Understanding Toxic Behavior in Gaming Communities Using AI to Promote Healthier Digital Spaces

Surendrabikram Thapa[1], Shuvam Shiwakoti[1], Siddhant Bikram Shah[2], Kritesh Rauniyar[3], Laxmi Thapa[4], Surabhi Adhikari[5], Kristina T. Johnson[2], Ali Hürriyetoğlu[6], Hristo Tanev[7], Usman Naseem[3]
[1]Virginia Tech, USA, [2]Northeastern University, USA,
[3]Macquarie University, Australia, [4]O.P. Jindal Global University, India
[5]Columbia University, USA, [6]Wageningen Food Safety Research, Netherlands,
[7]European Commission, Joint Research Centre, Italy
[1]{surendrabikram, shuvam}@vt.edu, [2]rauniyark11@gmail.com,
[6]ali.hurriyetoglu@wur.nl, [7]hristo.tanev@ec.europa.eu

Abstract

Online gaming communities are increasingly affected by toxic communication, including harassment, threats, hate speech, and extremist content. Detecting such behavior is challenging due to the short, noisy, multilingual, and highly imbalanced nature of gaming chat data. To advance research in this area, we organized the Shared Task on Fine-Grained Toxicity Detection in Online Gaming at EEUCA 2026, co-located with ACL 2026. The task is based on the GameTox dataset, containing approximately 53,000 annotated chat utterances from *World of Tanks* across six toxicity categories. A total of 102 participants took part, and 35 teams submitted systems exploring approaches such as domain-adaptive pretraining, multilingual transfer learning, contrastive learning, LLM-based augmentation, and ensemble methods. Systems were evaluated using macro-averaged F1-score, with the top system achieving 0.7041 Macro F1. This paper presents an overview of the shared task, dataset, evaluation framework, participant methods, and key findings.

1 Introduction

Online multiplayer gaming has become one of the most prevalent forms of digital social interaction, with billions of users worldwide engaging in real-time chat communication during gameplay (Crawford et al., 2013). Yet beneath the coordinated team play and casual banter, in-game chat channels also serve as fertile ground for some of the most toxic forms of online behavior: insults, harassment, identity-based hate speech, threats, and even extremist messaging (Naseem et al., 2025; Sanghvi et al., 2024). The anonymity afforded by gaming usernames, the high-stakes emotional intensity of competitive play, and the perceived ephemerality of chat exchanges combine to lower the threshold for toxic behavior, with measurable consequences for player well-being, community health, and platform governance (Wells et al., 2025).

Detecting toxic content in gaming chat presents a distinct set of computational challenges that distinguish it from toxicity detection on mainstream social media platforms. First, in-game chat utterances are characteristically short: messages average roughly twelve tokens in length and are densely populated with domain-specific slang, abbreviations, and obfuscated spellings that general-purpose pretrained language models struggle to interpret (Naseem et al., 2025). Second, toxicity in gaming spans a wide spectrum of severity and intent, ranging from casual flaming and competitive trash-talk to identity-based harassment, explicit threats, and extremist incitement—categories that demand fine-grained, multi-class differentiation rather than a coarse binary judgment. Third, gaming chat is heavily multilingual, with utterances in English, Russian, Polish, German, French, Spanish, and many other languages frequently appearing within a single match, often interleaved with code-switching and transliterations. Finally, real-world gaming toxicity datasets exhibit extreme long-tailed class distributions, with non-toxic communication dominating the corpus while the most consequential toxic categories—hate speech, threats, and extremism—collectively account for less than five percent of training samples.

These compounding challenges render general-purpose toxicity classifiers largely ineffective when transferred to the gaming domain. Models pretrained on formal text corpora suffer significant domain shift when applied to game chat's semantic sparsity and informal register; ensemble methods designed for high-resource settings can fail catas-

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 8–16

trophically under extreme data scarcity for minority classes; and evaluation metrics that prioritize overall accuracy obscure systemic failures on the rare-but-high-risk toxic categories that matter most for content moderation. Addressing these issues requires concerted research effort, robust benchmarks, and methodological innovations specifically tailored to the characteristics of gaming communication.

To advance research in this critical area, we present the **Shared Task on Fine-Grained Toxicity Detection in Online Gaming**, organized as part of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA 2026) (Hürriyetoğlu et al., 2026), co-located with ACL 2026. The task is built on the GameTox dataset (Naseem et al., 2025), comprising approximately 53,000 manually annotated chat utterances drawn from the online multiplayer game World of Tanks. Participating systems are required to classify each utterance into one of six fine-grained intent categories: *Non-Toxic* (0), *Insults and Flaming* (1), *Other Offensive Texts* (2), *Hate and Harassment* (3), *Threats* (4), and *Extremism* (5). The annotation schema, adapted from the CrisisHateMM framework (Bhandari et al., 2023), captures the gradient of toxic intent that characterizes real-world gaming communication. Following standard practice for imbalanced classification, systems are evaluated using macro-averaged F1-score, which assigns equal weight to all six categories and emphasizes performance on the high-risk minority classes that are most relevant to platform safety.

The shared task attracted strong participation, with 35 teams submitting systems exploring a wide range of approaches for fine-grained toxicity detection in gaming environments. Participating methods included transformer-based encoders, multilingual transfer learning, domain-adaptive pretraining, contrastive learning, large language model (LLM)-based augmentation, ensemble methods, and specialized architectures designed to address severe class imbalance and rare toxic categories. The diversity of submissions highlights both the growing interest in gaming toxicity detection and the methodological challenges posed by short, noisy, multilingual, and highly imbalanced chat data. This paper provides a comprehensive overview of the shared task, including a detailed description of the GameTox dataset and its annotation methodology, the evaluation protocol, summaries of the participating systems and their methodologies, and an analysis of the results. Through this shared task, we aim to advance the state of fine-grained toxicity detection in online gaming, foster methodological innovation under realistic class-imbalance and domain-shift conditions, and contribute to the development of more reliable AI-driven systems for promoting healthier digital spaces in gaming communities.

2 Related Works

Research on toxicity in gaming communities has progressively shifted from broad online-abuse frameworks toward examining the emergence of hostile communication within competitive multiplayer environments (Munn, 2023; Zsila et al., 2022). Recent work demonstrates that toxic behavior in games encompasses verbal harassment, hate speech, exclusionary conduct, and related hostile practices that diminish player enjoyment, harm psychological well-being, and normalize abusive interaction within gaming cultures (Wells et al., 2025; Zsila and Demetrovics, 2025). However, much of this literature remains sociological or psychological rather than computational, offering richer accounts of the causes and consequences of toxicity than deployable NLP methods for identifying fine-grained toxic intent in chat data (Munn, 2023; Zsila et al., 2022).

Early computational work on gaming toxicity demonstrated that supervised models could predict crowdsourced moderation decisions in League of Legends; however, such work largely treated toxicity as a coarse moderation outcome rather than a fine-grained taxonomy of harmful intents (Blackburn and Kwak, 2014). More recent gaming-specific research has begun to address this limitation by focusing directly on in-game chat, where utterances are short, noisy, context-dependent, and shaped by gaming slang (Naseem et al., 2025; Tereshchenko and Hämäläinen, 2025). GameTox is particularly relevant as it introduces a large-scale gaming-chat dataset annotated for toxicity detection via intent classification and slot filling, thereby advancing the field beyond coarse binary toxicity labels (Naseem et al., 2025). Nevertheless, existing gaming-focused datasets and systems still leave open challenges related to rare toxic classes, multilingual variation, context dependence, and robustness under severe class imbalance.

In broader NLP, transformer-based language models such as BERT (Devlin et al., 2019),

RoBERTa (Liu et al., 2019), XLM-R (Conneau et al., 2020), and HateBERT (Caselli et al., 2021) have become foundational to abusive-language detection, offering robust contextual, multilingual, and domain-adapted representations. Yet these models do not fully resolve gaming-toxicity detection, as strong performance on general abusive-language benchmarks does not reliably transfer to short, code-mixed, slang-heavy, and class-imbalanced game-chat data (Naseem et al., 2025; Caselli et al., 2021). Bias and explainability research further demonstrates that toxicity classifiers are susceptible to encoding social biases and may require rationales, target labels, or audit mechanisms to support equitable and transparent moderation decisions (Mathew et al., 2021; Sap et al., 2019; Rauniyar et al., 2023). This shared task is positioned at the intersection by providing a focused evaluation setting for fine-grained gaming-toxicity detection and by encouraging systems that explicitly address domain shift, rare harmful categories, and realistic class imbalance rather than relying on coarse binary toxicity classification.

3 Shared Task Description

This shared task focuses on the automated detection of toxic behavior in online gaming communities, targeting the classification of player intent in game chat utterances.

Task: Intent Classification. Given a game chat utterance, participating systems must classify its intent into one of six categories: *Non-toxic*, *Hate and Harassment*, *Threats*, *Extremism*, *Insults and Flaming*, and *Other Offensive Texts*. *Non-toxic* utterances correspond to normal, benign communication between players. *Hate and Harassment* includes abusive language targeting individuals or groups based on identity or personal attributes. *Threats* involve expressions of intent to cause harm. *Extremism* captures content related to extremist ideologies or propaganda. *Insults and Flaming* refers to offensive or aggressive language aimed at provoking or demeaning others, while *Other Offensive Texts* include inappropriate or offensive content that does not fall into the above categories.

The task was evaluated using a macro-averaged F1-score to ensure balanced performance across all classes, particularly in the presence of class imbalance and varying degrees of toxicity.

4 Dataset

The shared task is based on the *GameTox* dataset (Naseem et al., 2025), a large-scale collection of game chat utterances designed for toxicity detection through intent classification. The dataset consists of approximately 53,000 utterances collected from the online game *World of Tanks* via the WoT-Record database, capturing realistic player interactions in gaming environments.

4.1 Data Collection and Annotation

The dataset was constructed from publicly available chat logs, followed by preprocessing steps including language filtering, normalization, and removal of user identifiers to preserve privacy. Intent annotations were obtained through a human-LLM collaborative process, where initial pseudo-labels generated by large language models were verified and refined by human annotators. A multi-phase annotation protocol was employed to ensure consistency, including pilot annotation, guideline refinement, and consolidation (Bhandari et al., 2023). Each utterance was assigned one of six labels: *Non-toxic* (0), *Insults and Flaming* (1), *Other Offensive Texts* (2), *Hate and Harassment* (3), *Threats* (4), and *Extremism* (5). The annotation process achieved high reliability, with strong inter-annotator agreement.

4.2 Dataset Split

For the shared task, the dataset is divided into training, validation, and test sets using an approximate 80/10/10 split. The distribution preserves the naturally imbalanced nature of toxicity in gaming environments, where non-toxic utterances dominate.

Label	Train	Val	Test	Total
Non-toxic	34797	4349	4351	43497
Insults and Flaming	5925	740	742	7407
Other Offensive Texts	1874	234	235	2343
Hate and Harassment	279	34	36	349
Threats	60	7	8	75
Extremism	24	3	3	30
Total	42959	5367	5375	53701

Table 1: Dataset statistics for the shared task.

As shown in Table 1, class distribution is imbalanced, with majority utterances being non-toxic, while severe toxicity categories such as threats and extremism are relatively rare. This reflects real-world gaming environments and poses additional challenges for robust model development.

5 Evaluation and Competition

This section describes the structure of our competition, along with the methodology used to determine ranks and other relevant details.

5.1 Evaluation Metrics

To evaluate the effectiveness of the participants' contributions, we used four metrics: macro F1-score, accuracy, precision, and recall. The participants' final ranks were determined using the macro F1-score as the primary ranking metric.

5.2 Competition Setup

We used Codabench[1] to organize our competition. The competition consisted of two phases: a development phase, where participants could familiarize themselves with the Codabench platform and develop their methods, and a test phase, where performance was used to determine the final ranking on the leaderboard. The results from the development phase were made available to participants after the phase concluded, enabling them to further refine their approaches for the test phase.

5.2.1 Registration

A total of 102 participants registered, out of which 35 teams submitted their predictions. The leaderboard is shown in Table 2.

5.2.2 Competition Timelines

The competition commenced on December 10, 2025, when training and development data were made available, marking the start of the development phase. During this phase, participants familiarized themselves with the Codabench platform and began developing their systems. The test phase began on January 15, 2026, when test data was provided without any ground truth labels. The test phase concluded on March 18, 2026. The paper submission deadline was March 29, 2026. Notification of acceptance was scheduled for April 28, 2026, with camera-ready papers due by May 12.

6 Participants' Methods

syuhhh (Shi et al., 2026) proposed a three-stage progressive training framework on XLM-RoBERTa-large. The stages comprised: (1) gaming-domain adaptive MLM pre-training on a combined corpus of Dota 2 chat, multi-game balanced chat, and Twitter gaming toxicity datasets; (2) multilingual toxicity transfer fine-tuning on the Jigsaw 2018 dataset across five languages; and (3) SCL-enhanced end-to-end fine-tuning with a dual-head architecture (classification head and projection head) jointly optimized via class-balanced cross-entropy and supervised contrastive loss. The system was further enhanced with DeepSeek-driven short text augmentation, Claude API-generated long-tailed class synthesis for minority categories (classes 3–5), Nelder-Mead threshold optimization, and a minority-focused three-component ensemble combining the primary system with ToxicBERT and Claude API outputs. Their approach achieved a Macro F1 of 0.7041, ranking 1st among 35 teams, with ablation studies attributing the largest gains to domain alignment and toxic transfer (+10.37 points) and LLM-driven data augmentation (+4.71 points).

FNLP412 (Radulescu, 2026) approached the GameTox six-class toxicity classification task through a systematic comparison of seven model configurations built on top of a TF-IDF logistic regression baseline. Domain-specific preprocessing included URL and mention normalisation, repetition reduction, a manually curated slang map, and LLM-generated synthetic samples for the severely under-represented Threats and Extremism classes. The core architecture progressively moved from XLM-RoBERTa to MDeBERTa-V3, with the strongest variant first pre-trained on the Jigsaw Multilingual Toxic Comment dataset for one epoch before being fine-tuned on GameTox for five epochs; class-imbalance was further addressed through stratified five-fold cross-validation and severity-waterfall threshold optimisation at inference time. The final MDeBERTa-V3 system achieved a Macro F1 of 0.6725, placing 2nd on the shared task leaderboard.

thaulab (Guragain et al., 2026) presented a three-stage neural-symbolic pipeline combining an ensemble of DeBERTa-v3-base and XLM-RoBERTa-base with a Linguistically-Informed Mediator (LIM) that resolves inter-model disagreements through corpus-backed lexical normalization, class-conditional unigram scoring, multilingual profanity detection, and speech-act-theory-grounded agentive targeting analysis. To address extreme class imbalance, a two-stage augmentation strategy employed confusion-pair-driven and contrastive boundary generation using Claude

[1] https://www.codabench.org/competitions/12083/

Rank	Username	F1 Macro	Accuracy	Precision	Recall
1	syuhhh-637901 (Shi et al., 2026)	0.7041	0.8982	0.6400	0.7986
2	ramihai-572801 (Radulescu, 2026)	0.6725	0.8992	0.6636	0.6846
3	anmolguragain-637916 (Guragain et al., 2026)	0.6441	0.9062	0.6334	0.6601
4	srikarkashyap-635409 (Pulipaka, 2026)	0.6234	0.8800	0.5864	0.6814
5	akshyatshah-636282 (Shah et al., 2026)	0.6186	0.8902	0.6047	0.6497
6	yinloonkhor-636292	0.5932	0.8925	0.6098	0.5946
7	shrinep-637207	0.5883	0.9031	0.5540	0.6590
8	wangkongqiang-504685 (Wang et al., 2026)	0.5776	0.9075	0.6847	0.5343
9	dkhonker-536426	0.5749	0.8865	0.6214	0.5815
10	_alexcristea-610819	0.5632	0.8733	0.5652	0.5754
11	akking-609884	0.5563	0.8876	0.5239	0.6002
12	rukesh-shrestha-503743	0.5539	0.8932	0.5599	0.5557
13	nepalshr-637149	0.5512	0.8930	0.5201	0.6476
14	merrli-510969	0.5302	0.8603	0.4798	0.6137
15	xiaotian-518453	0.5301	0.8969	0.5402	0.5291
16	runick_allure-508659	0.5281	0.8772	0.5441	0.5328
17	rohanmainali-491803	0.5192	0.8893	0.5192	0.5221
18	linus-636500 (Ghimire et al., 2026)	0.5104	0.8716	0.5191	0.5134
19	xiaoyu666-603164	0.4984	0.8951	0.5156	0.4884
20	havnis-610798	0.4895	0.8794	0.4766	0.5083
21	giris-585517	0.4878	0.8964	0.5081	0.4895
22	shashi_sah-637803	0.4869	0.8999	0.5001	0.4774
23	wjyyyy-609715	0.4774	0.8953	0.4962	0.4732
24	justdoi-613394	0.4737	0.8973	0.4487	0.5071
25	barkion-610469	0.4726	0.8781	0.4538	0.5002
26	mestecha-623302	0.4686	0.8927	0.4763	0.4950
27	binayakkarki-589485 (Karki et al., 2026)	0.4645	0.8921	0.4647	0.4688
28	syhhh-610772	0.4641	0.7792	0.4198	0.5659
29	exterio-610602	0.4491	0.8443	0.4205	0.5084
30	zmin123-554678	0.4487	0.8506	0.4646	0.4568
31	aryankafle-524077	0.4421	0.8962	0.4490	0.4373
32	liutianyong-605718	0.4413	0.9036	0.4701	0.4219
33	quasar-501127	0.4169	0.6471	0.3943	0.5357
34	alexandru412-511289	0.3783	0.7068	0.3315	0.6432
35	wenbin-520996	0.1558	0.7784	0.1629	0.1653

Table 2: Leaderboard ranked by Macro F1-score. All scores are presented as percentages (%). Note that this leaderboard contains the score till the test deadline and does not consider further runs done by participants as a part of the system description paper.

Opus 4.6. The LIM concentrates corrections on safety-critical minority classes, yielding a Macro F1 of 0.6441 and the highest accuracy of 0.9062, ranking 3rd among 35 teams.

PSK (Pulipaka, 2026) fine-tuned Llama 3.1 8B with LoRA adapters and 4-bit quantization, augmented by GPT-4o-mini-generated paraphrases targeting minority classes at a carefully calibrated 5% synthetic data ratio. Structured prompt templates prepending class definitions were employed to sharpen category discrimination. The authors identified a "validation trap" phenomenon wherein models achieving high validation F1 via conservative majority-class predictions generalized poorly to the test set. The final system achieved a Macro F1 of 0.6234, ranking 4th among 35 teams.

TAGA (Shah et al., 2026) proposed a Token-Attribution Guided Attention (TAGA) architecture that augments a DeBERTa-v3-base encoder with externally computed toxicity signals to steer attention toward the most toxicity-indicative tokens. A leave-one-out perturbation method using the Detoxify scorer produces per-token attribution vectors across four channels (toxicity, threat, insult, identity attack), which are injected as learned biases into a content-based attention pooling layer; sentence-level features from two complementary Detoxify variants are concatenated to yield the final representation. Preprocessing handles gaming-specific obfuscation through leetspeak decoding, expansion of 22 gaming abbreviations, and regex-based uncensoring, while training combines focal loss with label smoothing, an auxiliary token-level MSE loss, and strategic class-specific oversampling of up to 15× with text augmentation. A five-phase ablation

study confirmed each component's incremental contribution, and the full system achieved a test Macro F1 of 0.618.

wangkongqiang (Wang et al., 2026) explored both transformer-based encoder fine-tuning and LLM instruction tuning. On the encoder side, multiple pre-trained models—including BERT, RoBERTa, ERNIE, ALBERT, and SimCSE-RoBERTa—were fine-tuned with task-specific classification heads, and a hard-voting ensemble was constructed over four RoBERTa variants augmented with LSTM and GRU layers. Additionally, Qwen2 1.5B and 7B Instruct variants were instruction-tuned on formatted triplets comprising instruction, input, and expected output. The Qwen2-7B configuration achieved the best performance with a Macro F1 of 0.5776, ranking 8th position.

LINUS (Ghimire et al., 2026) conducted a systematic benchmarking of multilingual transformer encoders - Toxic-XLM-RoBERTa, XLM-RoBERTa, m-DistilBERT, m-BERT, and mmBERT-base - on the GameTox fine-grained toxicity classification task. All models were fine-tuned using a customised `WeightedTrainer` that injects dynamically computed balanced class weights into the cross-entropy loss to counteract the severe class imbalance in the dataset. mmBERT-base, pre-trained on massively multilingual social media and informal web corpora, emerged as the best-performing architecture, achieving a validation Macro F1 of 0.5882 with a learning rate of 1e-5 and batch size of 64; however, a substantial $\sim$0.16 F1 generalisation gap on the official test set (0.4282) highlighted the difficulty posed by evolving gaming slang and distributional shift between validation and unseen test interactions. The system ranked 18th out of 35 participating teams.

ShriNep (Karki et al., 2026) presented RAKSHAK, a multi-task DeBERTa-v3-base framework for fine-grained toxic intent classification in gaming chat. The system integrated four key innovations: (1) rationale distillation from Qwen2.5-14B following the distill-then-train paradigm, where 5,000 teacher-generated natural-language rationales were concatenated with input messages during training but discarded at inference; (2) cross-domain transfer from the Jigsaw Toxic Comment dataset, with 16,225 samples mapped to GameTox Labels 1–4 via dual-LLM-validated label alignment; (3) 100 LLM-generated synthetic extremism samples produced through a four-step keyword-mining and placeholder-injection pipeline to circumvent LLM safety filters; and (4) dedicated rare-class binary heads for Threats and Extremism alongside Supervised Contrastive Loss on the shared embedding space, optimized jointly with Focal Loss. RAKSHAK achieved a Macro F1 of 0.5883, ranking 7th out of 35 teams, with a three-way ablation attributing +2.6 F1 points to Jigsaw cross-domain transfer and a further +3.7 points to the multi-task architectural implementation.

7 Discussion

The submissions to the shared task collectively demonstrated that fine-grained toxicity detection in gaming environments remains a challenging yet rapidly advancing research area. The diversity of approaches explored by participants highlighted several recurring methodological trends that proved effective under severe class imbalance, multilingual variation, and short noisy utterances.

A major observation across top-performing systems was the importance of domain adaptation. Systems that incorporated gaming-specific pretraining, multilingual toxicity transfer, or external toxicity corpora consistently outperformed generic fine-tuning approaches. In particular, teams leveraging staged training pipelines, domain-adaptive masked language modeling, or transfer learning from datasets such as Jigsaw achieved substantial gains in Macro F1-score. These findings reinforce the importance of aligning pretrained language models with the linguistic characteristics of gaming communication, including slang, abbreviations, obfuscations, and highly contextual expressions.

Another common trend among successful systems was the extensive use of data augmentation and synthetic sample generation for minority classes. Since categories such as *Threats* and *Extremism* were severely underrepresented, many teams relied on large language models to generate additional training examples, paraphrases, or rationale-based explanations. The strong performance of these approaches suggests that carefully designed augmentation pipelines can partially mitigate long-tail data scarcity. However, they also raise important questions regarding distributional realism, annotation consistency, and the risk of overfitting to synthetic patterns. Several partic-

ipants further showed the value of architectural specialization for rare toxic categories. Multi-task learning, dedicated binary heads, supervised contrastive learning, token-attribution guidance, and ensemble-based minority correction mechanisms all contributed to improved recognition of difficult classes. These approaches indicate that treating minority toxic categories as separate optimization objectives may be more effective than relying solely on standard multi-class classification losses.

Despite these advances, the leaderboard results also reveal that substantial challenges remain. Many systems exhibited high overall accuracy but comparatively lower Macro F1-scores, reflecting persistent difficulty in correctly identifying minority classes. Large generalization gaps between validation and test performance observed in several submissions further suggest that gaming toxicity remains highly sensitive to domain shift, evolving slang, multilingual variation, and contextual ambiguity. This highlights the need for more robust evaluation protocols and models capable of better generalization under realistic deployment conditions. Future work may explore incorporating conversational context, temporal interaction patterns, multimodal player signals, and retrieval-augmented reasoning to improve toxicity understanding beyond isolated utterance classification. Additionally, explainability, fairness, and bias mitigation remain important directions for future gaming moderation systems, particularly given the social consequences of automated moderation errors.

8 Conclusion

This shared task provided a comprehensive benchmark for fine-grained toxicity detection in online gaming environments using the GameTox dataset. The competition attracted strong participation and demonstrated a wide range of effective approaches, including domain-adaptive pretraining, multilingual transfer learning, LLM-based augmentation, contrastive learning, and specialized rare-class modeling strategies. The results highlight both the progress made and the remaining challenges in detecting nuanced toxic behavior under realistic class imbalance and domain-shift conditions. We hope this shared task encourages further research toward more robust, fair, and reliable toxicity detection systems for gaming communities.

Limitations

This shared task has several limitations. First, the GameTox dataset is derived primarily from *World of Tanks*, which may limit generalizability to other gaming communities and communication styles. Second, the dataset is highly imbalanced, with very limited samples for categories such as *Threats* and *Extremism*, making robust learning difficult. Third, toxicity is often context-dependent, and isolated utterances may not fully capture sarcasm, implicit abuse, or conversational intent. Finally, some participant systems relied on LLM-generated synthetic data, which may introduce artifacts or biases not present in authentic gaming interactions.

Ethical Considerations

Automated toxicity detection systems may produce both false positives and false negatives, potentially affecting moderation fairness and user experience. Biases in pretrained models, annotation processes, or synthetic augmentation may further impact system behavior across different linguistic communities and communication styles. To reduce privacy concerns, the dataset was constructed from publicly available chat logs and anonymized through preprocessing procedures. This shared task is intended solely for research purposes toward developing safer online gaming environments.

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.

Jeremy Blackburn and Haewoon Kwak. 2014. Stfu noob! predicting crowdsourced decisions on toxic behavior in online games. In *Proceedings of the 23rd international conference on World wide web*, pages 877–888.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. Hatebert: Retraining bert for abusive language detection in english. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised

cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Garry Crawford, Victoria K Gosling, and Ben Light. 2013. The social and cultural significance of online gaming. In *Online gaming in context*, pages 3–22. Routledge.

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.

Prajwal Ghimire, Aashish Mahato, and Sunil Regmi. 2026. Linus@eeuca 2026: Fine-grained toxicity detection in gaming chat using multilingual transformers. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Anmol Guragain, Marcos Estecha-Garitagoitia, and Luis Fernando D'Haro. 2026. thaulab@eeuca 2026: Who said what to whom? a targeting-aware neural-symbolic pipeline for gaming toxicity detection. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

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)*.

Binayak Karki, Aryan Kafle, and Pingala Ghimire. 2026. Shrinep@eeuca 2026: Rakshak – multi-task deberta with rationale distillation and jigsaw-augmented training for toxic intent classification. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

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*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Luke Munn. 2023. Toxic play: Examining the issue of hate within gaming. *First Monday*.

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.

Srikar Kashyap Pulipaka. 2026. Psk@eeuca 2026: Fine-tuning large language models with synthetic data augmentation for multi-class toxicity detection in gaming chat. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Mihai Radu Radulescu. 2026. Fnlp412@eeuca 2026: Understanding toxic behavioral intent in gaming chat logs using transfer learning and synthetic data augmentation. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Kritesh Rauniyar, Sweta Poudel, Shuvam Shiwakoti, Surendrabikram Thapa, Junaid Rashid, Jungeun Kim, Muhammad Imran, and Usman Naseem. 2023. Multi-aspect annotation and analysis of nepali tweets on anti-establishment election discourse. *IEEE Access*, 11:143092–143115.

Harshil Sanghvi, Rushir Bhavsar, Vini Hundlani, Lata Gohil, Tarjni Vyas, Anuja Nair, Shivani Desai, Nilesh Kumar Jadav, Sudeep Tanwar, Ravi Sharma, and 1 others. 2024. Metahate: Ai-based hate speech detection for secured online gaming in metaverse using blockchain. *Security and Privacy*, 7(2):e343.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Akshyat Shah, Shashi Sah, Aryan Gupta, and Kavinder Singh. 2026. Taga@eeuca 2026: Token-attribution guided attention for fine-grained toxic behaviour classification in online gaming communities. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Yuhao Shi, Yu Wang, and Shengjie Zhao. 2026. syuhhh@eeuca 2026: A three-stage progressive training framework for fine-grained toxicity detection in online gaming communities. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Yehor Tereshchenko and Mika K Hämäläinen. 2025. Efficient toxicity detection in gaming chats: A comparative study of embeddings, fine-tuned transformers and llms. *Journal of Data Mining & Digital Humanities*.

Kongqiang Wang, Peng Zhang, and Qingli Tan. 2026. wangkongqiang@eeuca 2026: Understanding toxic behavioral intent in gaming chat logs. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Garrison Wells, Ágnes Romhányi, and Constance Steinkuehler. 2025. Hate speech and hate-based harassment in online games. *Frontiers in Psychology*, 15:1422422.

Ágnes Zsila and Zsolt Demetrovics. 2025. Taxonomy of toxic behaviors in multiplayer gaming environments: An extension of the context of peer aggression. *Journal of behavioral addictions*.

Ágnes Zsila, Reza Shabahang, Mara S Aruguete, and Gábor Orosz. 2022. Toxic behaviors in online multiplayer games: Prevalence, perception, risk factors of victimization, and psychological consequences. *Aggressive Behavior*, 48(3):356–364.