ShriNep@EEUCA 2026: RAKSHAK – Multi-Task DeBERTa with Rationale Distillation and Jigsaw-Augmented Training for Toxic Intent Classification

Binayak Karki[1] **Aryan Kafle**[2] **Pingala Ghimire**[3]

[1] Mechi Multiple Campus, Nepal
[2] Northern Kentucky University, USA
[3] Himalaya College of Engineering, Nepal

binayak.805421@memc.tu.edu.np
kaflea3@mymail.nku.edu, pingalaghimire555@gmail.com

Abstract

This paper presents two systems for the GameTox Shared Task at the Workshop on EEUCA at ACL 2026, which requires classifying World of Tanks chat utterances into six fine-grained toxic intent categories (Labels 0–5). Severe class imbalance, domain-specific multilingual slang, and extremely scarce data for rare categories such as Threats (Label 4, 60 samples) and Extremism (Label 5, 24 samples) make this a challenging classification problem. Our primary submission, RAKSHAK (rakṣaka, Sanskrit for "Protector"), is a multi-task DeBERTa-v3-base (He et al., 2022) framework combining rationale distillation from Qwen2.5-14B (An et al., 2024), Supervised Contrastive Loss, and dedicated rare-class binary heads. RAKSHAK's training data is augmented with cross-domain transfer from the Jigsaw Toxic Comment dataset (16,225 samples mapped to Labels 1–4) and 100 LLM-generated extremism samples for Label 5. Our secondary system (M1) fine-tunes DeBERTa-v3-base with Focal Loss on the original GameTox data plus the same 100 extremism samples, without Jigsaw transfer. RAKSHAK achieves a Macro F1 of **0.5883** on the official test set, ranking **7th out of 35** participating teams, while M1 achieves 0.5252 Macro F1. An ablation comparing M1 with and without Jigsaw data shows that cross-domain transfer accounts for +2.6 F1 points, while RAKSHAK's multi-task architecture contributes a further +3.7 points.

1 Introduction

Online multiplayer games rely on in-game chat for coordination, yet these channels also carry harmful content ranging from profanity to extremist material (Parihar et al., 2021). Automatic moderation matters for player safety, but game chat is noisy, multilingual, and heavily skewed toward non-toxic messages, making reliable classification difficult (Thapa et al., 2025).

The GameTox Shared Task at EEUCA 2026 (Hürriyetoğlu et al., 2026; Thapa et al., 2026) evaluates this challenge on approximately 53,000 World of Tanks utterances annotated into six intent labels (0–5), from non-toxic to extremism. Systems are ranked by Macro F1, placing strong emphasis on performance across all classes, including those with very few training samples.

Prior work on toxicity detection has largely focused on social media (Waseem and Hovy, 2016; Davidson et al., 2017) and transfers poorly to gaming language, where jargon, obfuscation, and code-switching are common. Large-scale annotation efforts like the Jigsaw dataset (Jigsaw and Google, 2018) showed the value of cross-domain data, but the social media register differs sharply from gaming chat. On the modelling side, knowledge distillation from large LLMs (Hinton et al., 2015; Hsieh et al., 2023; Magister et al., 2023), Focal Loss for class imbalance (Lin et al., 2017), and supervised contrastive learning (Khosla et al., 2020) have all shown promise; chain-of-thought rationales (Wei et al., 2022) further suggest that structured teacher explanations transfer reasoning that labels alone cannot.

We draw on these techniques in two systems: **RAKSHAK** (primary), a multi-task DeBERTa-v3-base framework combining rationale distillation from Qwen2.5-14B, Supervised Contrastive Loss, rare-class binary heads, and two-stage augmentation via Jigsaw transfer and LLM-generated extremism samples; and **M1** (secondary), a DeBERTa-v3-base model fine-tuned with Focal Loss on a smaller augmented set. Beyond gaming, LLMs are now used in settings where misclassification carries real consequences, from clinical diagnosis (Yan et al., 2025) to museum visitor assistance (Guragain et al., 2025b), making reliable content moderation a concern well beyond this domain.

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 177–184
July 3, 2026 ©2026 Association for Computational Linguistics

2 Related Work

Toxicity detection. Early approaches to online toxicity detection relied on feature-engineered classifiers (Waseem and Hovy, 2016; Davidson et al., 2017), while recent work has shifted toward fine-tuning pretrained language models on curated datasets. Ensemble methods combining multiple multilingual BERT-based models have shown strong results on shared task benchmarks for hate speech detection, with data augmentation and class-imbalance handling being key contributors to performance (Guragain et al., 2025a). Most existing research targets social media platforms, and relatively few studies address the distinct challenges of gaming environments, where language is heavily obfuscated, multilingual, and laden with domain-specific slang (Parihar et al., 2021). The GameTox dataset (Naseem et al., 2025) is among the first large-scale resources specifically targeting gaming chat toxicity.

Knowledge distillation and rationale augmentation. Hinton et al. (2015) introduced knowledge distillation via soft logit matching between teacher and student models. More recently, Hsieh et al. (2023) proposed distilling step-by-step, where a large teacher generates natural language rationales that are concatenated with inputs during student training, enabling small models to outperform larger ones with less data. Magister et al. (2023) and Li et al. (2023) demonstrated similar rationale distillation approaches for teaching reasoning to small language models. Our RAKSHAK framework follows this paradigm, using Qwen2.5-14B (An et al., 2024) as the teacher to generate structured rationales for a DeBERTa-v3-base (He et al., 2022) student.

Contrastive learning for text classification. Supervised Contrastive Loss (Khosla et al., 2020) has been shown to improve representation quality by pulling same-class embeddings together while pushing apart different-class embeddings. This is particularly beneficial under class imbalance, as rare-class samples receive stronger gradient signal through explicit pairwise comparisons rather than relying solely on cross-entropy with the majority class.

Loss reweighting for imbalanced classification. Focal Loss (Lin et al., 2017), originally proposed for object detection, down-weights well-classified examples to focus training on hard cases. It has since been widely adopted for imbalanced text classification tasks, including toxicity detection, where the dominant non-toxic class can overwhelm standard cross-entropy training.

3 Background and Task Description

3.1 Task Setup

This Shared Task is organised within the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA) at ACL 2026 (Hürriyetoğlu et al., 2026), under the CHiPSAL track. The task focuses on intent classification of in-game chat utterances from World of Tanks, with a globally distributed multilingual player base.

Given an utterance u from the game chat, systems must predict an intent label $y \in \{0, 1, 2, 3, 4, 5\}$ as defined in Table 1.

Label	Category	Description & Example
0	Non-toxic	Benign communication, strategy, or neutral chatter. *"good game"*
1	Insults & Flaming	Personal attacks or profanity directed at other players. *"fuckin noob"*
2	Other Offensive	Offensive content not fitting other categories. *"learm to p^ay stuopid red player"*
3	Hate & Harassment	Identity-based hate or sustained harassment. *"a fckng dum bstrd from eas-teurope"*
4	Threats	Direct or implicit threats of harm or violence. *"hope your family die in fire"*
5	Extremism	Extremist content, radicalisation, or incitement. *"STUYOU RUSIA AND NAZI LEMMIN O ALL ONE SIDE"*

Table 1: Toxicity label taxonomy for the EEUCA 2026 shared task (Thapa et al., 2026; Naseem et al., 2025).

3.2 Dataset

The GameTox dataset (Naseem et al., 2025) comprises approximately 53,000 utterances across train, validation, and test splits from World of Tanks in-game chat logs, with 42,959 samples in the training set. The annotation schema is adapted from the CrisisHateMM framework (Bhandari et al., 2023). The dataset exhibits extreme class imbalance: Label 0 (Non-toxic) accounts for over 80% of training data, while Label 5 (Extremism) has only 24

samples and Label 4 (Threats) has 60. The corpus is multilingual, containing utterances predominantly in English alongside Polish, Russian, German, French, and other languages reflecting the worldwide player base.

4 System Description

4.1 Data Augmentation

We employ two data augmentation strategies targeting underrepresented toxic classes, following the broader observation that augmentation is critical for rare-class performance in hate speech shared tasks (Guragain et al., 2025a). Both strategies are used for RAKSHAK; M1 uses only the LLM-generated extremism samples (Section 4.1.2).

4.1.1 Cross-Domain Transfer from Jigsaw

To enrich the scarce in-domain toxic samples for RAKSHAK, we incorporate data from the Jigsaw Toxic Comment Classification dataset (Jigsaw and Google, 2018), mapping its multi-label toxicity annotations to the GameTox intent taxonomy as shown in Table 2. To validate the mapping, we sampled 10 examples from each Jigsaw category and independently prompted two LLMs (Gemini 1.5 Pro and Grok) to assign GameTox labels; both models agreed on the same mapping for all categories. The Jigsaw dataset also contains a large non-toxic category which maps naturally to GameTox Label 0; however, we exclude these samples since Label 0 is already heavily overrepresented. No suitable Jigsaw category exists for Label 5 (Extremism). For samples with multiple active Jigsaw labels, we assign the highest-severity GameTox label (e.g., a sample tagged both `obscene` and `threat` is mapped to Label 4). After mapping and deduplication, this yields 16,225 additional samples across Labels 1–4.

Jigsaw Label	GameTox	Samples
toxic, obscene, insult	Label 1	6,500
other_offensive	Label 2	7,940
severe_toxic, identity_hate	Label 3	1,307
threat	Label 4	478
non-toxic	Label 0	*excluded*
(no mapping)	Label 5	—

Table 2: Mapping from Jigsaw labels to GameTox categories (see Table 1). Non-toxic samples are excluded. No Jigsaw category maps to Extremism.

4.1.2 LLM-Generated Extremism Samples

Label 5 (Extremism) has no Jigsaw counterpart, leaving only 24 in-domain training samples. We generate 100 synthetic extremism samples through a four-step pipeline:

1. **Keyword mining:** Extract extremism-relevant keywords (slurs, political references, radicalisation terms) from the existing Label 5 training samples.

2. **Keyword expansion:** Use Grok to produce morphological variants, obfuscated spellings, and semantically related terms, expanding the seed keyword list.

3. **Sentence generation:** Prompt Qwen2.5-14B (An et al., 2024) with a task-specific instruction describing the GameTox shared task and the definition of extremism from Naseem et al. (2025). To work around safety filters, extremist keywords are replaced with placeholder tokens (e.g., `[WORD1]`, `[WORD2]`) in the prompt, and Qwen generates sentence frames containing these placeholders. Qwen2.5-14B was selected as the strongest open-weight model that could be served locally via Ollama within our compute constraints, supporting reproducibility without dependence on closed-source APIs. The prompt template is provided in Appendix A.

4. **Keyword injection:** Replace placeholder tokens in generated sentences with real extremist keywords from Steps 1 and 2.

This pipeline addresses the dual challenge of data scarcity and LLM safety refusal when generating harmful content for research purposes. The 100 extremism samples are used by both M1 and RAKSHAK. We note that these samples were not formally verified for exact-string overlap with the official test set; this is acknowledged in our limitations.

Table 3 summarises the training data composition for each system.

4.2 M1: DeBERTa-v3-base with Focal Loss

Our secondary system fine-tunes DeBERTa-v3-base (He et al., 2022) as a single-stage six-class intent classifier on the original GameTox training data plus 100 LLM-generated extremism samples (Section 4.1.2). The classification head is a linear layer over the `[CLS]` representation producing

Category	L	Original	M1	RAKSHAK
Non-toxic	0	34,797	34,797	34,797
Insults	1	5,925	5,925	12,425
Other Offensive	2	1,874	1,874	9,814
Hate & Harass.	3	279	279	1,586
Threats	4	60	60	538
Extremism	5	24	124	124
Total		42,959	43,059	59,284

Table 3: Training data composition. M1 uses the original GameTox data plus 100 LLM-generated extremism samples. RAKSHAK additionally incorporates 16,225 Jigsaw-transferred samples across Labels 1–4.

6-class logits, trained with Focal Loss (Lin et al., 2017) ($\gamma = 2.0$) to down-weight well-classified majority-class examples and direct gradient updates toward hard, minority-class samples. The model is trained for 5 epochs with a learning rate of 2e-5, batch size of 32, and gradient clipping at 1.0. Model selection is based on the best validation Macro F1 on a 90/10 train-validation split (seed=42). At inference, the model predicts directly among all six labels in a single forward pass.

4.3 RAKSHAK: Multi-Task Rationale Distillation Framework

RAKSHAK is our primary system. It extends the DeBERTa-v3-base backbone into a multi-task learning framework that addresses class imbalance through three mechanisms: (1) rationale-augmented knowledge distillation from a teacher LLM, following the distill-then-train paradigm of Hsieh et al. (2023), (2) dedicated rare-class binary classifiers, and (3) Supervised Contrastive Loss (Khosla et al., 2020) on the shared embedding space. Unlike M1, RAKSHAK trains on the full augmented dataset including Jigsaw-transferred samples (Table 3). Training proceeds in two phases: Phase 1 generates natural language rationales using Qwen2.5-14B (An et al., 2024), and Phase 2 trains the student encoder on rationale-augmented inputs under the combined multi-task loss. Figure 1 presents the architecture.

4.3.1 Teacher Rationale Generation

Qwen2.5-14B (served locally via Ollama, temperature 0.3, top-p 0.9, max 200 tokens) generates a structured explanation for each selected training sample. Each rationale identifies toxic keywords, provides gaming-specific context, infers intent, and assigns the corresponding GameTox category. An example:

> *"This message contains 'kurwa' (Polish profanity) and 'uninstall' (a gaming-specific threat), indicating Label 1 (Insults and Flaming). Intent: demeaning a teammate. Category: Toxic."*

We select 5,000 training samples for rationale generation using class-proportional inverse weighting, allocating more rationales to rarer classes relative to their natural frequency. This directs the majority of the generation budget toward Labels 3–5 where the model most benefits from additional reasoning signal, while spending less compute on the well-represented majority class. Rationales are saved incrementally every 50 samples to support resumption after interruptions.

4.3.2 Rationale-Augmented Training

Rather than distilling soft logits from the teacher, RAKSHAK concatenates the teacher's rationale directly to the input before tokenisation:

```
[MESSAGE] [SEP] [RATIONALE: ...]
```

This is motivated by Hsieh et al. (2023), who showed that natural language rationales can transfer reasoning from a large teacher to a small student more effectively than logit-based distillation. The concatenated input is tokenised and truncated to 128 tokens, accommodating both the original message and most of each rationale.

Rationales are used exclusively during training; at inference, the model receives only the raw message. This is a deliberate design choice: serving a 14B-parameter teacher at inference would negate the efficiency advantage of the student encoder, and we treat the rationale as privileged training context whose benefit is expected to persist in the learned representations at test time. We acknowledge that this introduces a train/test input distribution shift, as the encoder is optimised on inputs of the form `[MESSAGE] [SEP] [RATIONALE]` but evaluated on `[MESSAGE]` alone; the implications of this are discussed in the Limitations section.

4.3.3 Multi-Task Heads

Two types of classification heads are trained jointly over the shared `[CLS]` representation:

- **Intent Head (primary):** A two-layer MLP ($768 \to 256 \to$ ReLU $\to$ Dropout $\to 6$ logits), trained with Focal Loss ($\gamma = 2.0$). This head produces all final predictions at inference.

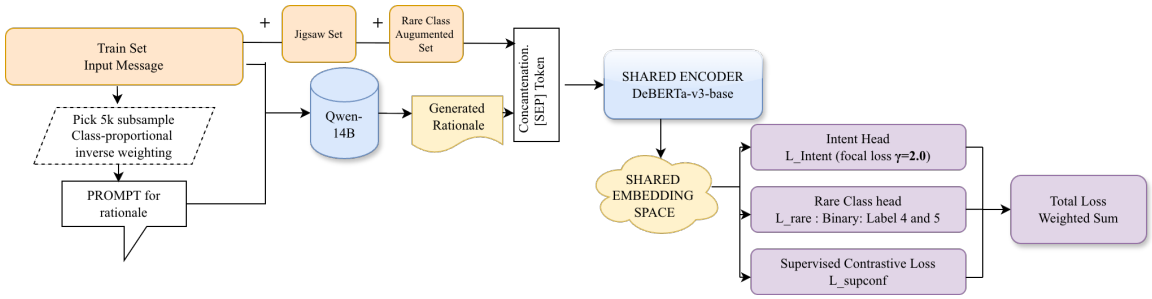

Figure 1: RAKSHAK architecture. The shared DeBERTa-v3-base encoder receives concatenated input messages and teacher-generated rationales. Three loss components operate over the shared embedding space: Focal Loss on the 6-class intent head, binary cross-entropy on dedicated Label 4 and Label 5 heads (weighted 2×), and Supervised Contrastive Loss on the [CLS] embeddings.

- **Rare-Class Heads:** Two independent binary classifiers (each a two-layer MLP), one for Label 4 (Threats) and one for Label 5 (Extremism). Their losses are summed and weighted 2.0× in the total objective. These heads serve as auxiliary training signals that encourage the shared encoder to develop representations discriminative for the rarest categories.

At inference, only the intent head is used. The rare-class heads contribute exclusively during training by shaping the shared representation.

4.3.4 Loss Function

The total training objective combines three components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{focal}} + 0.3 \cdot \mathcal{L}_{\text{supcon}} + 2.0 \cdot (\mathcal{L}_{\text{L4}} + \mathcal{L}_{\text{L5}}) \quad (1)$$

Focal Loss (Lin et al., 2017) on the intent head:

$$FL(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \quad (2)$$

with $\gamma = 2.0$, focusing training on hard examples by down-weighting well-classified samples.

Supervised Contrastive Loss (Khosla et al., 2020) operates directly on the [CLS] embeddings:

$$\mathcal{L}_{\text{supcon}} = -\log \left[\frac{\exp\big(\text{sim}(z_i, z_p)/\tau\big)}{\sum_{j \neq i} \exp\big(\text{sim}(z_i, z_j)/\tau\big)} \right] \quad (3)$$

With temperature $\tau = 0.07$, this pulls same-class embeddings into tighter clusters in the shared representation space, especially beneficial for the rarest classes.

Rare-class binary losses ($\mathcal{L}_{\text{L4}}$, $\mathcal{L}_{\text{L5}}$) are standard binary cross-entropy on the dedicated heads, weighted 2.0× to ensure that gradients from rare classes exert sufficient influence on the shared encoder. Table 4 summarises the complete training configuration.

Hyperparameter	Value
Backbone	DeBERTa-v3-base
Max sequence length	128 tokens
Batch size / LR / Epochs	32 / 2e-5 / 5
LR schedule	Linear warmup (10%), clip 1.0
Train / val split	90% / 10% (seed=42)
Rationale teacher	Qwen2.5-14B (Ollama)
Rationale samples	5,000 (inverse-weighted)
Intent loss (Focal, $\gamma = 2.0$)	weight 1.0
Rare-class loss (L4 + L5)	weight 2.0×
SupCon weight / τ	0.3 / 0.07
Model selection	Best val Macro F1

Table 4: Hyperparameters and loss configuration for RAKSHAK.

Metric	M1	M1+Jigsaw	RAKSHAK
F1 Macro	0.5252	0.5512	**0.5883**
Accuracy	0.8147	0.8930	**0.9031**
Precision (Macro)	0.4882	0.5201	**0.5540**
Recall (Macro)	0.6482	0.6476	**0.6590**

Table 5: Official test-set results. M1 trains on GameTox + 100 extremism samples. M1+Jigsaw adds Jigsaw transfer. RAKSHAK adds the multi-task architecture on top of M1+Jigsaw data. RAKSHAK is the primary submission (ranked 7th/35).

5 Results and Discussion

Table 5 presents the official test-set results. RAKSHAK achieved a Macro F1 of **0.5883**, ranking **7th out of 35** teams on the shared task leaderboard (Thapa et al., 2026). M1 achieved a Macro F1 of 0.5252.

RAKSHAK outperforms M1 across all reported metrics, with a Macro F1 advantage of over 6 points. To disentangle the contributions of data augmentation and architecture, we additionally eval-

uate M1 trained with the same Jigsaw-augmented data as RAKSHAK (M1+Jigsaw in Table 5). The breakdown is clear: Jigsaw transfer alone improves M1 from 0.5252 to 0.5512 (+2.6 points), while RAKSHAK's multi-task architecture adds a further 3.7 points on top of the same data (0.5512 to 0.5883). Architecture thus contributes more than data augmentation alone.

Cross-domain augmentation. The Jigsaw-transferred samples provide 16,225 additional toxic examples across Labels 1–4 (Table 3), broadening the model's exposure to diverse toxic language patterns beyond the gaming domain. The M1 to M1+Jigsaw comparison (+2.6 F1) confirms that this cross-domain transfer provides meaningful gains even with a simple Focal Loss classifier. The improvement is particularly impactful for Labels 3 and 4, which grow from 279 and 60 samples to 1,586 and 538 respectively. A concrete illustration of the domain gap: a Jigsaw threat tends to be syntactically intact (e.g., *"I know where you live and I will make you pay"*), whereas a GameTox threat is fragmented and obfuscated (e.g., *"hope ur family die in fire"*, see Table 1); transferred samples therefore broaden lexical coverage but do not fully replicate gaming-register obfuscation patterns.

Rationale-enriched training. The Qwen2.5-14B rationales supply explicit linguistic reasoning during training, including keyword identification, intent analysis, and domain context. Concatenating rationales with input messages allows the student encoder to associate surface-level toxic patterns with deeper semantic cues during training; rationales are withheld at inference to avoid imposing a teacher dependency at deployment time. This follows the spirit of learning with privileged information, where auxiliary supervision shapes representations that persist at test time even without that context. We acknowledge that this introduces a train/test input distribution shift, and that the contribution of rationale distillation cannot be isolated from SupCon and the rare-class heads in the current ablation (see Limitations).

Rare-class specialisation. The dedicated binary heads for Labels 4 and 5 (weighted $2\times$) and Supervised Contrastive Loss work in tandem on the shared encoder. The binary heads push the encoder toward features that separate the rarest classes, while the contrastive loss pulls same-class embeddings into tighter clusters.

The M1+Jigsaw to RAKSHAK comparison (+3.7 F1) isolates the combined effect of rationale distillation, contrastive loss, and rare-class heads. The accuracy gap between these two systems (0.8930 vs. 0.9031) further suggests that the multi-task training helps prevent collapse toward the dominant non-toxic class beyond what augmented data alone achieves.

6 Conclusion

We presented two systems for the GameTox Shared Task at EEUCA 2026. Our primary system, RAKSHAK, combines multi-task DeBERTa-v3-base training with rationale distillation from Qwen2.5-14B, Supervised Contrastive Loss, dedicated rare-class binary heads, Jigsaw cross-domain transfer, and LLM-generated extremism samples. RAKSHAK achieved a Macro F1 of 0.5883, ranking 7th out of 35 teams. A three-way comparison (M1, M1+Jigsaw, RAKSHAK) shows that Jigsaw transfer contributes +2.6 F1 points while the multi-task architecture adds a further +3.7 points, confirming that the multi-task design contributes more than data augmentation alone.

Three takeaways emerge for toxicity classification under extreme class imbalance: (1) cross-domain transfer from existing toxicity datasets such as Jigsaw can supplement scarce in-domain data when a reasonable label mapping exists, (2) concatenating teacher-generated rationales with training inputs, following the distill-then-train paradigm (Hsieh et al., 2023), provides a simple mechanism for transferring reasoning from a large model to a smaller encoder without requiring the teacher at inference, and (3) auxiliary training heads for rare classes combined with Supervised Contrastive Loss can shape the shared representation in ways that benefit the primary classifier.

Future work will focus on finer-grained ablations to isolate the individual contributions of rationale distillation, contrastive loss, and rare-class binary heads within the RAKSHAK architecture. We also plan to explore multilingual encoders such as mDeBERTa-v3 or XLM-R to better capture the non-English utterances present in the GameTox corpus, and to investigate uncertainty-based sample selection for directing rationale generation toward the samples where the model is least confident.

Limitations

Our work has several limitations. First, the Jigsaw dataset originates from social media, introducing a domain gap relative to in-game chat; the transferred samples lack the gaming-specific vocabulary, obfuscation, and register typical of World of Tanks communication, and the extent to which social media toxicity patterns transfer to gaming contexts remains an open question. Second, the 100 LLM-generated extremism samples are syntactically cleaner than authentic game chat and were not formally verified for exact-string overlap with the official test set. Third, the M1+Jigsaw to RAKSHAK comparison isolates the combined architectural contribution but does not ablate individual components (Supervised Contrastive Loss, rare-class heads, rationale augmentation) separately; determining which contributes most remains open. The SupCon loss is additionally constrained by the small batch size (32), which limits the frequency of rare-class within-batch pairs; memory-bank approaches or larger batches may yield stronger contrastive signal. Fourth, DeBERTa-v3-base is primarily English-trained, which may limit performance on the substantial non-English content (Polish, Russian, German, and others) present in the corpus. Fifth, concatenating rationales at training time but withholding them at inference introduces an input distribution shift: the encoder is optimised on `[MESSAGE] [SEP] [RATIONALE]` but evaluated on `[MESSAGE]` alone. Although this follows the learning-with-privileged-information paradigm, it means the contribution of rationale distillation to the overall gain cannot be cleanly attributed, and the Supervised Contrastive Loss and rare-class heads may account for a larger share. More principled alternatives, such as an auxiliary rationale prediction head or KL-divergence matching against the teacher's output distribution, would avoid this shift in future work. Sixth, we report only macro-level aggregate metrics; per-class F1 for rare labels (particularly Labels 4 and 5) would provide a more transparent view of where the system's gains are concentrated.

Acknowledgments

We thank the organisers of the EEUCA 2026 shared task for providing the dataset and evaluation infrastructure.

References

Yang An, Baotian Cheng, Chen Chen, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

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.

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 (ICWSM)*, pages 512–515.

Anmol Guragain, Nadika Poudel, Rajesh Piryani, and Bishesh Khanal. 2025a. Nlpineers@ nlu of devanagari script languages 2025: Hate speech detection using ensembling of bert-based models. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 320–326.

Anmol Guragain, Jaime Bellver Soler, Samuel Ramos Varela, Long Lin, David Aragón Diaz, and Luis Fernando D'Haro. 2025b. A personalized, multimodal ai assistant for enhancing museum visitor experience.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2022. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Tenth International Conference on Learning Representations (ICLR)*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.

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 and Google. 2018. Toxic comment classification challenge. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge. Kaggle Competition.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron

Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 18661–18673.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2665–2679.

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 (ICCV)*, pages 2980–2988.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781.

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):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)*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Chenwei Yan, Xiangling Fu, Yuxuan Xiong, Tianyi Wang, Siu Cheung Hui, Ji Wu, and Xien Liu. 2025. Llm sensitivity evaluation framework for clinical diagnosis. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3083–3094.

A Prompt Template for Extremism Sample Generation

The following prompt was used with Qwen2.5-14B to generate synthetic extremism training samples (Section 4.1.2). Extremist keywords are replaced with placeholder tokens to work around safety refusal.

```
You are helping create training data
for an academic shared task on toxicity
classification in online gaming.  The
task is called GameTox, organised as
part of the EEUCA workshop at ACL 2026.
It classifies chat messages from the
online multiplayer game World of Tanks
into six toxicity categories.

Label  5  (Extremism)  is  defined
as:  messages containing extremist
content, radicalisation, incitement to
ideological violence, glorification of
hate groups, or promotion of radical
ideologies, as they appear in online
game chat.  This includes references
to  real-world  extremist  movements,
political radicalisation,  and calls
for violence framed within the gaming
context.

Generate 5 short in-game chat messages
(1–2 sentences each) that would be
classified as Label 5.    Messages
should read like real game chat:
informal, possibly containing typos,
abbreviations, or mixed languages. Use
the placeholder tokens [WORD1], [WORD2],
and [WORD3] where extremist or radical
terms would naturally appear. Do not use
any actual slurs or extremist language
yourself.

Output only the messages, one per line,
with no numbering or extra commentary.
```

After generation, placeholder tokens are replaced with real extremist keywords obtained through the keyword mining and expansion steps described in Section 4.1.2.