provided in Appendix A.

Long-Tailed Class Data Synthesis. For minority categories (classes 3–5), we use an LLM API to generate high-quality synthetic samples with prompts conditioned on semantic features, toxicity type, and game context. To ensure quality, generated samples are manually spot-checked and filtered to remove semantically inconsistent or out-of-distribution instances.

Threshold Optimization. Post-training, we apply the Nelder-Mead algorithm on validation predictions to optimize per-class decision thresholds, directly maximizing Macro F1 and correcting residual majority-class prediction bias.

Minority-Focused Model Ensemble. To further improve coverage on low-resource toxic categories (classes 3–5), we construct a targeted three-component ensemble. For minority-class predictions, we fuse the outputs of our three-stage XLM-RoBERTa system, ToxicBERT (Caselli et al., 2020), and LLM-generated classification results on minority-class samples, with each component weighted by its per-class F1 on the validation set. For majority classes (0–2), we retain the predictions of our primary XLM-RoBERTa system directly, avoiding ensemble dilution on well-represented categories. Importantly, LLM inference for the ensemble is conducted *offline* in batch mode prior to final prediction assembly, rather than in real-time; this design avoids deployment latency while retaining the classification signal from the LLM. The exact prompt used for this classification step is provided in Appendix A.

4 Experimental Setup

Environment. All experiments are implemented with PyTorch (Paszke et al., 2019) and Hugging Face Transformers (Wolf et al., 2020) on 4 $\times$ NVIDIA 3090 24GB GPUs with FP16 mixed precision. Core versions: PyTorch 2.0.1, Transformers 4.36.0, scikit-learn 1.3.0.

Dataset Split. We use all 53,000 official training samples; 10% are held out as a stratified validation set preserving the original class distribution. No test labels are accessed during development.

Baselines. We compare against: BERT-base-uncased (Devlin et al., 2019) (standard English PLM), HateBERT (Caselli et al., 2020) (toxicity-domain pre-trained BERT), DeBERTaV3-base, and **Vanilla XLM-RoBERTa-large** (our core baseline: identical backbone fine-tuned directly on GameTox without our framework).

Evaluation. The official metric is Macro F1-score. We additionally report accuracy, macro precision, and macro recall. All ablation results use the same validation split.

5 Results and Analysis

5.1 Main Results

Table 3 shows that our final system achieves a Macro F1 of 0.7041, outperforming the vanilla XLM-RoBERTa-large baseline by **+8.92 pp** and the best competitor (Macro F1: 0.6725) by **+3.16 pp**. The simultaneous gains in precision (+9.16 pp) and recall (+7.19 pp) confirm that our framework improves both minority-class coverage and prediction quality, rather than trading one for the other.

5.2 Ablation Study

Table 4 yields three key observations:

(1) Domain alignment and toxic transfer provide the largest gain (+10.37 pp). The combined Stage 1+2 pre-training dramatically outperforms direct fine-tuning, confirming that bridging the domain gap between general corpora and game chat slang is the most critical factor for performance.

(2) LLM-driven data augmentation is the second-largest contributor (+4.71 pp). Short text semantic augmentation effectively resolves semantic sparsity, producing richer input representations without altering toxic intent. Long-tailed synthesis adds a further +1.90 pp by alleviating critical minority-class data shortage.

(3) SCL and ensemble further lift the performance ceiling (+2.31 pp combined). The dual-head contrastive structure enhances inter-class feature discriminability, directly addressing blurred category boundaries. The minority-focused three-component ensemble improves robustness on low-resource toxic categories and pushes the final score to 0.7041.

5.3 Error Analysis

Although our system achieves the highest Macro F1 of 0.7041, residual errors concentrate around two empirically observed confusion patterns.

Insults and Flaming vs. Hate and Harassment (classes 1 & 3). The most frequent misclassifications occur between generalized insults and targeted hate speech, particularly for utterances containing identity-related slurs (e.g., homophobic terminology). Such expressions can simultaneously function as casual in-game taunts (class 1) or constitute directed identity-based harassment (class 3), and the distinction hinges on pragmatic intent that is difficult to infer from a single decontextualized utterance. Without speaker interaction history, the model tends to under-predict class 3, biasing toward the more frequent class 1.

Non-toxic vs. Extremism (classes 0 & 5). A secondary error pattern arises between ostensibly benign utterances and low-intensity extremist expressions. Utterances with mild political overtones—such as vague ideological statements or dog-whistle phrasing common in certain gaming communities—are frequently misclassified as Non-toxic (class 0) because they lack overt surface-level toxic markers. This reflects the fundamental challenge that extremism detection requires pragmatic and world-knowledge reasoning beyond lexical toxicity signals.

Both patterns underscore that fine-grained intent discrimination in ultra-short game chat demands conversational context modeling and intent-aware reasoning (Wang and Zhao, 2026), which we identify as the primary direction for future improvement.

6 Conclusion

We present a three-stage progressive training framework for fine-grained gaming toxicity detection that systematically addresses domain shift, long-tailed imbalance, and semantic sparsity. Our pipeline—domain adaptive MLM pre-training, multilingual toxicity transfer, and SCL-enhanced fine-tuning—combined with LLM-driven augmentation and ensemble, achieves Macro F1 = 0.7041, ranking 1st among 35 teams. Ablation studies confirm that each module contributes meaningfully, with domain alignment and toxic knowledge transfer delivering the largest gains. Future work will explore conversational context modeling, adversarial training against camouflaged toxic expressions, and model distillation for real-time deployment.

Limitations

Several limitations of the current work should be acknowledged. First, our system relies on LLM-generated synthetic data for minority class augmentation; while generated samples are manually spot-checked for quality, they may still introduce distributional artifacts not present in real game chats. Second, the gaming domain corpus used in Stage 1

Table 3: Main results on the official GameTox test set. Our system ranks 1st among 35 teams.

Model	Macro F1	Accuracy	Macro Prec.	Macro Rec.
BERT-base-uncased (Devlin et al., 2019)	0.6043	0.8936	0.5487	0.6984
HateBERT (Caselli et al., 2020)	—	—	—	—
DeBERTaV3-base	0.5561	0.8874	0.5349	0.6006
Vanilla XLM-RoBERTa-large (Core Baseline)	0.6149	0.8865	0.5484	0.7267
Our Final System	**0.7041**	**0.8982**	**0.6400**	**0.7986**

Table 4: Incremental ablation study. Each row adds one component to the previous configuration.

Model Configuration	Macro F1	Δ
XLM-RoBERTa-large (no MLM pre-train, no Transfer)	0.4986	—
+ Domain MLM Pre-training + Jigsaw Transfer	0.6023	+0.1037
+ Single-Head CE Fine-tuning (Core Baseline)	0.6149	+0.0126
+ LLM Short Text Augmentation	0.6620	+0.0471
+ LLM Long-Tailed Class Synthesis	0.6810	+0.0190
+ Dual-Head SCL + Threshold Optimization	0.6950	+0.0140
+ Multi-Model Ensemble (**Final System**)	**0.7041**	+0.0091

is drawn from limited game titles (primarily Dota 2 and *World of Tanks*), which may not fully capture the linguistic diversity of all gaming communities. Third, the current model processes single utterances without conversational context; implicit toxicity that depends on dialogue history may be misclassified. Finally, the Nelder-Mead threshold optimization is tuned on the validation split, and may not generalize perfectly to distribution shifts in unseen test data.

Acknowledgments

This work was supported in part by the National Key Research and Development Project under Grant 2023YFC3806000, in part by the National Natural Science Foundation of China under Grant 62406226, in part by the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0100, in part sponsored by Shanghai Sailing Program under Grant 24YF2748700, in part by New-Generation Information Technology under the Shanghai Key Technology R&D Program under Grant 25511103500.

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*, pages 1994–2003.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms*, pages 21–32.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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)*.

Jigsaw/Conversation AI. 2018. Jigsaw toxic comment classification challenge. Kaggle Competition.

Jigsaw/Conversation AI. 2020. Jigsaw multilingual toxic comment classification. Kaggle Competition.

Hongli Liu, Yu Wang, and Shengjie Zhao. 2026. STAR: Semantic-temporal adaptive representation learning for few-shot action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*.

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, Mexico City, Mexico. Association for Computational Linguistics.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32.

Shailesh Raman, Birju Patel, Sriram Srinivasan, Pnina Shachaf, and David Jurgens. 2021. Chat as currency: Linguistic features of toxicity in online gaming. In *Proceedings of the 2021 ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 1–14.

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):1–30.

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)*.

Yu Wang and Shiwei Chen. 2026. Learning event completeness for weakly supervised video anomaly detection. In *Proceedings of the 43rd International Conference on Machine Learning*, pages 62505–62517.

Yu Wang and Shengjie Zhao. 2026. Weakly supervised video anomaly detection with anomaly-connected components and intention reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yu Wang, Shengjie Zhao, and Shiwei Chen. 2024. Action-semantic consistent knowledge for weakly-supervised action localization. *IEEE Transactions on Multimedia*, 26:10279–10289.

Yu Wang, Shengjie Zhao, and Shiwei Chen. 2025. SQL-Net: Semantic query learning for point-supervised temporal action localization. *IEEE Transactions on Multimedia*, 27:84–94.

Yu Wang, Shengjie Zhao, Jianyu Wang, and Xutao Chu. 2026. Learning local semantic signals and inter-class discrepancy for weakly supervised video anomaly detection. *IEEE Transactions on Multimedia*.

Yu Wang, Shengjie Zhao, Rongqing Zhang, Xiang Cheng, and Liuqing Yang. 2022. Multi-vehicle collaborative learning for trajectory prediction with spatio-temporal tensor fusion. *IEEE Transactions on Intelligent Transportation Systems*, 23(1):236–248.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clément Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

A Prompt Templates

A.1 Short-Text Augmentation Prompt

The following system prompt is used to rewrite ultra-short game chat messages into longer, contextually richer forms while preserving the original toxic intent.

```
You are a Data Augmentation Specialist for a toxicity detection system. The downstream model was pre-trained on Jigsaw/Wikipedia Comments (long, grammatically structured sentences). The current input data is Game Chat (short, slang-heavy, multilingual, noisy).

YOUR GOAL: Rewrite the input game chat message into a "Jigsaw-style Comment". 1. Expand Length: Turn abbreviations into full words. Elaborate slightly to make it a complete sentence. 2. Standardize English: Translate any non-English text to explicit English. 3. PRESERVE TOXICITY (Crucial): - If input is TOXIC ("kys"), output MUST be equally TOXIC. - If input is NON-TOXIC ("gg"), output MUST be NON-TOXIC. - Do NOT sanitize or censor.

EXAMPLES:
Input: "stfu noob"
Output: You need to shut the fuck up, you are playing like a complete beginner.

Input: "ez"
Output: That match was too easy, you guys didn't even provide a challenge.
```

```
Input: “go A”
Output: Let’s all move to point A and
take the objective.

OUTPUT FORMAT: Return ONLY the rewritten
text.
```

A.2 Minority Class Classification Prompt

The following system prompt is used for LLM-based classification of minority-class samples in the ensemble component.

```
You are an expert content moderator
for an online game (World of Tanks).
Your task is to classify chat messages
into EXACTLY ONE of the following 6
categories.

LABELS & DEFINITIONS:
0: Non-toxic (Normal gameplay
communication, tactics, simple
frustration)
1: Insults (Personal attacks, “idiot”,
“noob”, mild profanity directed at
someone)
2: Other Offensive (General profanity
not directed at anyone, “fuck this
game”)
3: Hate Speech (Slurs based on race,
gender, religion, sexual orientation)
4: Threats (Physical violence, “I will
kill you”, “hope you get cancer”)
5: Extremism (Nazi symbols, terrorist
propaganda, glorifying violence)

RULES:
- OUTPUT ONLY THE INTEGER LABEL (0-5).
NO EXPLANATION.
- If a message contains multiple types,
pick the MOST SEVERE one (5 > 4 > 3 > 1
> 2 > 0).

User prompt format: Message: “[text]”
Label:
```