EEUCA 2026

The 9th Workshop on Event Extraction and Understanding: Challenges and Applications

Proceedings of the Workshop

July 3, 2026

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-402-6

Preface by the EEUCA organizers

Welcome to the Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA 2026), held in conjunction with the Annual Meeting of the Association for Computational Linguistics (ACL 2026).

EEUCA, formerly known as CASE (Challenges and Applications of Automated Extraction of Socio-political Events from Text), continues to serve as a premier venue for researchers working on event extraction, event understanding, information extraction, computational social science, and related areas. Over the years, the workshop has evolved alongside the rapid advancement of natural language processing technologies, expanding its scope from traditional event extraction pipelines to encompass multilingual, multimodal, and generative approaches for understanding complex real-world events.

This year's edition reflects several important developments in the field. First, large language models (LLMs) have become a central theme across both research papers and shared task submissions. Many contributions explore how LLMs can support event extraction through prompting, reasoning, weak supervision, agent-based architectures, and structured generation. At the same time, authors critically examine the limitations of current generative systems, including challenges related to reliability, interpretability, schema adherence, and low-resource settings.

Second, EEUCA 2026 highlights the growing importance of multimodal event understanding. As information about events is increasingly communicated through images, memes, videos, and other forms of multimedia content, researchers are developing methods that move beyond text-only analysis toward richer representations that combine visual and linguistic information. This trend is particularly evident in the workshop's shared tasks, which address socially relevant problems involving multimodal vaccine discourse and toxicity detection in online gaming communities.

The workshop received submissions covering a diverse range of topics, including low-resource event extraction, benchmark creation, symbolic reasoning, geopolitical event analysis, reflective multi-agent systems, and generative event extraction. The accepted papers collectively demonstrate the breadth of current research directions and the increasing interdisciplinarity of the field. EEUCA 2026 also had strong participation in its shared tasks. Compared to previous editions, this year's shared tasks place greater emphasis on multimodal reasoning, socially impactful applications, and real-world challenges such as misinformation, online harms, and nuanced behavioral intent detection. The enthusiasm and diversity of approaches demonstrated by participating teams illustrate the growing interest in event-centered AI research and provide valuable benchmarks for future work.

We would like to express our sincere gratitude to the authors, shared task participants, program committee members, reviewers, keynote speakers, and organizers whose efforts made this workshop possible. Their contributions ensure the continued success of EEUCA as a collaborative forum for advancing research on event extraction and understanding. We look forward to the continued growth of this research community and to future editions of EEUCA.

The EEUCA 2026 Organizing Committee

Organizing Committee

Workshop Organizers

Ali Hürriyetoğlu, Wageningen University & Research, Netherlands
Hristo Tanev, Joint Research Centre, European Commission, Italy
Surendrabikram Thapa, Virginia Polytechnic Institute and State University, USA
Surabhi Adhikari, Columbia University, USA

Program Committee

Program Chairs

Ali Hürriyetoğlu, Wageningen University & Research
Surendrabikram Thapa, Virginia Polytechnic Institute and State University
Hristo Tanev, Joint Research Centre, European Commission

Program Committee

Ehsan Barkhordar

Nischal Reddy Chandra

Farhana Ferdousi Liza

Sujal Maharjan, Osman Mutlu

Rajat Patel

Kritesh Rauniyar, Sunil Regmi

Siddhant Bikram Shah, Raghav Sharma, Jiazhao Shi, Shuvam Shiwakoti, Astha Shrestha

Peratham Wiriyathammabhum

Reyyan Yeniterzi, Suveyda Yeniterzi

Vanni Zavarella

Table of Contents

Program

Friday, July 3, 2026

09:00 - 09:10 *Welcome and Opening Remarks*

09:10 - 09:50 *Keynote*

09:50 - 10:30 *Session 1: Multilingual and Low-Resource Event Extraction*

10:30 - 11:00 *Coffee Break*

11:00 - 12:30 *Session 2: LLMs, Generative Methods, and Advanced Event Understanding*

12:30 - 12:40 *Closing Remarks*

Overview of the Workshop on Event Extraction and Understanding: Challenges and Applications

Ali Hürriyetoğlu[1], Surendrabikram Thapa[2], Hristo Tanev[3], Laxmi Thapa[4], Surabhi Adhikari[5]
[1]Wageningen Food Safety Research, Netherlands, [2]Virginia Tech, USA,
[3]European Commission, Joint Research Centre, Italy
[4]O.P. Jindal Global University, India, [5]Columbia University, USA
[1]ali.hurriyetoglu@wur.nl, [2]sbt@vt.edu, [3]hristo.tanev@ec.europa.eu

Abstract

This paper presents an overview of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA 2026), held in conjunction with ACL 2026. Formerly known as CASE, the workshop continues its mission of bringing together researchers from natural language processing, machine learning, computational social science, and related disciplines to advance research on event extraction and understanding. This year's edition particularly emphasized the growing influence of large language models (LLMs), multimodal learning, and weakly supervised methodologies in event extraction research. The workshop featured six regular research papers covering topics such as low-resource event extraction, reflective multi-agent architectures, symbolic auditing of procedural events, geopolitical event extraction, and generative event extraction strategies. In addition, EEUCA 2026 hosted two shared tasks focusing on toxicity detection in gaming communities and multimodal vaccine-critical meme analysis, attracting broad international participation and encouraging research on socially impactful applications of AI. The workshop highlights current advances, emerging challenges, and future directions in multilingual, multimodal, and socially aware event extraction systems.

1 Introduction

The increasing availability of large-scale digital data has transformed the study of events across social, political, economic, and public health domains (Chen et al., 2024; Thapa et al., 2025a). News articles, social media posts, online discussions, multimodal content, and user-generated media continuously document evolving real-world events, creating new opportunities for automated event extraction and understanding (Liu et al., 2020; Shah, 2016; Thapa et al., 2025b; Hey et al., 2025). At the same time, modern societal challenges such as misinformation, political polarization, online extremism, public health crises, and humanitarian emergencies have increased the need for reliable AI systems capable of identifying, structuring, and interpreting complex event information from heterogeneous data sources (Hürriyetoğlu et al., 2025).

Recent advances in large language models (LLMs), multimodal transformers, and instruction-tuned generative systems have significantly reshaped the event extraction landscape (Meng et al., 2024; Liu et al., 2025b; Chen et al., 2024). Contemporary models are increasingly capable of performing event detection, argument extraction, temporal reasoning, and cross-document event understanding with reduced supervision and improved generalization. At the same time, these systems introduce new research questions related to hallucination, interpretability, schema flexibility, multilingual robustness, and ethical deployment. In parallel, multimodal learning has expanded the scope of event analysis beyond traditional text-based pipelines by enabling systems to jointly reason over textual, visual, and contextual information in domains such as misinformation analysis, social media discourse, and crisis monitoring (Liu et al., 2025a; Li et al., 2024; Suwannahong et al., 2026; Ma et al., 2025).

Against this backdrop, the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA 2026), formerly known as CASE, continues to provide an interdisciplinary venue for researchers working at the intersection of NLP, machine learning, computational social science, and AI-driven event analytics. Building on previous editions of CASE, now EEUCA, this year's workshop places particular emphasis on multilingual event extraction, multimodal reasoning, generative AI, low-resource settings, and societally relevant applications of event understanding systems (Hürriyetoğlu et al., 2025, 2024, 2023).

EEUCA 2026 featured six regular research papers spanning a broad range of topics, including

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 1–7

weakly supervised Vietnamese event extraction, benchmark creation for Nepali event extraction, reflective multi-agent event extraction architectures, symbolic auditing of procedural event datasets, generative event extraction training strategies, and LLM-driven geopolitical event extraction pipelines. Collectively, these papers reflect the increasing diversity of methodologies and applications emerging in the event extraction community.

In addition to regular papers, the workshop hosted two shared tasks designed to encourage research on socially relevant and multimodal problems. The first shared task focused on understanding toxic behavioral intent in gaming chat logs using the GameTox dataset, while the second addressed multimodal identification of vaccine-critical content on social media using the VaxMeme dataset. These shared tasks attracted broad international participation and highlighted the growing importance of multimodal reasoning, domain adaptation, and robust classification under severe class imbalance and noisy real-world conditions.

This overview paper summarizes the accepted papers, shared tasks, and broader research themes represented at EEUCA 2026. We further discuss emerging trends in event extraction research, including the role of LLMs, multimodal systems, weak supervision, and reflective reasoning architectures, while outlining future directions for building more reliable, interpretable, and socially responsible event understanding systems.

2 Accepted Papers

This year, 6 regular papers were accepted. Below, we provide brief descriptions of accepted papers:

Hieu et al. (2026) proposed a weakly supervised framework for Vietnamese event extraction that addresses the scarcity of annotated data by constructing a large-scale silver corpus from unlabeled news articles. Their approach first pseudo-labels Vietnamese news data using an existing BKEE event extraction model, then applies a cross-document n-ary relation filtering strategy that retains event structures consistently observed across multiple articles discussing the same topic, thereby reducing noisy annotations. The authors further improve diversity through schema-based data augmentation, where event schemas containing triggers and arguments are diversified and instantiated using LLMs to generate structurally valid synthetic training examples. Built on top of the FourIE joint event extraction architecture, their framework expands the training data to over 46,000 silver-labeled sentences and achieves consistent improvements on the Vietnamese BKEE benchmark, particularly for entity mention detection and event argument extraction.

Maharjan et al. (2026) introduced NepEE, a manually annotated benchmark dataset for low-resource Nepali event extraction, focusing on trigger phrase identification and event type classification in morphologically complex Devanagari text. The dataset contains 10,226 Nepali sentences annotated by five native speakers through a rigorous three-phase annotation protocol involving pilot annotation, instruction refinement, and conflict resolution, resulting in high inter-annotator agreement scores of Fleiss' κ = 0.812 for trigger identification and $\kappa = 0.855$ for event classification. The authors defined eight event categories, including political, economic, health, disaster, and education events, and developed detailed guidelines to capture complex Nepali trigger structures such as nominalized triggers and compound verb constructions. They further benchmarked a broad range of approaches, including classical machine learning models, multilingual and Indic-specific Transformer encoders, and instruction-tuned LLMs under zero-shot and few-shot prompting settings. Experimental results showed that Indic-specialized Transformer models achieved the strongest performance for event classification, while generative LLMs struggled with exact trigger span extraction due to Nepali's morphological complexity and ambiguity in trigger boundaries.

Rim et al. (2026) presented a symbolic audit framework for procedural event annotations by introducing Entity Qualia Structure (EQS), a lexical-semantic representation grounded in Generative Lexicon theory to distinguish semantically central entity state changes from incidental ones in procedural text. Using lexical resources such as the Brandeis Semantic Ontology, CoreLex, and WordNet, the authors categorized entities into coarse sortal types including natural, artifactual, and instrument classes, and applied this framework to the OpenPI food-domain procedural dataset. Their analysis revealed that only 51.1% of OpenPI transformation annotations corresponded to actual

food entities, while 30.2% tracked incidental instrument-related state changes such as bowls, knives, or ovens, highlighting substantial annotation noise in procedural state tracking datasets. The study further compared EQS-based filtering with prior human and LLM-based cleanup methods, demonstrating that the symbolic approach uniquely identified 15.6% of problematic annotations missed by both human re-annotation and LLM salience scoring approaches. Additionally, the authors analyzed the AGENTIVE quale feature and showed that most agentive-positive annotations involved instruments rather than food entities, emphasizing that procedural state interpretation requires compositional reasoning between entity qualia and verb semantics.

Ravikumar and Batista-Navarro (2026) conducted a systematic study of training strategies for generative event extraction, focusing on how event detection (ED) and event argument extraction (EAE) should be coordinated when fine-tuning LLMs. The authors proposed a taxonomy of seven training strategies spanning three paradigms: disjoint training, fully shared training, and hybrid parameter-sharing approaches, and evaluated them across ACE2005 and RichERE using multiple instruction-tuned LLMs ranging from 3B to 12B parameters. Their framework formulated ED, EAE, and joint event extraction as conditional text generation tasks using structured JSON outputs, enabling direct comparison between pipeline and joint generative approaches under consistent settings. Experimental results demonstrated that training strategy substantially affects extraction performance, with the strongest overall results achieved by an "ED Backward Transfer" approach that initializes event detection adapters from pretrained event argument extraction adapters. In contrast, fully joint modeling approaches that generated complete event structures in a single pass consistently underperformed, particularly for trigger classification. The study further showed that event detection benefits from cross-task transfer and partial parameter sharing, whereas argument extraction performs best with dedicated task-specific adapter capacity, highlighting the importance of carefully balancing parameter sharing and specialization in generative event extraction systems.

Tanev et al. (2026) proposed MAREA, a reflective multi-agent architecture for Semi-Open Event Extraction (SOEE) that combines fixed event schema fields with dynamically generated event attributes inferred through self-reflective reasoning using LLMs. Unlike traditional closed-schema event extraction systems, their SOEE framework preserves a core set of standardized fields such as event type, date, and location while allowing the system to iteratively expand templates with context-specific attributes generated at runtime. The proposed architecture consists of three layers: an expert layer that generates initial event templates and answers follow-up questions, a reflective layer that formulates questions to uncover missing or implicit event information, and a coordination layer that manages interactions among agents. The reflective component employs multiple strategies, including BERT-based question mapping, prompt-driven question generation, and keyword-based reasoning, to discover new event attributes beyond the predefined schema. Evaluated on health-related news articles using LLaMA-3.1-70B-Instruct, MAREA achieved strong extraction performance on core event fields such as event type, actors, disease, and mitigation measures, while also generating additional semantically relevant attributes that improved template completeness and contextual richness. The study demonstrates how reflective multi-agent reasoning can support flexible, extensible, and semantically richer event extraction beyond rigid fixed-schema approaches.

Dell'Orto and Kommandeur (2026) introduced GENOME, a continuously updated geopolitical event extraction pipeline and dataset designed to capture both conflictual and cooperative international interactions using LLMs and the PLOVER ontology. Addressing limitations in existing resources such as POLECAT (Halterman et al., 2023), the authors proposed a two-stage extraction and classification framework that leverages GPT-based one-shot prompting with enforced structured outputs to extract events from large-scale English-language newswire data. GENOME extends the traditional Actor–Recipient event representation by introducing a novel Third Party role, enabling richer multi-entity geopolitical representations that better capture complex international interactions and contextual participants. The pipeline further incorporates entity normalization and embedding-based clustering for resolving geopolitical actors,

as well as a multi-criteria deduplication module that merges duplicate reports of the same event across multiple sources. Evaluated against the POLECAT dataset over a five-month overlap period, GENOME demonstrated strong alignment on conflict-related event types while capturing a substantially more balanced distribution of cooperative and verbal interactions, particularly diplomatic consultations and agreements that were largely absent in POLECAT. The study also showed that GENOME more accurately associates events with their inferred occurrence dates rather than publication dates and provides finer-grained geopolitical entity resolution for organizations such as NATO, IMF, and WTO. Overall, the work highlights the potential of LLM-based structured extraction pipelines for building scalable and temporally grounded geopolitical event databases for international relations research and early-warning applications.

3 Shared Task on Understanding Toxic Behavioral Intent in Gaming Chat Logs

The shared task on Understanding Toxic Behavioral Intent in Gaming Chat Logs[1] addressed the challenge of fine-grained toxicity detection in online gaming communities using the GameTox dataset, a large-scale corpus of approximately 53,000 annotated chat utterances collected from the multiplayer game *World of Tanks* (Naseem et al., 2025; Thapa et al., 2026c). Participants were required to classify each utterance into one of six intent categories: *Non-toxic*, *Insults and Flaming*, *Other Offensive Texts*, *Hate and Harassment*, *Threats*, and *Extremism*, reflecting the diverse and highly imbalanced nature of toxic communication in gaming environments. The task highlighted several key challenges specific to gaming chat, including short and noisy utterances, multilingual and code-switched communication, gaming-specific slang, and severe long-tail class imbalance where high-risk categories such as threats and extremism were extremely rare. A total of 102 participants registered for the competition, with 35 teams submitting systems that explored a broad range of approaches, including domain-adaptive pretraining, multilingual transfer learning, supervised contrastive learning, token-attribution guided architectures, ensemble methods, and LLM-based synthetic data augmentation for minority classes. Systems were evaluated using macro-averaged F1-score to emphasize balanced performance across all toxicity categories, and the best-performing system achieved a Macro F1-score of 0.7041. Overall, the shared task provided a comprehensive benchmark for studying toxicity detection in gaming communities and highlighted the importance of domain adaptation, rare-class modeling, and robust multilingual learning for developing safer and healthier online gaming environments.

4 Shared Task on Multimodal Identification of Vaccine Critical Content on Social Media

The shared task on Multimodal Identification of Vaccine Critical Content on Social Media focused on detecting vaccine stance in social media memes[2] using the VaxMeme dataset, a large-scale multimodal collection of over 10,000 vaccination-related memes containing both images and associated textual content (Naseem et al., 2023; Thapa et al., 2026b,a). Participants were tasked with classifying each meme into one of three categories: *Vaccine-critical*, *Neutral*, or *Pro-vaccine*, requiring systems to jointly reason over visual cues, embedded OCR text, sarcasm, humor, and multimodal context. The task highlighted the challenges of multimodal public health misinformation analysis, where stance is often conveyed implicitly through image-text interactions, cultural references, and visual metaphors rather than explicit textual claims alone. A total of 77 participants registered for the competition, with 25 teams submitting systems that explored a wide range of approaches, including transformer-based multimodal architectures, vision-language models, cross-modal attention mechanisms, ensemble strategies, OCR-enhanced pipelines, and instruction-tuned LLMs. Systems were evaluated using macro-averaged F1-score to ensure balanced performance across stance categories despite moderate class imbalance, and the best-performing system achieved a Macro F1-score of 0.8494. Overall, the shared task provided a benchmark for multimodal vaccine stance detection and offered insights into the strengths and limitations of current multimodal AI systems for analyzing vaccine-related discourse, misinformation, and public health narratives on social media platforms.

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

[2] https://www.codabench.org/competitions/12085/

5 Future Direction

The rapid evolution of LLMs, multimodal AI systems, and agentic reasoning frameworks continues to redefine the future of event extraction and understanding research. Future editions of EEUCA will further expand beyond traditional text-centric pipelines toward systems capable of integrating information across multiple modalities, languages, and sources while maintaining robustness, interpretability, and scalability.

One important future direction involves multilingual and low-resource event extraction. Despite recent progress, many languages still lack sufficiently large annotated corpora, standardized schemas, and benchmark datasets. Future research must continue to explore weak supervision, synthetic data generation, transfer learning, and cross-lingual adaptation techniques that can improve event extraction capabilities for underrepresented languages and regions. Building multilingual and culturally aware event extraction systems remains essential for ensuring equitable global coverage of socio-political and public health events.

Another major research direction concerns multimodal event understanding. Increasingly, important real-world events are communicated not only through text but also through images, videos, memes, and multimodal social media content. The success of this year's shared tasks demonstrates both the promise and the difficulty of multimodal reasoning in socially sensitive settings such as misinformation detection and online toxicity analysis. Future work should focus on more robust cross-modal alignment, multimodal temporal reasoning, sarcasm and implicit intent detection, and multimodal explainability techniques capable of identifying how visual and textual signals jointly contribute to model predictions.

The workshop also highlights growing interest in reflective and semi-open event extraction architectures powered by LLMs. Future systems may increasingly move beyond rigid fixed-schema extraction toward adaptive frameworks capable of dynamically discovering new event attributes, reasoning over incomplete information, and interacting with external knowledge sources. Agentic AI systems combining retrieval, reasoning, verification, and self-reflection may play an important role in improving event completeness, factual grounding, and temporal consistency.

Another critical challenge concerns reliability, fairness, and evaluation. While LLM-based systems have demonstrated impressive generative capabilities, they remain susceptible to hallucination, bias, instability, and poor calibration in high-stakes applications. Future research should therefore prioritize more rigorous evaluation methodologies, uncertainty estimation, bias auditing, and interpretable reasoning frameworks for event extraction systems. Human-in-the-loop evaluation, symbolic validation, and hybrid neuro-symbolic approaches may become increasingly important for ensuring trustworthy event extraction pipelines.

Future research could also explore tighter integration of event extraction with related NLP tasks highlighted in this workshop edition, such as abusive language and hate speech detection, misinformation analysis, stance detection, question answering, and sentiment analysis. Future EEUCA shared tasks will continue to focus on realistic and societally relevant challenges involving multimodal analysis, multilingual event understanding, misinformation, public health communication, and emerging online harms. We also aim to further support participation from early-career researchers and underrepresented communities through mentorship opportunities, collaborative initiatives, and accessible benchmark resources.

6 Conclusion

The 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA 2026) reflects the continued growth and diversification of the event extraction research community. This year's workshop showcased advances spanning low-resource event extraction, multimodal reasoning, generative architectures, symbolic auditing, geopolitical event analysis, and socially grounded shared tasks involving gaming toxicity and vaccine-related misinformation. The accepted contributions demonstrate how modern event extraction research is increasingly shaped by LLMs, multimodal AI, and interdisciplinary applications that extend far beyond traditional information extraction settings.

The workshop further highlighted both the opportunities and the challenges introduced by rapidly evolving AI technologies. While LLMs and multimodal systems have substantially expanded the capabilities of event extraction pipelines, important questions remain regarding reliability, interpretability, fairness, multilingual inclusivity, and responsible deployment. Through its combination of

regular papers, shared tasks, and interdisciplinary collaboration, EEUCA continues to provide a platform for advancing research on robust, scalable, and socially responsible event understanding systems.

Broader Impact

EEUCA 2026 contributes to the broader advancement of socially impactful AI research by promoting event extraction technologies that support applications in public health monitoring, misinformation analysis, online safety, humanitarian response, and socio-political understanding. The workshop encourages interdisciplinary collaboration between NLP researchers, computational social scientists, and domain experts working on real-world societal challenges.

This year's shared tasks particularly emphasized socially relevant applications involving online toxicity in gaming communities and vaccine-related misinformation on social media, both of which have direct implications for digital well-being, public discourse, and public health communication. By fostering research on multilingual, multimodal, and low-resource event understanding systems, the workshop also supports the development of more inclusive AI technologies capable of addressing global and culturally diverse contexts.

At the same time, the workshop recognizes the ethical challenges associated with automated event extraction and large-scale content analysis. Event extraction systems may inherit societal biases, misinterpret context, or produce harmful outputs when deployed without appropriate safeguards. Multimodal and LLM-based systems are additionally vulnerable to hallucination, misinformation propagation, privacy concerns, and unfair treatment of marginalized groups. Accordingly, EEUCA promotes responsible AI practices, transparent evaluation methodologies, and research that prioritizes fairness, accountability, and human-centered deployment considerations in event extraction technologies.

References

Ruirui Chen, Chengwei Qin, Weifeng Jiang, and Dongkyu Choi. 2024. Is a large language model a good annotator for event extraction? In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17772–17780.

Alessandro Dell'Orto and Jesse Kommandeur. 2026. Genome: A new geopolitical event methodology and dataset using large language models. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Andrew Halterman, Benjamin E Bagozzi, Andreas Beger, Phil Schrodt, and Grace Scraborough. 2023. Plover and polecat: A new political event ontology and dataset. In *International Studies Association Conference Paper*.

Spencer Phillips Hey, Julie Walsh, and Eni Mustafaraj. 2025. Comparing human and llm ethical analyses: A case study in computational social science research. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 1245–1254.

Pham Xuan Hieu, Tuan Vu Minh, Mai-Vu Tran, and Hoang-Quynh Le. 2026. Constructing a silver corpus for weakly supervised vietnamese event extraction using cross-document n-ary relation filtering. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Ali Hürriyetoğlu, Hristo Tanev, Osman Mutlu, Surendrabikram Thapa, Fiona Anting Tan, and Erdem Yörük. 2023. Challenges and applications of automated extraction of socio-political events from text (CASE 2023): Workshop and shared task report. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 167–175, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Ali Hürriyetoğlu, Surendrabikram Thapa, Hristo Tanev, and Surabhi Adhikari. 2025. Findings and insights from the 8th workshop on challenges and applications of automated extraction of socio-political events from text. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts*, pages 1–5.

Ali Hürriyetoğlu, Surendrabikram Thapa, Gökçe Uludoğan, Somaiyeh Dehghan, and Hristo Tanev. 2024. A concise report of the 7th workshop on challenges and applications of automated extraction of socio-political events from text. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 248–255.

Haoxuan Li, Zhengmao Yang, Yunshan Ma, Yi Bin, Yang Yang, and Tat-Seng Chua. 2024. Mm-forecast: a multimodal approach to temporal event forecasting with large language models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 2776–2785.

Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, Xin Meng, Fei Richard Yu, Xiangyang Ji, and Ming Li. 2025a. Eventgpt: Event stream understanding with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29139–29149.

Tianpeng Liu, Feng Xue, Jian Sun, and Xiao Sun. 2020. A survey of event analysis and mining from social multimedia. *Multimedia Tools and Applications*, 79(45):33431–33448.

Wenxuan Liu, Zixuan Li, Long Bai, Yuxin Zuo, Daozhu Xu, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. 2025b. Towards event extraction with massive types: Llm-based collaborative annotation and partitioning extraction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34365–34387.

Junxiao Ma, Jingjing Wang, Jiamin Luo, Peiying Yu, and Guodong Zhou. 2025. Sherlock: Towards multi-scene video abnormal event extraction and localization via a global-local spatial-sensitive llm. In *Proceedings of the ACM on Web Conference 2025*, pages 4004–4013.

Sujal Maharjan, Astha Shrestha, Lakshmojee Koduru, Sweta Poudel, Shuvam Shiwakoti, Rabin Thapa, Kritesh Rauniyar, and Surendrabikram Thapa. 2026. Benchmarking models for low-resource nepali event extraction with trigger phrase identification and event classification. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Zihao Meng, Tao Liu, Heng Zhang, Kai Feng, and Peng Zhao. 2024. Cean: Contrastive event aggregation network with llm-based augmentation for event extraction. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 321–333.

Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G Dunn. 2023. A multimodal framework for the identification of vaccine critical memes on twitter. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 706–714.

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, Albuquerque, New Mexico. Association for Computational Linguistics.

Rishi Ravikumar and Riza Batista-Navarro. 2026. When tasks share structure: A comparative study of training strategies for generative event extraction. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Kyeongmin Rim, Marc Verhagen, and James Pustejovsky. 2026. A qualia-based audit of procedural event annotations. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Rajiv Ratn Shah. 2016. Multimodal analysis of user-generated content in support of social media applications. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 423–426.

Phattaradanai Suwannahong, Piphatpong Wannapanon, Aphisara Klayburee, Takafumi Nakanishi, and Ponlawat Chophuk. 2026. Explainable multimodal earthquake event classification using aime attribution and llm-based narrative reasoning. In *2026 18th International Conference on Knowledge and Smart Technology (KST)*, pages 598–603. IEEE.

Hristo Tanev, Michel de Bollivier, and Bertrand De Longueville. 2026. A self-reflective llm-based architecture for semi-open event extraction. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Laxmi Thapa, Aryaman Jain, Lakshmojee Koduru, Surabhi Adhikari, Junaid Rashid, Jungeun Kim, Surendrabikram Thapa, and Usman Naseem. 2026a. Concept-grounded detection of vaccine misinformation in multimodal content using interpretable vision-language models. In *Companion Proceedings of the ACM on Web Conference 2026*.

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

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Kritesh Rauniyar, Laxmi Thapa, Surabhi Adhikari, Kristina T. Johnson, Ali Hürriyetoğlu, Hristo Tanev, and Usman Naseem. 2026b. Multimodal identification of vaccine content stance on social media. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Kritesh Rauniyar, Laxmi Thapa, Surabhi Adhikari, Kristina T. Johnson, Ali Hürriyetoğlu, Hristo Tanev, and Usman Naseem. 2026c. 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)*.

Surendrabikram Thapa, Hariram Veeramani, Liang Hu, Qi Zhang, Wei Wang, and Usman Naseem. 2025b. A multimodal prompt-based framework for analyzing code-mixed and low-resource memes. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1913–1923.

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.

Multimodal Identification of Vaccine Content Stance on Social Media

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

Vaccination-related memes on social media play an increasingly influential role in shaping public perception of immunization, often spreading both supportive messaging and vaccine-critical narratives through multimodal communication. Detecting such content is challenging due to the combined use of images, embedded text, sarcasm, humor, and cultural references. This paper presents an overview of the Shared Task on Multimodal Identification of Vaccine Critical Content on Social Media, organized as part of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA 2026) at ACL 2026. The task is based on the VaxMeme dataset, a large-scale collection of vaccination-related memes annotated into three classes: *Vaccine-critical*, *Neutral*, and *Pro-vaccine*. A total of 77 participants registered for the competition, with 25 teams submitting systems for evaluation. Participating approaches included transformer-based multimodal architectures, vision-language models, ensemble methods, and instruction-tuned large language models. The best-performing system achieved a macro F1-score of 0.8494. This shared task provides insights into the strengths and limitations of current multimodal approaches for vaccine stance detection and highlights future directions for robust public health misinformation analysis.

1 Introduction

Social media platforms have become primary arenas for public health discourse, where information about vaccines, treatments, and disease prevention spreads with unprecedented speed and reach (Shah et al., 2024b). Among the many forms of digital communication, memes—multimodal artifacts that fuse images and embedded text into compact, virally shareable units—have emerged as particularly influential vehicles for shaping public attitudes toward vaccination (Naseem et al., 2023; Ahmad et al., 2025). The COVID-19 pandemic dramatically illustrated both the promise and the peril of this medium: while memes proved effective at promoting awareness and disseminating accurate health information, they also became powerful conduits for vaccine misinformation, conspiracy theories, and skepticism that contributed to vaccine hesitancy and eroded public trust in immunization programs (Thapa et al., 2024b).

The challenge of identifying vaccine-critical content in memes is fundamentally compounded by the medium's inherent ambiguity. Vaccine-related memes frequently rely on sarcasm, irony, visual metaphor, and culturally specific references that obscure intent and complicate automated analysis (Pramanick et al., 2021). The boundary between legitimate critical commentary, satirical humor, and harmful misinformation is often deliberately blurred, with hateful or misleading content embedded within seemingly benign visual frames (Shah et al., 2024a). Single-modality approaches—whether text-only or image-only—consistently fail to capture this layered communicative intent, as the meaning of a vaccine-related meme typically emerges from the interplay between its visual and textual components rather than from either modality in isolation. This makes vaccine-critical content detection a paradigmatic case for multimodal reasoning, requiring systems to jointly interpret visual cues, embedded text, surrounding captions, and broader sociocultural context.

Despite the public health importance of this problem, computational resources for vaccine-critical meme detection remain limited. Most prior work on health misinformation has focused on text-only

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 17–25

analysis of social media posts (Karafillakis et al., 2021). The scarcity of large-scale, publicly accessible, and richly annotated multimodal benchmarks has hindered systematic progress, particularly for capturing the diverse and evolving repertoire of vaccine-critical narratives that circulate across social media platforms.

To address these gaps, we present the **Shared Task on Multimodal Identification of Vaccine Critical Content on Social Media**, 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 upon the VaxMeme dataset (Naseem et al., 2023), a corpus of over 10,000 manually annotated vaccination-related memes spanning multiple platforms and timelines, designed specifically to support the development of multimodal vaccine-critical content detection systems. Participating systems are required to classify each meme into one of three categories: (1) *Vaccine Critical*: memes that criticize vaccines, contain vaccine misinformation, propagate conspiracy theories, or argue against vaccination, (2) *Neutral*: memes that report vaccine-related events or opinions objectively without taking a stance, or (3) *Pro-Vaccine*: memes that advocate for vaccination, promote awareness, or support immunization efforts. The task is evaluated using macro-averaged F1-score to ensure balanced performance across the three classes despite their natural distributional differences.

The shared task attracted a diverse range of participating teams employing both text-centric and multimodal approaches for vaccine stance classification. Submitted systems explored a variety of strategies, including transformer-based architectures, vision-language modeling, multimodal fusion techniques, ensemble methods, and instruction-tuned large language models. The diversity of submissions highlights the growing interest in multimodal public health content analysis and provides useful insights into the strengths and limitations of current approaches for vaccine-related meme understanding. This paper provides a comprehensive overview of the shared task, including a detailed description of the VaxMeme dataset and its annotation protocol, the evaluation methodology, 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 multimodal vaccine-critical content detection, support myth-debunking efforts, and contribute to the design of more effective public health communication strategies on social media platforms.

2 Related Works

Stance detection classifies whether a given text or image expresses support, opposition, or neutrality toward a specified target, distinguishing it from general sentiment analysis in that the inferred position is inherently relational and target-conditioned (Shiwakoti et al., 2024; Thapa et al., 2024a). Early work showed that stance can benefit from discourse-level information, since agreement and disagreement between utterances provide useful evidence beyond isolated lexical features (Thomas et al., 2006). The field has since expanded to social media, where short, informal, and context-dependent posts make stance recognition challenging due to implicit targets, sarcasm, and limited conversational context (Küçük and Can, 2020). However, most detection tasks remain primarily text-based, which limits their applicability to memes where stance may be encoded through visual framing, image-text incongruity, or culturally specific references (Küçük and Can, 2020; Kiela et al., 2020).

Research on misinformation detection has shown that misleading content can often be identified through linguistic, stylistic, and factuality-related cues, but such approaches usually focus on veracity rather than stance toward a public-health target (Rashkin et al., 2017). Vaccine misinformation is particularly consequential, as exposure via social media and coordinated online disinformation campaigns have been empirically associated with diminished vaccine confidence and elevated vaccine hesitancy (Wilson and Wiysonge, 2020). Experimental evidence further shows that exposure to COVID-19 vaccine misinformation can reduce vaccination intent, demonstrating the public-health importance of detecting harmful vaccine narratives online (Loomba et al., 2021). Nevertheless, vaccine-critical content subsumes misinformation, encompassing expressions of distrust, opposition, conspiracy framing, sarcasm, and satire that do not necessarily advance a directly verifiable factual claim (Rashkin et al., 2017; Loomba et al., 2021).

Multimodal meme analysis has established that image and text require joint interpretation, as each modality in isolation may appear benign while their combination conveys harmful or oppositional

meaning (Kiela et al., 2020; Thapa et al., 2025). Pramanick et al. (2021) demonstrated that harmful meme detection is improved by jointly modeling global meme-level context and local visual-textual cues, underscoring the necessity of fine-grained multimodal representations. In the vaccine domain, VaxMeme introduced a large manually annotated dataset of vaccine-critical memes and demonstrated that multimodal modeling is necessary for capturing the contextual and visual-textual signals present in vaccine discourse (Naseem et al., 2023). However, prior multimodal meme studies have largely focused on hate, harm, or binary misinformation labels, leaving a gap for systematic evaluation of multi-class vaccine content in social media memes (Kiela et al., 2020; Pramanick et al., 2021; Naseem et al., 2023).

Transformer-based language models such as BERT and RoBERTa established strong general-purpose representations for downstream NLP classification through large-scale pretraining and task-specific fine-tuning (Devlin et al., 2019; Liu et al., 2019). Vision-language transformers such as ViLBERT (Lu et al., 2019), VisualBERT (Li et al., 2019), and UNITER (Chen et al., 2020) extended this paradigm to multimodal learning by jointly encoding visual regions and textual tokens through cross-modal attention or image-text pretraining objectives. Contrastive and generative vision-language models such as CLIP (Radford et al., 2021), Flamingo (Alayrac et al., 2022), and BLIP-2 (Li et al., 2023) further improved transferability, few-shot adaptation, and instruction-following capabilities by scaling image-text supervision and integrating pretrained language models with visual encoders. Despite these advances, the classification task in the vaccine dataset remains non-trivial, as OCR noise, sarcasm, visual metaphor, and rapidly evolving sociopolitical narratives collectively undermine models trained on generic image-text corpora. The shared task features a multimodal dataset designed to engage the research community and encourage investigation into the identification and analysis of vaccine-critical content on social media platforms.

3 Shared Task Description

This shared task focuses on the automated understanding of vaccination-related memes through a multimodal lens, targeting the detection of vaccine stance in online content.

Task: Vaccine Stance Classification. Given a meme consisting of an image and associated textual content, participating systems must classify its stance towards vaccination into one of three categories: *Pro-vaccine*, *Vaccine-critical*, or *Neutral*. *Pro-vaccine* memes promote vaccination, highlight its benefits, or encourage positive health behaviors. *Vaccine-critical* memes express skepticism, opposition, or criticism towards vaccines, which may include misinformation, conspiracy narratives, or sarcastic undermining of vaccination efforts. *Neutral* memes present vaccination-related content without a clear stance, often conveying informational or ambiguous messages.

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

4 Dataset

The shared task is based on the *VaxMeme* dataset (Naseem et al., 2023; Thapa et al., 2026), a large-scale multimodal collection of vaccination-related memes. The dataset consists of 10,244 memes collected from Twitter, where each instance contains both an image and associated textual content (including OCR-extracted text when embedded in images).

4.1 Data Collection and Annotation

The dataset was constructed by collecting tweets containing both images and text between October 2020 and April 2021 using the Twitter API. Non-English content was excluded. Each meme was annotated by multiple human annotators with strong linguistic proficiency, following detailed annotation guidelines. Disagreements were resolved via majority voting with an additional annotator when required. The annotation quality is high, with substantial inter-annotator agreement (Fleiss' $\kappa = 0.85$).

Each meme is labeled into one of three stance categories: *Vaccine-critical* (0), *Neutral* (1), and *Pro-vaccine* (2).

4.2 Dataset Split

For the shared task, the dataset is split into training, validation, and test sets using an 80/10/10 ratio. The splits are designed to maintain a realistic and slightly imbalanced class distribution.

As shown in Table 1, class distribution reflects real-world conditions, where pro-vaccine content

Label	Train	Val	Test	Total
Vaccine-critical	2535	308	314	3157
Neutral	2461	327	316	3104
Pro-vaccine	3199	389	395	3983
Total	8064	1025	1024	10113

Table 1: Dataset statistics for the shared task.

is slightly more prevalent, while vaccine-critical and neutral memes appear in comparable proportions. This subtle imbalance makes the task more realistic and encourages the development of robust multimodal models.

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 77 participants registered, out of which 25 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 were 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, 2026.

[1] https://www.codabench.org/competitions/12085/

6 Participants' Methods

LilyMeme (Li, 2026) built upon the MemeCLIP framework (Shah et al., 2024a), introducing a series of targeted enhancements for the VaxMeme vaccine stance detection task. The input is restructured using a `[POST]/[IMG]` template that explicitly separates post text from OCR-extracted image text, with `[NO_POST]` and `[NO_OCR]` markers for missing modalities, and the original element-wise fusion is replaced by a lightweight two-layer, eight-head cross-modal Transformer that models token-level image–text interactions. Training is further strengthened through noise-aware sample weighting, which derives per-instance confidence scores via nearest-neighbour consistency analysis and downweights ambiguous or likely mislabelled samples, and an auxiliary LLM description branch using Qwen2.5-VL-7B-Instruct that supplements memes with poor OCR quality. Inference-stage refinement combines test-time augmentation with a retrieval-augmented k-nearest-neighbour prior interpolated against the parametric model output, and the final submission ensembles multiple complementary variants trained across different cross-validation folds and visual backbones (CLIP ViT-L/14 and EVA02-L-14). The system achieved a macro F1 of 0.8494, securing 1st place overall.

CUET_SYNTHETICA (Zaman et al., 2026) proposed a gated cross-modal attention framework combining Twitter-RoBERTa for text encoding with CLIP ViT-L/14 for visual feature extraction. Textual inputs concatenated post-text with OCR-extracted meme overlay text, while both modalities were projected into a shared 512-dimensional fusion space. A learned scalar gate dynamically balanced cross-attended image representations against raw text features, suppressing uninformative visual signals. Final predictions were produced via a three-model weighted ensemble incorporating a text-only classifier, the full multimodal model, and a variant retrained on combined training and validation data. Their system achieved a test

Rank	Username	F1 Macro	Accuracy	Precision	Recall
1	lili12-637947 (Li, 2026)	0.8494	0.8517	0.8494	0.8517
2	wangxiuxian-637268	0.8389	0.8420	0.8386	0.8409
3	rishta_19-611897 (Zaman et al., 2026)	0.8357	0.8390	0.8383	0.8359
4	_alexcristea-636983 (Cristea and Ionescu, 2026)	0.8340	0.8380	0.8338	0.8351
5	sumaiya_110-594217 (Zaman et al., 2026)	0.8332	0.8361	0.8345	0.8340
6	anchy-637928	0.8308	0.8341	0.8309	0.8309
7	myname-637930	0.8308	0.8341	0.8309	0.8309
8	quasar-637336 (Chowdhury and Chowdhury, 2026)	0.8306	0.8322	0.8331	0.8324
9	wenbin-634065 (Shen, 2026)	0.8205	0.8244	0.8205	0.8218
10	naturia_beast-636958	0.8201	0.8244	0.8212	0.8209
11	vinaybabu-637935	0.8184	0.8215	0.8216	0.8190
12	ratpier-637076	0.8150	0.8176	0.8170	0.8161
13	yjwong1999-494691	0.8122	0.8137	0.8189	0.8141
14	linus-637363 (Acharya and Regmi, 2026)	0.8105	0.8137	0.8106	0.8123
15	havnis-636808	0.8067	0.8117	0.8080	0.8083
16	alishba-wazir-604227	0.8067	0.8088	0.8132	0.8071
17	zmin123-553584	0.7997	0.8039	0.8005	0.8013
18	lin123-637530	0.7994	0.8039	0.7992	0.8007
19	barkion-636765	0.7976	0.7990	0.8080	0.7986
20	merrli-636903	0.7972	0.7990	0.8058	0.7982
21	exterio-636705	0.7861	0.7912	0.7964	0.7846
22	abs123-504332	0.7846	0.7912	0.7868	0.7864
23	thatgrass-519137	0.7754	0.7844	0.7858	0.7802
24	wangkongqiang-637899 (Wang et al., 2026)	0.7552	0.7600	0.7652	0.7560
25	kannanrrk-615633	0.7436	0.7502	0.7435	0.7437

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.

Macro F1 of 0.8357, ranking 3rd overall.

_alexcristea (Cristea and Ionescu, 2026) proposed a text-only early-fusion pipeline that skips visual encoders, instead extracting embedded meme text via OCR and concatenating it with the social media post before processing the unified sequence through an ERNIE-2.0-Large encoder. To reduce overfitting on noisy, label-ambiguous meme data, the standard classification head was replaced with a Multi-Sample Dropout architecture using five parallel dropout masks, acting as an implicit ensemble within a single forward pass. Trained with inverse class-weighted Cross-Entropy loss, the system achieved a Macro F1 of 0.8340, ranking 4th overall.

Quasar (Chowdhury and Chowdhury, 2026) presented a comprehensive ablation-driven system for three-class vaccine stance detection in social media memes, systematically evaluating text-only models (TF-IDF, BERT, RoBERTa, and DeBERTa variants), image-only models (ResNet-50, ViT, Swin, ConvNeXt, EfficientNet, CLIP Vision), and multimodal models (CLIP, BLIP, LLaVA) across multiple preprocessing and augmentation configurations. A domain-specific text normalisation pipeline preserves stance-indicative tokens such as emojis and hashtags, images are uniformly enhanced via contrast, brightness, and sharpness scaling, and balanced class oversampling — identified as the single most impactful intervention, adding approximately 4–5 macro F1 points across all model families — is applied to address the moderate class imbalance in the VaxMeme dataset. The final system combines DeBERTa-v3-large, RoBERTa-large, and CLIP multimodal (ViT-B/32) via soft voting with weights proportional to individual validation macro F1, achieving a macro F1 of 0.8306 and placing 8th out of 25 participating teams.

wenbin-634065 (Shen, 2026) introduced MoEs-VaxAgent, a hybrid discriminative-generative pipeline addressing both standard and boundary-ambiguous meme samples. Feature extraction draws on RoBERTa, ViT, CLIP, and Sentence-BERT, with the last encoder processing domain-relevant passages retrieved from MMCoVaR as external knowledge, producing five modality-specific expert representations dynamically aggregated via a learnable Top-2 gating network. Samples where the Mixture-of-Experts (MoE) classifier yields low-confidence predictions are subsequently re-evaluated by a

trio of LLM agents, namely a text agent, a visual agent, and a judge agent for conflict resolution, before a final label is assigned. The framework achieved a Macro F1 of 0.8205, ranking 9th overall.

Linus (Acharya and Regmi, 2026) compared text-only and multimodal late-fusion approaches for vaccine-critical meme classification using a shared three-layer feedforward classification head across all configurations. The multimodal systems combined CLIP ViT-B/32 image features with BERT-family text encoders via L2-normalized concatenation, while the text-only systems fine-tuned five encoders, namely BERT-base-uncased, RoBERTa-base, ModernBERT-base, DistilBERT-base, and DeBERTa-v3-base, on post text alone. Contrary to expectations, text-only models consistently outperformed their multimodal counterparts, with BERT-base-uncased achieving the best test Macro F1 of 0.8102, ranking 14th overall.

wangkongqiang (Wang et al., 2026) explored a wide range of supervised learning approaches for the multimodal identification of vaccine-critical content, evaluating both fine-tuned pre-trained transformer encoders and instruction-tuned large language models. The pre-trained model branch included ALBERT, BERT, ERNIE, and RoBERTa variants, with additional architectural augmentations such as RNN, CNN, and LSTM layers stacked on top of RoBERTa, and a hard voting ensemble over the four strongest variants. The LLM branch fine-tuned Qwen2-1.5B, Qwen2-7B, Llama2-7B, and Llama3-8B using the Llama-Factory framework with prompts composed of post text, image text, and selectable label types. The best-performing system was the fine-tuned Qwen2-1.5B LLM, achieving a Macro F1 of 0.8153, accuracy of 0.8185, and ranking 12th overall, demonstrating that smaller instruction-tuned LLMs can compete with larger variants when computational resources are constrained.

CSECU-Learners (Ahmad and Uddin, 2026) proposed a two-stage early fusion framework integrating three transformer-based encoders for vaccine-critical meme detection. The architecture combined Twitter-RoBERTa for textual encoding, Vision Transformer (ViT) for visual feature extraction, and Vision-and-Language Transformer (ViLT) for joint cross-modal representations. In Stage 1, the pooler outputs of RoBERTa and ViT were combined via performance-weighted summation with weights derived from validation rankings; in Stage 2, this visual-contextualized representation was concatenated with the ViLT pooler output and passed through a linear classification layer. To mitigate class imbalance, the system was trained with Focal Loss. Their approach achieved a Macro F1 of 0.8308 and accuracy of 0.8341, ranking 6th overall. Ablation studies confirmed that the Stage 1 RoBERTa-ViT fusion contributes most substantially to performance, partly by compensating for ViLT's restrictive 40-token sequence length limit.

7 Discussion

The submitted systems demonstrate the continued importance of multimodal reasoning for understanding vaccine-related discourse on social media. Most high-performing teams combined textual and visual representations through transformer-based architectures, cross-modal attention mechanisms, or ensemble strategies, highlighting that vaccine stance in memes is rarely conveyed through a single modality alone. In many cases, the interaction between image context, embedded OCR text, and accompanying captions was necessary for correctly identifying sarcasm, misinformation, or subtle stance cues.

A notable trend among top-performing submissions was the strong reliance on pretrained vision-language models and domain-adapted language encoders. Several teams incorporated CLIP-based visual representations, Twitter-domain RoBERTa encoders, or instruction-tuned large language models, suggesting that pretrained multimodal knowledge transfers effectively to vaccine-critical meme analysis. Additionally, ensemble methods and hybrid fusion strategies consistently improved robustness, particularly for ambiguous or noisy samples.

Interestingly, some text-only systems remained highly competitive, occasionally outperforming more complex multimodal architectures. This suggests that OCR-extracted textual content and associated captions contain substantial stance-related information in the VaxMeme dataset. However, purely text-based approaches may struggle in cases where stance is conveyed implicitly through visual symbolism, irony, or image-text incongruity. The results therefore indicate that while textual

information remains dominant in many instances, multimodal integration provides complementary contextual signals that improve generalization and robustness.

The competition also highlighted several persistent challenges. Vaccine-related memes often contain sarcasm, cultural references, visual metaphors, and low-quality OCR text, all of which complicate reliable classification. Furthermore, the evolving nature of online vaccine discourse means that models trained on static datasets may face distributional shifts over time. Many systems also relied heavily on large pretrained models, raising concerns regarding computational efficiency, accessibility, and reproducibility.

Future work can explore several promising directions. First, retrieval-augmented and knowledge-grounded systems may help models reason about evolving public health narratives and misinformation trends. Second, finer-grained explainability methods could improve transparency by identifying which textual or visual elements contribute most strongly to predictions. Third, multilingual and cross-cultural extensions of vaccine meme datasets would improve the applicability of these systems beyond English-speaking contexts. Finally, more robust handling of sarcasm, implicit stance, and adversarial meme constructions remains an important open research challenge for multimodal public health content analysis.

8 Conclusion

This shared task presented a benchmark for multimodal vaccine stance classification using the VaxMeme dataset and attracted a diverse range of approaches spanning transformer ensembles, vision-language models, and instruction-tuned LLMs. The results demonstrate that multimodal modeling remains highly effective for identifying vaccine-critical content, while also revealing the continued strength of carefully designed text-centric approaches. Through this shared task, we hope to encourage further research into multimodal public health content understanding, misinformation detection, and socially responsible AI systems for online discourse analysis.

Limitations

This shared task has several limitations. First, the dataset consists only of English-language memes collected from Twitter during a specific period of the COVID-19 pandemic, limiting generalizability across languages, cultures, and platforms. Second, vaccine-related memes often rely on sarcasm, humor, and cultural references that remain difficult for current multimodal systems to interpret reliably. OCR quality and noisy embedded text may also affect model performance. Finally, leaderboard metrics such as macro F1-score do not fully capture robustness, fairness, or real-world deployment challenges.

Ethical Considerations

This shared task involves the analysis of vaccine-related social media content, including misinformation and conspiracy-oriented memes. While such systems may support public health research and misinformation analysis, incorrect predictions could misclassify satire, political commentary, or legitimate criticism. The dataset contains publicly shared social media content and should be used responsibly and only for research purposes. We also acknowledge that multimodal content analysis systems may introduce societal risks if used for surveillance or automated censorship without appropriate human oversight.

References

Darwin Acharya and Sunil Regmi. 2026. Linus@eeuca 2026: Multimodal and text-only approaches to vaccine-critical meme detection. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Monir Ahmad and Md. Saif Uddin. 2026. Csecu-learners@eeuca 2026: Vaccine critical memes identification using two-stage early fusion of transformers. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Syed Talal Ahmad, Haohui Lu, Sidong Liu, Annie Lau, Amin Beheshti, Mark Dras, and Usman Naseem. 2025. Vaxguard: A multi-generator, multi-type, and multi-role dataset for detecting llm-generated vaccine misinformation. *arXiv preprint arXiv:2503.09103*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text

representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Adiba Fairooz Chowdhury and MD Sagor Chowdhury. 2026. Quasar@eeuca 2026: Multimodal deep learning for vaccine stance detection in memes. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Alexandru-Marian Cristea and Costin Ionescu. 2026. _alexcristea@eeuca 2026: A robust early-fusion ernie pipeline for multimodal covid-19 vaccine meme classification. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

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)*.

Emilie Karafillakis, Sam Martin, Clarissa Simas, Kate Olsson, Judit Takacs, Sara Dada, and Heidi Jane Larson. 2021. Methods for social media monitoring related to vaccination: systematic scoping review. *JMIR public health and surveillance*, 7(2):e17149.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Yixuan Li. 2026. Lilymeme@eeuca 2026: Multimodal vaccine meme stance detection with task-adapted memeclip and complementary ensembling. 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*.

Sahil Loomba, Alexandre De Figueiredo, Simon J Piatek, Kristen De Graaf, and Heidi J Larson. 2021. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature human behaviour*, 5(3):337–348.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G Dunn. 2023. A multimodal framework for the identification of vaccine critical memes on twitter. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 706–714.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Momenta: A multimodal framework for detecting harmful memes and their targets. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 4439–4455.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024a. Memeclip: Leveraging clip representations for multimodal meme classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17320–17332.

Siddhant Bikram Shah, Surendrabikram Thapa, Ashish Acharya, Kritesh Rauniyar, Sweta Poudel, Sandesh Jain, Anum Masood, and Usman Naseem. 2024b. Navigating the web of disinformation and misinformation: Large language models as double-edged swords. *IEEE Access*.

Wenbin Shen. 2026. wenbin-634065@eeuca 2026: Moes-vaxagent, a two-stage framework for multimodal vaccine critical meme detection. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Shuvam Shiwakoti, Surendrabikram Thapa, Kritesh Rauniyar, Akshyat Shah, Aashish Bhandari, and Usman Naseem. 2024. Analyzing the dynamics of climate change discourse on twitter: A new annotated corpus and multi-aspect classification. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 984–994.

Laxmi Thapa, Aryaman Jain, Lakshmojee Koduru, Surabhi Adhikari, Junaid Rashid, Jungeun Kim, Surendrabikram Thapa, and Usman Naseem. 2026. Concept-grounded detection of vaccine misinformation in multimodal content using interpretable vision-language models. In *Companion Proceedings of the ACM on Web Conference 2026*.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Jafri, Shuvam Shiwakoti, Hariram Veeramani, Raghav Jain, Guneet Singh Kohli, Ali Hürriyetoğlu, and Usman Naseem. 2024a. Stance and hate event detection in tweets related to climate activism-shared task at case 2024. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 234–247.

Surendrabikram Thapa, Kritesh Rauniyar, Hariram Veeramani, Aditya Shah, Imran Razzak, and Usman Naseem. 2024b. Did you tell a deadly lie? evaluating large language models for health misinformation identification. In *International Conference on Web Information Systems Engineering*, pages 391–405. Springer.

Surendrabikram Thapa, Siddhant Bikram Shah, Kritesh Rauniyar, Shuvam Shiwakoti, Surabhi Adhikari, Hariram Veeramani, Kristina T Johnson, Ali Hürriyetoğlu, Hristo Tanev, and Usman Naseem. 2025. Multimodal hate, humor, and stance event detection in marginalized sociopolitical movements. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts*, pages 20–31.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 327–335.

Kongqiang Wang, Peng Zhang, and Qingli Tan. 2026. wangkongqiang@eeuca 2026: Multimodal identification of vaccine critical content on social media. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Steven Lloyd Wilson and Charles Wiysonge. 2020. Social media and vaccine hesitancy. *BMJ global health*, 5(10).

Sumaiya Zaman, Miftahul Jannat Rishta, and Shiti Chowdhury. 2026. Cuet_synthetica@eeuca 2026: Gated cross-modal attention with domain-adapted text encoding for vaccine-critical meme detection. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Constructing a Silver Corpus for Weakly Supervised Vietnamese Event Extraction using Cross-Document N-ary Relation Filtering

Xuan-Hieu Pham*, **Minh-Tuan Vu***, **Mai-Vu Tran, Hoang-Quynh Le**†
Faculty of Information Technology, VNU University of Engineering and Technology, Vietnam
{20020125, 21020664, vutm, lhquynh}@vnu.edu.vn

Abstract

Event extraction for low-resource languages such as Vietnamese is limited by the lack of large-scale annotated data. To address this, we propose a weakly supervised framework that constructs a silver corpus via pseudo-labeling. We introduce a cross-document n-ary relation filtering strategy to reduce noise by leveraging consistency across multiple articles describing the same event, and further enhance data diversity with schema-based augmentation. Experiments on the BKEE benchmark show consistent improvements, demonstrating the effectiveness of our approach. Data is available at: https://github.com/Larken1612/VietEE2.

1 Introduction

Event Extraction (EE) is a fundamental and challenging task in Information Extraction, aiming to identify structured representations of events described in unstructured text (Xiang and Wang, 2019). An event is typically defined as an occurrence that takes place at a specific time and location, involving one or more entities and often associated with a change of state. Accordingly, EE seeks to detect and structure event-related information from text, including event triggers and their associated arguments (Jurafsky and Martin, 2026).

A common end-to-end formulation of EE decomposes the task into three subtasks (Walker et al., 2006; Liu et al., 2020; Xiang and Wang, 2019): (i) *Entity Mention Detection (EMD)*, which identifies and classifies mentions of real-world entities such as persons, organizations, locations, and temporal expressions; (ii) *Event Detection (ED)*, which identifies event triggers-words or phrases that indicate the occurrence of events, and classifies them into predefined event types. (iii) *Event Argument Extraction (EAE)*, which identifies entities participating in each event and assigns them semantic roles. Figure 1 illustrates an example of EE.

EE remains challenging due to the complexity of event structures and the diverse interactions among their components, motivating a variety of approaches (Kontostathis et al., 2004; Xiang and Wang, 2019). Traditional pipeline-based methods decompose EE into sequential subtasks, i.e., EMD, ED and EAE, where each component is modeled independently. While this modular design allows for task-specific optimization, it suffers from error propagation, as mistakes in earlier stages (e.g., incorrect entity or trigger detection) can adversely affect downstream predictions. To mitigate this issue, recent studies have explored joint learning approaches that model entities, event triggers, and arguments simultaneously, thereby capturing their interdependencies and reducing cascading errors (Nguyen et al., 2021; Wadden et al., 2019; Lin et al., 2020). In this work, we follow this line of research as a strong EE baseline.

Vietnamese EE remains challenging due to its linguistic characteristics, including the lack of explicit word boundaries, strong contextual ambiguity, and the prevalence of multi-word event triggers. These factors make accurate detection of entities and events more difficult. Beyond linguistic challenges, a major bottleneck for Vietnamese EE lies in the scarcity of large-scale annotated data. The recently proposed BKEE dataset (Nguyen et al., 2024), although pioneering, is relatively small and lacks sufficient diversity to cover complex real-world event structures. This limitation significantly restricts the performance of data-driven approaches and motivates the need for scalable alternatives.

Our contributions are as follows. First, we construct a large-scale silver corpus for Vietnamese EE using pseudo-labeling, enhanced by a cross-document n-ary relation filtering strategy to improve label quality. Building upon this resource, we propose a weakly supervised joint learning frame-

* Co-first authors.
† Corresponding authors.

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 26–37

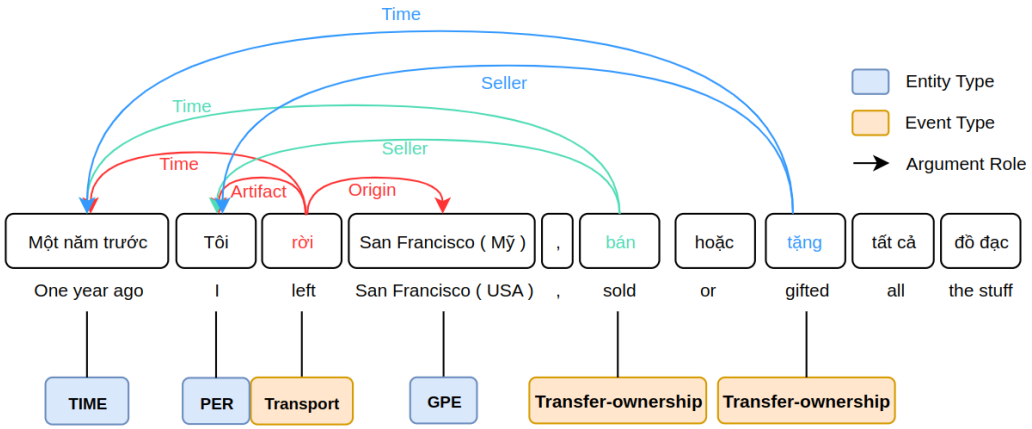


Figure 1: An example of event extraction.

work and demonstrate its effectiveness through baseline experiments on the BKEE benchmark.

2 Related Work

Existing approaches for EE can be broadly categorized into pipeline and joint learning methods (Xie et al., 2021). Pipeline approaches decompose the task into subtasks such as entity mention detection, event detection, and argument extraction, but suffer from error propagation across stages. To address this issue, joint learning models such as DyGIE++ (Wadden et al., 2019), OneIE (Lin et al., 2020) have been proposed to jointly model multiple components of EE. However, these models still have limitations in capturing complex dependencies across tasks and instances. FourIE (Nguyen et al., 2021) further improves joint learning by explicitly modeling inter-task dependencies through instance interaction and type dependency graphs. In this work, we adopt FourIE as the backbone model due to its strong performance and ability to capture structured dependencies, and focus on improving data quality via weak supervision.

Although EE has been extensively studied, most prior work focuses on high-resource languages such as English and Chinese, supported by large-scale datasets like MAVEN (Wang et al., 2020), RAMS (Ebner et al., 2020), and WikiEvents (Li et al., 2021). Multilingual benchmarks such as ACE 2005 (Walker et al., 2006) and TAC KBP (Mitamura et al., 2017) have further facilitated cross-lingual research. However, resources for low-resource languages remain limited. In particular, Vietnamese lacks large-scale annotated datasets for EE, with BKEE (Nguyen et al., 2024) being one of the few available resources, which significantly restricts the development of data-driven approaches.

Weakly supervised learning has emerged as a promising paradigm for EE, leveraging various forms of incomplete, inexact, or imprecise supervision signals to reduce annotation costs. *Distant supervision* generates training data by aligning text with external knowledge sources (e.g., Araki and Mitamura (2018)). While scalable, this approach depends on the availability of knowledge bases and often introduces noisy labels, as the presence of trigger words does not necessarily imply actual events. *Semi-supervised approaches* leverage both labeled and unlabeled data to improve EE. Huang and Ji (2020) learn latent representations for unseen event types, Ferguson et al. (2018) use self-training methods to generate pseudo-labels for unlabeled data. However, these methods are sensitive to pseudo-label quality, as errors can be reinforced during training. Recent work explores *pseudo-labeling strategies* to construct silver corpus. Yao et al. (2020) generate seed event pairs using heuristic patterns and refine them via semantic consistency before expanding with a trained classifier. However, these approaches mainly rely on local contextual signals, which may introduce noise for complex event structures. *Data augmentation* techniques have also been explored to improve model robustness and data diversity, including back-translation (Xie et al., 2020), synonym replacement (Dai and Adel, 2020), contextual rewriting (Yang et al., 2019), and schema-based generation (Jin and Ji, 2024). These methods typically preserve existing labels and operate at the sentence level, without explicitly addressing the quality of supervision signals. Our proposed method combines pseudo-labeling and self-training idea, reduce noise by exploiting cross-document consistency

and further incorporate a schema-based data augmentation strategy to improve data diversity while preserving structural validity.

3 Silver Corpus Construction

The silver corpusset construction process is illustrated in Figure 2. It consists of three main phases: data preparation, cross-document filtering and schema-based data augmentation.

3.1 Data Preparation

We collect a large-scale corpus of unlabeled Vietnamese news articles and organize them into groups based on shared topics or underlying events to support silver corpus construction. We leverage two widely used news aggregation platforms, *Báo Mới*[1] and *Google News*[2], which continuously collect and categorize news articles from multiple sources, providing topic-level grouping of semantically related documents.

We retain their topic assignments to form document groups and segment each article into sentences. As a result, we obtain a collection of sentence sets, where each set corresponds to a group of documents discussing a similar topic. In total, our dataset comprises approximately 72,000 articles, organized into more than 300 topic-based groups, yielding 2,673,796 unlabeled sentences.

To obtain initial annotations, we use BKEE event extraction model (Nguyen et al., 2024) to produce *coarse-grained annotations* for each sentence, including candidate event triggers and argument roles. This step results in a pseudo-labeled corpus that serves as input for subsequent refinement via cross-document filtering.

3.2 Cross-document Filtering

Pseudo-labeled data obtained from the previous stage inevitably contains noise due to model errors and domain mismatch. Based on the observation that real-world events are often reported by multiple news sources, resulting in multiple mentions of the same event across different articles, we propose a cross-document filtering strategy to improve annotation quality (see Table 5 in Appendix A for an example of sentences from different articles within the same topic may describe the same event with consistent triggers and arguments). This cross-document consistency provides a strong signal for distinguishing reliable event structures from noisy predictions.

Given a group of documents discussing the same topic, we first extract *n-ary relations* from pseudo-labeled sentences. Each relation is defined by an event trigger and its associated arguments (e.g., time, location, participants), as predicted by the BKEE model (see Table 6 in Appendix A for examples of n-ary relations extracted from pseudo-labeled sentences within a topic).

We aggregate these relations across documents within the same group and retain those that appear frequently. To implement this approach, we apply *frequency-based filtering* at two levels. First, for each event type et_j, we retain only those appearing in at least μ sentences within the group, i.e.,

$$count(|R_j|, m) \geq \mu. \tag{1}$$

Second, for each n-ary relation r_i associated with event type et_j, we retain it only if its occurrence count exceeds an adaptive threshold θ_j:

$$count(|r_i|, R_j) \geq \theta_j. \tag{2}$$

Let A_j denote the occurrence counts of all candidate n-ary relations associated with et_j. We compute the interquartile range (IQR) of A_j to measure the variability of relation frequencies. If the IQR is sufficiently small relative to the minimum count, i.e.,

$$IQR(A_j) \leq \frac{\min(A_j)}{\lambda}, \tag{3}$$

where $\lambda = 3$, we consider the relation frequencies to exhibit low variability and set $\theta_j = 0$. Otherwise, the threshold is defined as:

$$\theta_j = \frac{\min(A_j) + \max(A_j)}{2}. \tag{4}$$

Finally, we select all sentences that contain at least one such relation to construct the *filtered corpus*. This process effectively filters out noisy pseudo-labels while preserving frequently observed and contextually consistent event structures, resulting in a higher-quality silver-standard dataset. After cross-document filtering phase, we obtain a total of 15,260 qualified sentences in the filtered corpus.

3.3 Schema-based Data Augmentation

While filtering improves data quality, it does not increase structural diversity. We therefore introduce schema-based augmentation to generate diverse yet structurally valid event instances. This

[1] https://baomoi.com
[2] https://news.google.com/home?hl=vi&gl=VN

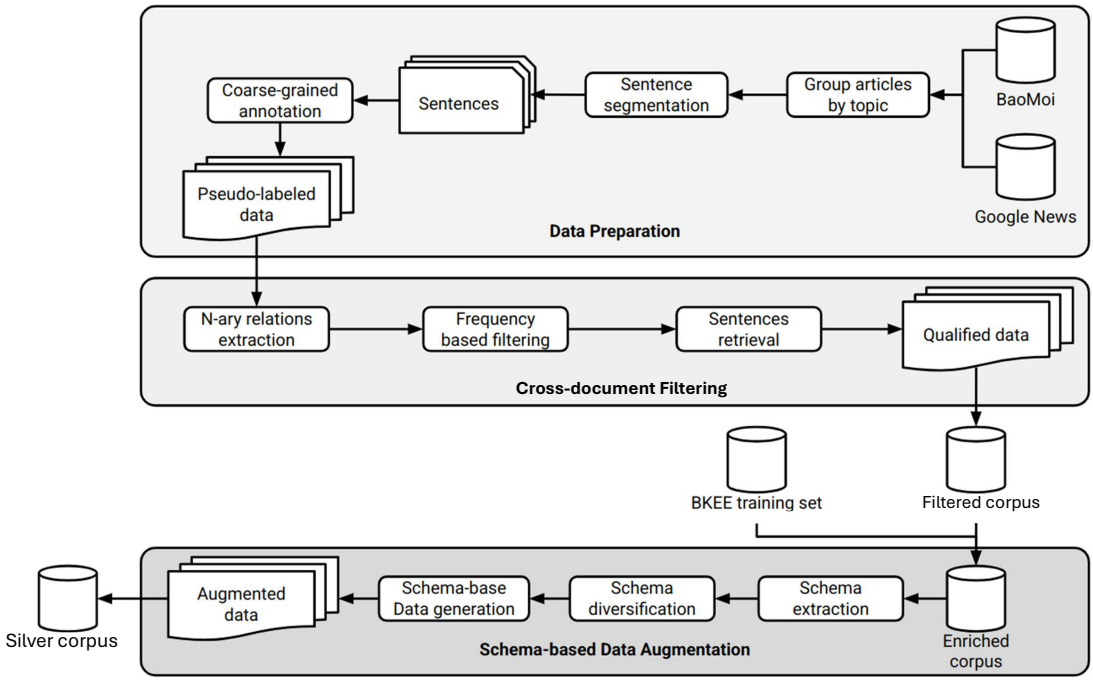


Figure 2: Silver corpus construction process.

phase adopts a sentence-level schema-based data augmentation strategy that introduces structural constraints during generation. Inspired by prior work (Jin and Ji, 2024), we adapt it to a setting where events are expressed within isolated sentences rather than across consecutive sentences.

Schema extraction: Since a sentence may contain multiple event triggers, we extract an event schema based on the triggers and their associated arguments. Figure 3 illustrates an example of such a schema, capturing the structural pattern of events and their contextual arguments, serving as an abstract template for data augmentation.

Schema diversification: To increase diversity, we construct new schemas from existing ones. For sentences containing multiple events, we decompose them into substructures and recombine them to form new event configurations. For sentences with a single event, we extend the schema by introducing additional arguments (e.g., *Time*, *Place*) when they are absent, ensuring that the resulting structures remain contextually valid. We represent schemas as structured graphs, where nodes correspond to event triggers and argument roles. To instantiate these schemas, we build a mapping M from event and entity types to candidate surface forms, collected from both internal (labeled data) and external sources. We sample from this pool to assign concrete values to each node, producing diverse realizations of the same structural pattern.

Schema-based data generation: The instantiated schema is serialized into a structured format (e.g., JSON) and used as input to an LLM (GPT-4o) for sentence generation. For schemas containing multiple events, the order of event instances is randomized to increase diversity. The generated sentences are then automatically annotated by aligning them with the schema, where matched spans are assigned their corresponding event and argument labels, ensuring structural consistency. This process ensures that the generated data remains structurally valid while introducing diverse surface realizations. The final silver corpus is expanded to 46,240 sentences.

4 Proposed Weakly Supervised Event Extraction Model

We build our framework upon the FourIE architecture (Nguyen et al., 2021), which jointly performs event mention detection (EMD), event detection (ED), and event argument extraction (EAE) over an input sentence $\mathbf{w} = [w_1, \ldots, w_n]$. We adopt this model as a backbone without modifying its core architecture, and focus on improving performance through enhanced training data constructed via weak supervision. The architecture comprises three phases: Span Detection, Instance Interaction, and Type-aware Regularization, as illustrated in Figure 4.

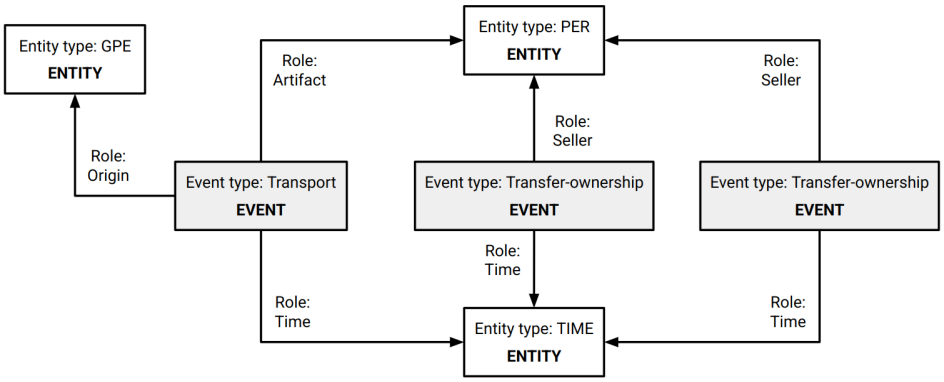


Figure 3: An example of an event schema extracted from a sentence. The schema represents event triggers and their associated arguments as a structured graph.

4.1 Span Detection

This module identifies entity mentions and event triggers from an input sentence to form nodes for the subsequent interaction graph. We formulate this step as a sequence labeling task using the BIO scheme, where each token is assigned one of three tags: B (Begin), I (Inside), or O (Outside). Unlike conventional named entity recognition, this stage does not predict specific entity or event types.

Given an input sentence $\mathbf{w}$, a pretrained encoder (i.e, PhoBert (Nguyen and Nguyen, 2020) or XLM-RoBERTa (Conneau et al., 2019)) produces contextual representations $X = [x_1, \ldots, x_n]$. These representations are then fed into two Conditional Random Fields layers to decode the optimal BIO tag sequences for entity mentions and event triggers, respectively. The model is trained by minimizing the negative log-likelihood losses $L_{\text{span}}^{\text{entity}}$ and $L_{\text{span}}^{\text{trigger}}$.

4.2 Instance Interaction

Given two span sets (entities and event triggers), the Instance Interaction module captures and enhances interactions between instances across tasks to improve prediction accuracy.

First, a representation vector for each span (i, j) $(1 \leq i \leq j \leq n)$ in these two sets is computed using the representation vectors $x_i, \ldots, x_j$. Let $R^{\text{entity}} = \{e_1, e_2, ..., e_{n_{\text{entity}}}\}$ $(n_{\text{entity}} = |R^{\text{entity}}|)$ and $R^{\text{trigger}} = \{t_1, t_2, ..., t_{n_{\text{trigger}}}\}$ $(n_{\text{trigger}} = |R^{\text{trigger}}|)$ denote the sets of span representation vectors for entities and event triggers in $\mathbf{w}$, respectively. The use of these two sets will be described in the following sub-sections.

Once R^{entity} and R^{trigger} are formed, we construct instance representations for the three IE tasks (EMD, ED, and EAE). Entity and trigger instances are directly derived from R^{entity} and R^{trigger}, respectively.

For argument prediction, which involves both a trigger and an entity, argument instances are defined as:

$$R^{\text{argument}} = \{\, arg_{ij} = [t_i, e_j] \mid t_i \in R^{\text{trigger}}, \; e_j \in R^{\text{entity}} \,\}. \tag{5}$$

The initial representation vectors for argument instances are constructed accordingly.

To model interactions between related instances, a graph G^{inst} is constructed, consisting of nodes N^{inst} and edges E^{inst}. The node set is defined as $N^{\text{inst}} = R^{\text{entity}} \cup R^{\text{trigger}} \cup R^{\text{argument}}$. Each entity node e_i is connected to all argument nodes $arg_{ij} = [t_j, e_i]$, and each trigger node t_j is also connected to these argument nodes, enabling information sharing among related instances.

This graph is then processed by a Graph Convolutional Network (GCN) to enrich instance representations. Let the initial node representations be $\{r_1, r_2, ..., r_{n_i}\}$ and the adjacency matrix be A^{inst}, where $A_{ij}^{\text{inst}} = 1$ indicates a connection between nodes r_i and r_j. The enriched representations are computed as:

$$r_1^{\text{inst}}, r_2^{\text{inst}}, ..., r_{n_i}^{\text{inst}} = \text{GCN}(A^{\text{inst}}; r_1, r_2, ..., r_{n_i}; N_i). \tag{6}$$

Finally, the enriched vectors are used to perform EMD, ED, and EAE. Let $\mathcal{T} = \mathcal{T}^{\text{entity}} \cup \mathcal{T}^{\text{trigger}} \cup \mathcal{T}^{\text{argument}}$, where $\mathcal{T}^{\text{entity}}$ denotes the set of entity types, and similarly for triggers and arguments. Let $t_k \in \{\text{entity}, \text{trigger}, \text{argument}\}$ be the task index and y_k the ground-truth label. Each type is associated with an embedding vector v, forming the set $\mathcal{V}$, with $\mathcal{V}^{t_k}$ corresponding to task t_k.

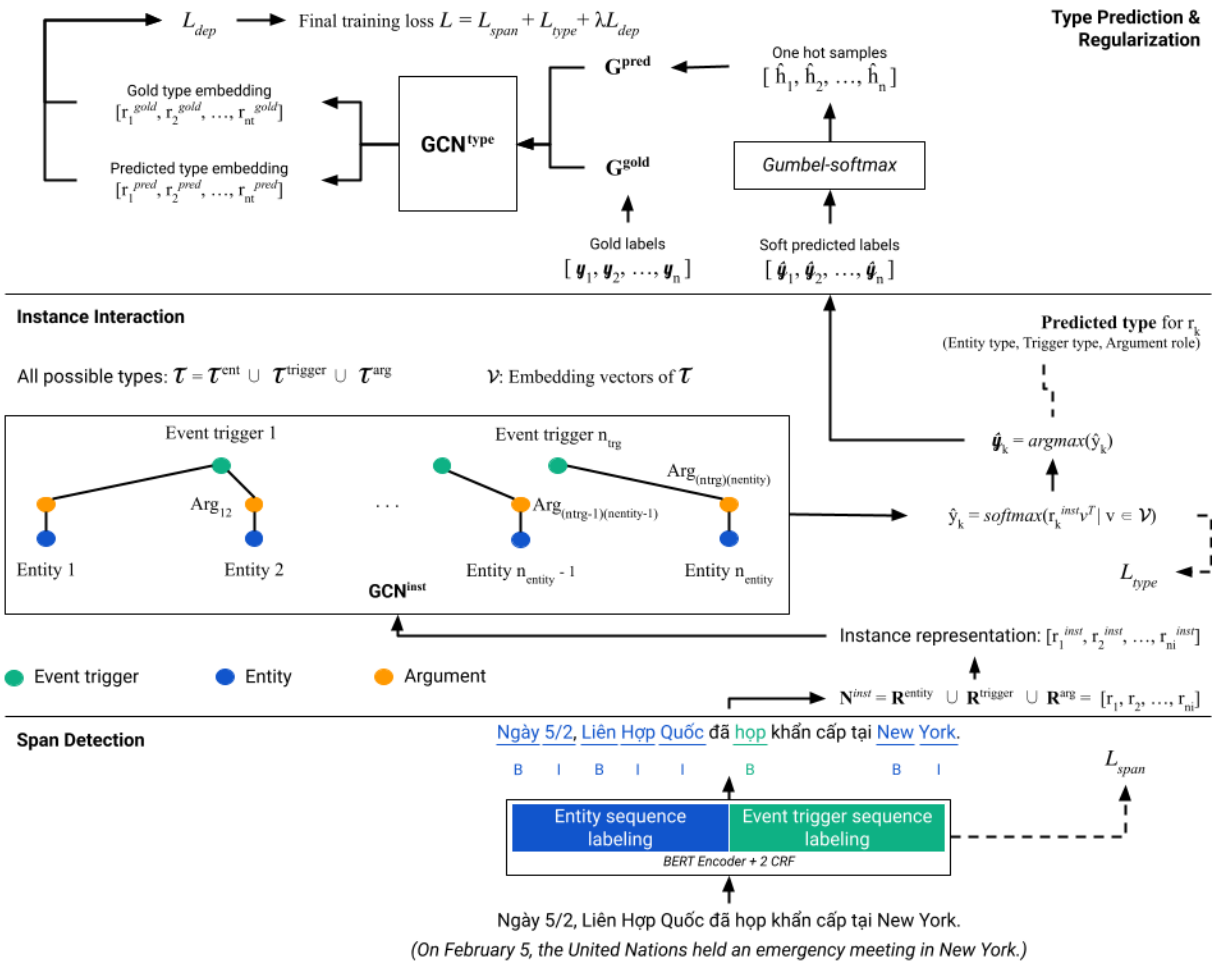


Figure 4: Overall architecture of the event extraction model.

For each instance $r_k \in N^{\text{inst}}$, the probability distribution over types is computed as:

$$\hat{y}_k = \text{softmax}(r_k^{\text{inst}} v^T \mid v \in \mathcal{V}^{t_k}), \tag{7}$$

and the predicted label is:

$$\hat{y}_k = \text{argmax}(\hat{y}_k). \tag{8}$$

4.3 Type Prediction and Regularization

This component models global type dependencies across the three IE tasks (EMD, ED, and EAE) to refine instance representations and improve prediction consistency.

Two dependency graphs, G^{gold} and G^{pred}, are constructed based on the gold types $y = \{y_1, y_2, ..., y_{n_i}\}$ and predicted types $\hat{y} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_{n_i}\}$. The nodes correspond to types in $\mathcal{T}$, while edges encode relations between types (e.g., entity–argument compatibility). Their adjacency matrices are denoted as A^{gold} and A^{pred}, respectively.

Type representations are computed via a GCN over the initial type embeddings $\mathcal{T} = [v_1, ..., v_{n_t}]$:

$$r_1^{\text{gold}}, r_2^{\text{gold}}, ..., r_{n_t}^{\text{gold}} = \text{GCN}(A^{\text{gold}}; v_1, ..., v_{n_t}; N_t), \tag{9}$$

and similarly for the predicted representations. The dependency loss is defined as:

$$L_{\text{dep}} = \sum_{i=1}^{n_t} \| r_i^{\text{gold}} - r_i^{\text{pred}} \|_2^2. \tag{10}$$

Since G^{pred} is derived from discrete predictions, direct backpropagation is not feasible. To address this, A^{pred} is approximated by a differentiable matrix $\hat{A}^{\text{pred}}$:

$$\hat{A}^{\text{pred}} = \sum_{(i,j) \in I^{\text{inst}}} \exp\left(-\beta \left(B - \text{int}_t - j\right)^2\right), \tag{11}$$

where $I^{\text{inst}} = \{(i,j) \mid A_{ij}^{\text{pred}} = 1\}$, $B = \{b_{ij}\}_{i,j=1,...,n_t}$, and β is a large constant. In addition, we employ the Gumbel-Softmax trick to

approximate categorical predictions with continuous relaxations for gradient-based optimization.

The final training objective is:

$$L = L_{\text{span}}^{\text{entity}} + L_{\text{span}}^{\text{trigger}} + L_{\text{type}} + \lambda L_{\text{dep}}. \quad (12)$$

where λ balances the regularization term.

5 Experimental Results and Discussion

5.1 Experimental Settings

Dataset. We conduct experiments on the BKEE dataset, a benchmark for Vietnamese event extraction collected from 11 news domains. The dataset covers 12 entity types, 8 event types, 33 event subtypes, and 28 argument roles, with nearly 9,000 annotated event mentions and over 16,000 annotated entity mentions and arguments (Nguyen et al., 2024). Following the standard data split, it contains 10,959 training instances, 4,301 development instances, and 3,736 test instances.

Metrics. We report F1 scores for the three tasks, including entity mention detection (EMD), event detection (ED), and event argument extraction (EAE).

Environments and Configurations. We implement our model using Python 3.10.16, PyTorch 2.0.1+cu117, and Transformers 4.47.1. Experiments are conducted on an NVIDIA GeForce RTX 2080 Ti GPU with 12GB VRAM. We use the pretrained FacebookAI/xlm-roberta-large model (Conneau et al., 2019) and PhoBert (Nguyen and Nguyen, 2020) for encoding and VnCoreNLP (Vu et al., 2018) for word tokenization.

The model is configured with a dropout rate of 0.4, sigmoid activation, and $N_i = 2$ hidden layers. We set $\beta = 1000$, $\lambda = 0.5$, and the learning rate to $5\text{e}{-6}$. The maximum number of training epochs is 25, and the batch size is 1 due to resource constraints.

5.2 Model Performance and Comparisons

We compare our method with three baseline models: (i) Pipeline models that perform each task independently, (ii) FourIE (Nguyen et al., 2021) as a joint learning baseline originally designed for English event extraction, and (iii) OneIE (Nguyen et al., 2024), an extension of joint learning approaches adapted for Vietnamese. We also evaluate two embedding settings, including multilingual XLM-RoBERTa (Conneau et al., 2019) and Vietnamese-specific PhoBERT (Nguyen and Nguyen, 2020).

Model	EMD	ED	EAE
(1) Pipeline + XLM-RoBERTa	55.0	60.3	44.9
(2) FourIE + XLM-RoBERTa	56.4	61.5	51.6
(3) OneIE + XLM-RoBERTa	56.3	60.0	51.7
(4) Pipeline + PhoBERT	54.4	61.8	44.4
(5) FourIE + PhoBERT	57.6	61.9	53.4
(6) OneIE + PhoBERT	55.8	**62.8**	53.0
(7) Proposed model	**59.1**	62.5	**55.2**

Table 1: Model performance comparison. Results are reported in F1 (%). The best results are highlighted in bold.

Table 1 shows that joint learning approaches (FourIE and OneIE) consistently outperform pipeline models across all tasks, highlighting the importance of modeling interdependencies between EMD, ED, and EAE. In addition, PhoBERT-based models generally achieve better performance than XLM-RoBERTa, confirming the advantage of language-specific representations for Vietnamese.

Compared to these baselines, our proposed model achieves the best overall performance, obtaining the highest F1 scores on EMD (59.1) and EAE (55.2), while remaining competitive on ED (62.5). In particular, compared to OneIE with PhoBERT, our model improves EMD and EAE by $+3.3$ and $+2.2$ F1, respectively, with a slight decrease of 0.3 F1 on ED. These results suggest that the proposed weak supervision strategy is especially effective for entity and argument extraction, where richer contextual and structural signals are required, while offering limited gains for trigger detection, which is less dependent on additional data. Overall, this demonstrates the effectiveness of leveraging enhanced training data for improving joint event extraction performance.

5.3 Impact of Silver Corpus Construction and Training Strategies

We evaluate different strategies for constructing and utilizing Silver corpus:

- (0) *w/o Silver corpus*: the model is trained only on the golden BKEE training set.
- (1) *w/o augmentation*: silver corpus is used but without schema-based augmentation and LLM-based generation.
- (2) *w/o filtering*: silver corpus is constructed without applying cross-document n-ary relation filtering.

Table 2: Impact of silver corpus Construction and Training Strategies. Results are reported in F1 (%). The best results are highlighted in bold.

Training setting	EMD	ED	EAE
(0) w/o silver corpus	55.5	**62.6**	52.9
(1) w/o augmentation	56.59	60.33	51.32
(2) w/o filtering	57.46	60.83	52.80
(3) Full corpus (scratch)	57.43	62.22	53.09
(4) Full + PM_1	56.88	61.24	52.50
(5) Full + PM_2	**59.11**	61.73	**55.17**

Full corpus: BKEE + silver corpus. PM_1: pretrained on BKEE. PM_2: pretrained on BKEE + 1/5 Silver corpus.

- (3) *Full corpus (scratch)*: the golden training data and the full Silver corpus are combined and used to train the model from scratch.
- (4) *Full + PM1*: the model is first pre-trained on the golden training set, and then further trained on the full Silver corpus.
- (5) *Full + PM2*: a small corpus is first constructed by combining the golden training set with 20% of the Silver corpus to pre-train the model, which is then further trained on the remaining silver corpus.

The results on Table 2 shows that incorporating silver corpus consistently improves performance on EMD and EAE compared to training only on the golden data (row (0)), demonstrating the effectiveness of weakly supervised data for span and argument extraction. However, ED does not benefit as much, indicating that trigger detection is less sensitive to additional data. Both schema-based augmentation and filtering play important roles in improving data quality. Removing augmentation (row (1)) or filtering (row (2)) leads to noticeable drops in performance compared to the best setting (row (5)), showing that both diversity and noise reduction are crucial for constructing effective silver corpus. Finally, training strategy has a significant impact on performance. Simply training on the combined data from scratch (row (3)) does not yield the best results, and pretraining only on the golden data (row (4)) is also suboptimal. The best performance is achieved by progressively leveraging silver corpus (row (5)), where the model is first exposed to a smaller, mixed corpus before being trained on the full dataset. This suggests that a curriculum-style training strategy is more effective than directly mixing all data or relying solely on golden pretraining.

5.4 Error Analysis

To improve the proposed model and incentivize future research, the model output has been analyzed to find out errors that need to be taken into account. For errors examples, please refer to Appendix B.

Span errors. In EMD and ED, the model sometimes predicts incomplete or over-extended spans, especially for long or nested mentions. For example, *"Ngân hàng Nông nghiệp và Phát triển nông thôn Việt Nam Agribank"* may be partially detected as *"Agribank"*.

Isolated entity detection failure. Entities without explicit trigger associations are occasionally missed, suggesting that the model relies heavily on trigger-aware context for entity recognition.

Polysemy. Words with multiple meanings may lead to incorrect entity type predictions, e.g., *"Jordan"* being classified as PER instead of GPE.

Ambiguous context in EAE. In complex sentences with overlapping contextual cues, the model may assign incorrect semantic roles, indicating limitations in contextual reasoning.

Overall, these errors suggest that the main limitations lie in boundary detection and contextual reasoning. Future work may benefit from stronger structural modeling and more fine-grained semantic supervision.

6 Conclusion

In this paper, we address the data scarcity problem in Vietnamese event extraction by proposing a weakly supervised framework for constructing a qualified silver corpus. Our approach combines pseudo-labeling with a cross-document n-ary relation filtering strategy to improve annotation quality, and a schema-based data augmentation method to enhance data diversity. Built upon a strong joint learning backbone, the proposed framework effectively leverages the constructed silver corpus of 46,240 sentences to improve event extraction performance. Experimental results on the BKEE benchmark demonstrate consistent improvements, achieving gains of +3.3% F1 on EMD and +2.2% F1 on EAE compared to strong baselines, while maintaining competitive results on ED. Overall, our findings suggest that a carefully constructed silver corpus, together with appropriate training strategies, can serve as an effective alternative to costly manual annotation for low-resource event extraction.

Limitations

Despite the effectiveness of our approach, several limitations remain.

First, the quality of the constructed silver corpus still depends on the initial pseudo-labeling model. Errors introduced in this stage may propagate through subsequent filtering and training, especially for rare or complex event types.

Second, the cross-document n-ary relation filtering strategy relies on the assumption that important events are reported multiple times across different documents. As a result, infrequent or emerging events may be underrepresented or filtered out, limiting coverage.

Third, the schema-based data augmentation process depends on predefined schema structures and LLM-based generation, which may introduce noise or generate less natural sentences in some cases.

Finally, our framework builds upon an existing backbone model without modifying its architecture. While this allows us to focus on data-centric improvements, it may limit the potential gains achievable through model-level innovations.

Addressing these limitations, particularly improving pseudo-label quality and better handling rare events, remains an important direction for future work.

References

Jun Araki and Teruko Mitamura. 2018. Open-domain event detection using distant supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 878–891. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867. International Committee on Computational Linguistics.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077. Association for Computational Linguistics.

James Ferguson, Colin Lockard, Daniel Weld, and Hannaneh Hajishirzi. 2018. Semi-supervised event extraction with paraphrase clusters. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 359–364. Association for Computational Linguistics.

Lifu Huang and Heng Ji. 2020. Semi-supervised new event type induction and event detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 718–724. Association for Computational Linguistics.

Xiaomeng Jin and Heng Ji. 2024. Schema-based data augmentation for event extraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14382–14392. ELRA and ICCL.

Daniel Jurafsky and James H. Martin. 2026. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd edition. Online manuscript released January 6, 2026.

April Kontostathis, Leon M. Galitsky, William M. Pottenger, Soma Roy, and Daniel J. Phelps. 2004. *A Survey of Emerging Trend Detection in Textual Data Mining*, pages 185–224. Springer New York.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 894–908.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009. Association for Computational Linguistics.

Kang Liu, Yubo Chen, Jian Liu, Xinyu Zuo, and Jun Zhao. 2020. Extracting events and their relations from texts: A survey on recent research progress and challenges. *AI Open*, 1:22–39.

Teruko Mitamura, Zhengzhong Liu, and Eduard H Hovy. 2017. Events detection, coreference and sequencing: What's next? overview of the tac kbp 2017 event track. In *TAC*.

Dat Quoc Nguyen and Anh-Tuan Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 1037–1042.

Minh Van Nguyen, Viet Dac Lai, and Thien Huu Nguyen. 2021. Cross-task instance representation interactions and label dependencies for joint information extraction with graph convolutional networks.

In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 27–38. Association for Computational Linguistics.

Thi-Nhung Nguyen, Bang Tien Tran, Trong-Nghia Luu, Thien Huu Nguyen, and Kiem-Hieu Nguyen. 2024. BKEE: Pioneering event extraction in the Vietnamese language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2421–2427. ELRA and ICCL.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese natural language processing toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 56–60. Association for Computational Linguistics.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789. Association for Computational Linguistics.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. (URL: https://catalog. ldc. upenn. edu/LDC2006T06).

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. Maven: A massive general domain event detection dataset. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1652–1671.

Wei Xiang and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2020. Unsupervised data augmentation for consistency training.

QunLi Xie, JunLan Pan, Tao Liu, BeiBei Qian, XianChuan Wang, and Xianchao Wang. 2021. A survey of event relation extraction. In *International Conference on Frontier Computing*, pages 1818–1827. Springer.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. 2019. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5284–5294. Association for Computational Linguistics.

Wenlin Yao, Zeyu Dai, Maitreyi Ramaswamy, Bonan Min, and Ruihong Huang. 2020. Weakly supervised subevent knowledge acquisition. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)*, pages 5345–5356.

#	Vietnamese	English
1	Ngày 12/02/2025, Bộ Tư pháp đã tổ chức **họp thẩm định** đề nghị xây dựng Luật thuế TNCN.	(On February 12, 2025, the Ministry of Justice held an **appraisal meeting** on the proposal to draft the Personal Income Tax Law.)
2	Trên cơ sở đó, ngày 12/2 Bộ Tư pháp đã tổ chức **họp thẩm định** đề nghị xây dựng Luật thuế thu nhập cá nhân.	(On that basis, on February 12, the Ministry of Justice held an **appraisal meeting** on the proposal to draft the Personal Income Tax Law.)
3	Ngày 12/2/2025, Bộ Tư pháp đã tổ chức **họp thẩm định** với mục đích đề nghị xây dựng Luật Thuế thu nhập cá nhân.	(On February 12, 2025, the Ministry of Justice held an **appraisal meeting** with the aim of proposing the drafting of the Personal Income Tax Law.)

Figure 5: Sentences from different documents describing the same event.

A Cross-document Filtering Examples

Figure 5 illustrates an example of sentences from different articles within the same topic may describe the same event with consistent triggers and arguments.

Figure 6 presents examples of n-ary relations extracted from pseudo-labeled sentences within a topic group. Each relation is characterized by an event type and its associated arguments, along with its occurrence count across sentences.

B Model Error Examples

Some representative examples of model errors are provided in Table 3.

Span errors. This type of error occurs in both EMD and ED when the predicted span does not exactly match the gold annotation. The model may either partially detect a mention or over-extend the span beyond its correct boundary. For instance, the entity *"Ngân hàng Nông nghiệp và Phát triển nông thôn Việt Nam Agribank"* is only partially detected as "*Agribank*", while complex mentions such as *"Bí thư tỉnh ủy và Chủ tịch Hội đồng nhân dân tỉnh Bến Tre Hồ Thị Hoàng Yến"* should be split into multiple entities but are instead merged into a single span. These errors typically arise when entity mentions or event triggers have long and nested structures, making it difficult for the model to accurately determine span boundaries.

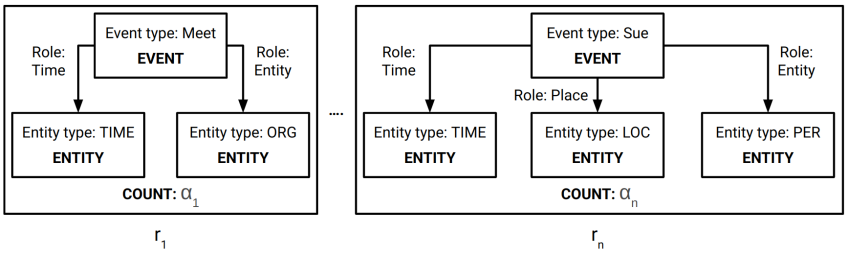


Figure 6: Examples of n-ary relations extracted from sentences within the same topic group. Each relation consists of an event type and its associated arguments with occurrence counts.

Isolated entity detection failure. This error occurs when entities that are not explicitly associated with an event trigger are not detected. For example, in sentence (3), entities such as *"ông Hùng"* and *"năm 2023"* are missed because they do not directly participate in a clearly expressed event. This suggests that the model relies heavily on trigger-aware context to identify relevant entities, and struggles to recognize standalone entities when contextual cues are limited.

Polysemy. Ambiguous words with multiple semantic meanings can lead to incorrect entity type predictions. For example, in sentence (4), the word *"Jordan"* is incorrectly classified as a person (PER) instead of a geo-political entity (GPE). Such errors occur when the model fails to effectively leverage contextual signals to disambiguate between different meanings of the same surface form. Although relatively less frequent, these errors highlight limitations in semantic understanding.

Ambiguous context in EAE. In the EAE task, errors often arise from complex or ambiguous contexts where multiple entities and actions are present within the same sentence. In sentence (5), the entity "*An*" is incorrectly assigned the Agent role in a Transport event, even though it does not perform the action of being transported. This type of error indicates that the model may misinterpret semantic roles when contextual signals are dense or overlapping, leading to incorrect argument-role assignments.

Table 3: Some notable errors on the test set output of the model.

#	Vietnamese	English	Error
1	*Lãnh đạo Ngân hàng Nông nghiệp và Phát triển nông thôn (**Agribank**) Chi nhánh Bắc Yên Bái trao tiền hỗ trợ gia đình chị Nông Thị Đông ở thị trấn Yên Bình.*	*Leaders of the Agricultural Bank and Rural Development (**Agribank**) North Yen Bai Branch donated financial support to the family of Ms. Nong Thi Dong in Yen Binh Town.*	**Wrong entity detected**: Agribank
2	***Bí thư tỉnh ủy và Chủ tịch Hội đồng nhân dân tỉnh Bến Tre Hồ Thị Hoàng Yến** chủ trì phiên họp.*	***Secretary of the Provincial Party Committee and Chairwoman of the People's Council of Ben Tre Province, Ho Thi Hoang Yen** chaired the meeting.*	**Wrong entity detected**: full span incorrectly merged
3	*Trước đây, gia đình ông **Hùng** thuộc diện hộ nghèo, đến năm 2023 là hộ cận nghèo.*	*In the past, Mr. **Hung**'s family was classified as a poor household, but by 2023, they became a near-poor household.*	**Missing entity**: Hung
4	***Jordan** bày tỏ lo ngại về những căng thẳng gia tăng trong khu vực.*	***Jordan** expressed concerns over the escalating tensions in the region.*	**Wrong entity type**: Jordan (PER → LOC)
5	*Hải đến nhà **An**, và **An** dùng điện thoại để giữ liên lạc với Hải trong quá trình di chuyển.*	*Hai went to **An**'s house, and **An** used his phone to stay in touch with Hai during the journey.*	**Wrong argument**: An incorrectly labeled as *Agent* in *Transport* event

When Tasks Share Structure: A Comparative Study of Training Strategies for Generative Event Extraction

Rishi Ravikumar and Riza Batista-Navarro
Department of Computer Science, University of Manchester, United Kingdom
{rishi.ravikumar, riza.batista}@manchester.ac.uk

Abstract

Event extraction requires performing two interdependent subtasks: event detection and event argument extraction. While prior work has explored pipelined and joint training approaches, the question of how best to coordinate training across these subtasks in generative LLM-based systems remains open. We present a systematic study comparing three training paradigms: disjoint, fully shared and hybrid weight allocation, instantiated as eight concrete strategies and evaluated on ACE2005 and RichERE across multiple instruction-tuned LLMs. Our findings show that training strategy has a consistent and meaningful effect on extraction accuracy, and that a clear best-performing strategy emerges across models and benchmarks. We believe that these findings could extend beyond event extraction to other information extraction tasks that decompose into interdependent subtasks.

1 Introduction

Event extraction (EE) aims to identify structured event information from text. It is typically decomposed into two subtasks: event detection (ED), which identifies event triggers and their types, and event argument extraction (EAE), which identifies event participants and their roles (LDC, 2005). A long-standing question in EE is how these interdependent subtasks should be coordinated during training: should they be learned independently in a pipeline, jointly or through some intermediate form of interaction?

Both pipeline and joint formulations have been extensively explored, each with well-known trade-offs. Pipeline approaches offer modularity but suffer from error propagation, while joint approaches can capture cross-task dependencies at the cost of increased complexity. The advent of large language models (LLMs) has introduced a new paradigm, where EE is framed as a conditional text generation problem. This shift brings the question of training strategy into a new setting, one that has received little systematic attention: what is the best strategy for fine-tuning LLMs for generative event extraction?

In this work, we systematically investigate how different training strategies affect extraction performance in generative event extraction. We study three paradigms: disjoint parameter allocation, fully shared parameters and hybrid configurations with partial parameter sharing, instantiated as eight **computationally equivalent** training strategies. Models are fine-tuned using LoRA (Low-Rank Adapters) (Hu et al., 2021) and evaluated on ACE2005 (Doddington et al., 2004) and RichERE (Song et al., 2015) across three instruction-tuned LLMs ranging from 3B to 12B parameters.

Our results show that training strategy has a consistent and meaningful effect on extraction accuracy across models and benchmarks. We find that a disjoint approach in which the ED adapters are initialised from pre-trained EAE adapters consistently outperforms both de facto approaches: fully independent training and joint modelling. Joint modelling, where both tasks are handled within a single pass, on the contrary, proves to be the weakest configuration overall. We also find that robustness to strategy choice increases with model size. Our contributions are as follows:

- We propose a taxonomy of training strategies for generative event extraction based on parameter sharing and task interaction, spanning disjoint, fully shared and hybrid paradigms.
- We conduct a controlled empirical study of eight strategies across three LLMs and two benchmarks, providing a comprehensive comparison of training strategies for generative event extraction[1].

[1]Our code is available at www.github.com/rishi-ravikumar/GenerativeEventExtractionTrainingStrategies.

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 38–48

- We provide insights into how task decomposition, transfer and parameter sharing affect extraction accuracy across models and datasets.

2 Related Work

2.1 Event Extraction: Pipeline and Joint Approaches

Pipeline and joint approaches represent the two dominant paradigms for event extraction, each with well-established trade-offs. Pipeline approaches train separate models for ED and EAE, applying them sequentially at inference time, offering modularity but suffering from error propagation: mistakes in event detection directly degrade argument extraction (Xiang and Wang, 2019). Joint approaches model both subtasks within a unified framework, enabling cross-task interaction and interdependencies to be exploited, thereby alleviating error propagation at the cost of increased model complexity (Lin et al., 2020; Xiang and Wang, 2019). The relative merits of each paradigm remain setting-dependent, with leading systems spanning both: joint models such as OneIE (Lin et al., 2020) and DyGIE++ (Wadden et al., 2019) alongside pipeline approaches such as TagPrime (Hsu et al., 2023), as evidenced by the standardised evaluation in Huang et al. (2024).

2.2 Generative Event Extraction

The reframing of event extraction as a conditional text generation problem has gained significant traction with the advent of pre-trained language models. Early generative approaches include Du and Cardie (2020), who formulate EAE as question answering, and Paolini et al. (2021), who cast a range of structured prediction tasks, including EE, as translation between augmented natural languages. Lu et al. (2021) propose Text2Event, which directly generates structured event records from text using constrained decoding, handling ED and EAE jointly in a single pass. Hsu et al. (2022) propose DEGREE, which frames EE as prompt-based conditional generation, and they explicitly evaluate both pipeline (DEGREE-PIPE) and joint (DEGREE-E2E) configurations, finding that the joint configuration generally outperforms the pipeline configuration on ACE2005, particularly in low-resource scenarios. More recently, work on large instruction-tuned LLMs has examined the upper bound of the generative paradigm: Gao et al. (2023) evaluate an earlier version of ChatGPT on few-shot EE and find that it considerably falls short of supervised approaches, while Srivastava et al. (2025) systematically study instruction tuning strategies for event extraction. Across this body of work, the question of how training should be coordinated across ED and EAE in generative models has received no systematic attention, with most work committing to a single fixed configuration. This paper addresses this gap directly.

2.3 Multi-Task and Transfer Learning for Information Extraction

The question of how to coordinate learning across related tasks has a long history in NLP. Caruana (1997) establishes that jointly training related tasks provides inductive biases that improve generalisation, and subsequent work has explored both hard parameter sharing, where all tasks share the same parameters, and soft sharing, where tasks maintain separate parameters with regularisation encouraging similarity (Ruder, 2017). In the context of IE, multi-task learning over related subtasks, where shared encoders capture common linguistic features, has consistently outperformed single-task baselines (Wadden et al., 2019; Lin et al., 2020). Paolini et al. (2021) extend this to a broader range of structured prediction tasks within a unified generative framework, demonstrating that related IE tasks can benefit from shared representations at scale. Sequential transfer between related tasks has also proven effective, with pre-training on auxiliary tasks providing beneficial initialisations for downstream extraction objectives (Pruksachatkun et al., 2020). Our work draws directly on these insights, systematically investigating how hard parameter sharing, task-specific parameters and sequential transfer interact in the specific context of generative EE under LoRA-based fine-tuning.

3 Methodology

3.1 Task Formulation

We adopt a generative formulation of event extraction, casting ED, EAE and joint EE as conditional text generation tasks within a prompt-completion framework. In each case, the model takes a natural language sentence as input and generates a structured JSON representation of the predicted event information. We deliberately adopt a minimal prompting strategy, providing no task instructions or additional context beyond the input sentence. Since all models are fine-tuned, the training

process itself is sufficient for the model to learn the target behaviour. This choice avoids confounding effects introduced by prompt engineering and ensures consistency across all experimental conditions.

Event Detection. Given a raw input sentence, the model generates a list of all event triggers present in the sentence along with their corresponding event types, in a single generation step.

Event Argument Extraction. Given the input sentence along with a specific event trigger and its type, provided using simple delimiters, the model generates all arguments associated with that event and their semantic roles. Since argument extraction is conditioned on a single trigger, EAE is run once per trigger. This formulation reflects the dependence of EAE on ED outputs.

Joint Event Extraction. Given a raw input sentence, the model generates complete event structures in a single pass, including triggers, event types and all associated arguments.

The prompt-completion formats for all three formulations are illustrated in Table 1, with all completions structured as JSON arrays with consistent field names across tasks.

3.2 Training Strategies

We investigate eight training strategies organised into three paradigms based on how adapters are allocated across ED and EAE. Unless otherwise stated, inference follows a pipeline: ED is run first to identify triggers and event types, which are then passed as input to EAE. During training, EAE is conditioned on gold triggers and event types, whereas at inference it relies on predictions from ED.

3.2.1 Disjoint Training

Disjoint strategies assign separate sets of adapters to ED and EAE, with no parameter sharing between the two tasks at inference time.

Independent (S1). Separate adapter sets are trained for ED and EAE independently, with no interaction between tasks at any stage. This represents the standard pipeline baseline.

Forward Transfer (S2). The ED adapters are trained first and used to initialise the EAE adapters, which are then trained. This examines whether ED supervision provides a useful starting point for argument extraction.

Backward Transfer (S3). The EAE adapters are trained first and used to initialise the ED adapters, which are then trained. This examines whether exposure to argument-level supervision benefits ED.

3.2.2 Fully Shared Training

Fully shared strategies use a single set of adapters for both tasks, with no task-specific parameters.

Joint Modelling (S4). A single set of adapters is trained to generate complete event structures, including triggers, types and arguments, in a single pass. Unlike the other strategies, inference is also performed in a single pass. This represents the standard joint baseline.

Mixed Training (S5). A single set of adapters is trained on interleaved batches of ED and EAE instances. Despite sharing a single adapter set, inference remains pipelined: the model first performs ED, then EAE.

3.2.3 Hybrid Training

Hybrid strategies partition adapters into two sets: a shared set applied to the lower layers, and task-specific sets applied to the upper layers. Since lower layers of transformer-based LLMs encode general linguistic and semantic representations while upper layers encode increasingly task-specific information (Rogers et al., 2021), this design allows the model to leverage common event semantics in the shared layers while retaining task-specific capacity in the upper layers.

Partial Sharing (S6). Adapters in the lower layers are shared between ED and EAE, while the upper layers maintain separate task-specific adapter sets. This design interpolates between fully disjoint and fully shared training: at 0% sharing the strategy reduces to Independent training (S1), and at 100% sharing it reduces to Mixed Training (S5). We experiment with three configurations: sharing the lower 25% of layers (**S6.1**), 50% (**S6.2**) and 75% (**S6.3**). An overview of this architecture is shown in Figure 1.

3.3 Benchmark Datasets

We evaluate on ACE2005 and RichERE, the two most widely used benchmarks for event extraction, both focusing on the news domain. We follow the TextEE standardised benchmarking framework (Huang et al., 2024) to download, pre-process and split the data, adopting the standard train/dev/test splits. In particular, we use Split 1 of the five

Task	Example
ED	**Prompt:** A bomb exploded in the city center. **Completion:** `[{'trigger': 'exploded','event_type': 'Conflict:Attack'}]`
EAE	**Prompt:** A bomb exploded in the city center. <trigger> exploded </trigger> <type> Conflict:Attack </type> **Completion:** `[{'argument': 'city center', 'argument_role': 'Place'}]}]`
EE (ED+EAE)	**Prompt:** A bomb exploded in the city center. **Completion:** `[{'trigger': 'exploded','event_type': 'Conflict:Attack', 'arguments': [{'argument': 'city center', 'argument_role': 'Place'}]}]`

Table 1: Prompt-completion formats for different task formulations.

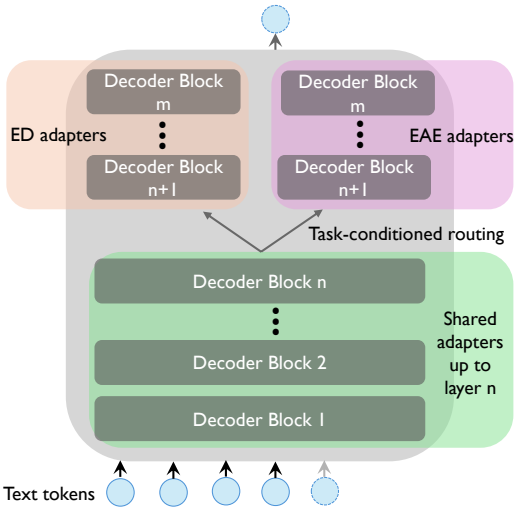


Figure 1: Partial Sharing (S6): lower layers use a shared adapter set across ED and EAE, while upper layers maintain separate task-specific adapter sets.

Dataset	# Event Types	# Argument Types	Train	Dev	Test
ACE05	33	22	3234	348	416
RichERE	38	21	2721	291	403

Table 2: Dataset statistics for ACE2005 and RichERE. Samples containing no events are excluded.

splits provided by TextEE. Additionally, *we exclude samples containing no events* due to the class imbalance in both datasets: fewer than 20% of ACE2005 samples and fewer than 30% of RichERE samples contain at least one event. Retaining null-event samples would therefore heavily skew evaluation towards trivial empty predictions, obscuring meaningful differences in extraction quality across strategies. We note that this filtering precludes direct comparisons with prior work that retain such samples; however, since all strategies are evaluated under identical conditions, cross-strategy comparisons remain fully valid. The data is also pre-processed into the prompt-completion format described in Section 3.1.

ACE2005 is a popular benchmark for event extraction, featuring a well-established ontology of event types and argument roles. RichERE extends this setting with a broader ontology, potentially presenting additional challenges due to its increased schema complexity and smaller size. Together, the two benchmarks allow us to assess the robustness of our findings across different levels of schema complexity and data availability. Key statistics are reported in Table 2.

3.4 Models

We experiment with three open-weight, instruction-tuned decoder-only LLMs: Qwen2.5-3B-Instruct (Qwen et al., 2025), Llama-3.1-8B-Instruct (Meta et al., 2024) and Mistral-Nemo-Instruct-2407 (Mistral AI, 2024), spanning 3B to 12B parameters. This selection provides diversity across both model family and scale, allowing us to assess the consistency of our findings across these dimensions.

3.5 Hardware and Hyperparameters

Hyperparameter	Value
Sampling Parameters	Greedy Sampling
Number of Epochs	2
Precision	bfloat16
Optimizer	AdamW
Learning Rate	2.00E-04
Learning Rate Scheduler	cosine with warm-up
Warm-up Ratio	0.05
Batch Size	4
LoRA Dropout	0.05
Gradient Accumulation Steps	4
Weight Decay	0.01
Gradient Clipping	1.0
Maximum Seq. Length	512
Projections Modified	'q_proj', 'k_proj', 'v_proj', 'o_proj','gate_proj', 'up_proj', 'down_proj'

Table 3: Hyperparameters used for training.

Table 3 lists the key hyperparameters used for fine-tuning. We fine-tune all models using LoRA on NVIDIA RTX A5000, A40 and A100 GPUs, with a cumulative training time of approximately 200 GPU hours. Loss is computed exclusively over completion tokens, with input tokens masked during training. We use a fixed training schedule of 2 epochs across all experiments to ensure comparable

compute across strategies. This choice is further supported by prior QLoRA results (Dettmers et al., 2023), which report stronger performance with 2-epoch fine-tuning compared to 1 and 3 epochs. Additionally, the small dataset sizes and the number of model components adapted by LoRA make 2 epochs a reasonable choice.

3.6 Computational Equivalence

All eight strategies are designed to incur identical training cost, measured as the total number of completion tokens processed under prompt loss masking. Let D_{ED} and D_{EAE} denote the ED and EAE training sets with mean completion lengths m and k respectively. In the disjoint and transfer strategies (S1-S3), two adapter sets are trained independently, one per task, for two epochs each, contributing $2(|D_{\text{ED}}| \cdot m + |D_{\text{EAE}}| \cdot k)$ loss tokens in total. In Joint Modelling (S4), ED and EAE completions are concatenated into a single sequence per sample, and a single adapter set is trained for two epochs, yielding the same total cost. In Mixed Training (S5), the two datasets are interleaved and processed by a single adapter set over two epochs, again resulting in the same total. In the hybrid strategies (S6.1-S6.3), the shared lower-layer adapters are trained on the interleaved combined dataset, while each task-specific upper-layer adapter set is trained on its respective task; the contributions sum to the same quantity. *Training cost is thus held constant across all strategies by construction.*

3.7 Evaluation

We adopt two evaluation metrics following standard practice in event extraction literature (Huang et al., 2024), all reported as micro-averaged F1. A true positive for ED requires both the trigger span and event type to exactly match the gold annotation; we report this as the **TC** (Trigger Classification') score. A true positive for *end-to-end EE* requires the trigger span, event type, argument span and argument role to all exactly match the gold annotation; we report this as the **AC** (Argument Classification') score. Since EAE models are trained on gold triggers but evaluated on predicted triggers, AC is subject to exposure bias. To address this, we additionally propose **AC|TC**, which evaluates argument extraction exclusively over correctly identified triggers; by construction, this ensures that EAE receives input consistent with its training conditions, eliminating the exposure bias gap. However, AC|TC should be interpreted with care: as it is computed over only the subset of instances where ED succeeds, it is sensitive to the composition of that subset and provides only a rough signal of EAE performance in isolation. Together, TC measures ED quality, AC captures end-to-end EE performance, and AC|TC offers a rough measure of EAE quality.

4 Results and Discussion

Table 4 presents TC, AC and AC|TC scores for all eight training strategies across three models and two benchmarks, with each configuration run over three random seeds and averaged. To facilitate comparison across strategies independently of absolute score differences between models, we rank strategies by score within each model and benchmark combination, then average these ranks across models to produce a consolidated ordering for each metric and benchmark; results are reported in Table 5. To assess the consistency of these rankings across models, we compute Kendall's coefficient of concordance (W) for each metric–benchmark combination. In 8 of 9 cases, W ranges from 0.503 to 0.677, indicating moderate to substantial agreement according to commonly used interpretation guidelines adapted from Landis and Koch (1977); The Comprehensive R Archive Network. The remaining case, AC|TC on ACE2005 (W=0.323), indicates fair agreement, although two of the three model pairs still exhibit substantial pairwise concordance. We organise our discussion around the key trends that emerge.

Performance improves consistently with model scale. Across all strategies and both benchmarks, TC and AC scores increase with model size. This trend holds regardless of training strategy, suggesting that the absolute gains from scaling generalise across different forms of task interaction.

Larger models are more robust to training strategy variation. As show in Figure 2, the TC and AC score gap (margin) between the best and worst performing training strategies generally narrows with model size across benchmarks. On ACE2005 TC, this margin is approximately 4.5 points for Qwen2.5 3B, compared to around 2.8 points for Mistral Nemo. A similar trend holds on RichERE and for AC. Training strategy remains a meaningful factor at all scales, but its effect is more pronounced in smaller models. This is likely attributable to two complementary factors: larger models bring

ACE05	Qwen2.5 3B Instruct			Llama 3.1 8B Instruct			Mistral Nemo Instruct 2407		
Strategies	TC	AC	AC|TC	TC	AC	AC|TC	TC	AC	AC|TC
S1: Independent	71.86 ± 0.41	46.15 ± 0.41	66.21 ± 0.34	75.52 ± 0.19	54.56 ± 0.17	72.93 ± 0.12	76.78 ± 0.38	54.83 ± 0.14	72.29 ± 0.23
S2: Forward Transfer	71.86 ± 0.41	47.23 ± 0.63	**67.71 ± 0.59**	75.52 ± 0.19	54.49 ± 0.35	**72.95 ± 0.50**	76.78 ± 0.38	55.55 ± 0.39	**73.15 ± 0.49**
S3: Backward Transfer	73.62 ± 0.27	47.28 ± 0.57	65.87 ± 0.40	**77.20 ± 0.43**	**55.91 ± 0.36**	72.25 ± 0.52	**78.15 ± 0.21**	**55.71 ± 0.25**	71.48 ± 0.19
S4: Joint Modelling	69.26 ± 0.24	43.74 ± 0.16	64.62 ± 0.08	75.07 ± 0.13	52.75 ± 0.61	71.02 ± 0.64	75.38 ± 0.39	54.59 ± 0.35	72.20 ± 0.59
S5: Mixed Training	72.67 ± 0.37	47.21 ± 0.80	66.07 ± 0.80	76.64 ± 0.41	53.99 ± 0.39	70.70 ± 0.24	76.89 ± 0.28	54.16 ± 1.07	70.97 ± 0.92
S6.1: Partial Sharing (25%)	**73.72 ± 0.31**	**47.65 ± 0.10**	65.95 ± 0.33	76.31 ± 0.28	54.83 ± 0.85	72.27 ± 0.47	76.67 ± 0.65	55.41 ± 0.94	72.07 ± 0.74
S6.2: Partial Sharing (50%)	73.58 ± 0.45	46.12 ± 0.55	64.45 ± 1.07	75.70 ± 0.18	53.70 ± 0.24	71.70 ± 0.03	77.34 ± 0.45	54.75 ± 0.86	70.99 ± 0.37
S6.3: Partial Sharing (75%)	72.16 ± 0.44	46.12 ± 1.08	65.42 ± 0.81	75.57 ± 0.69	53.52 ± 0.49	71.01 ± 0.10	76.95 ± 0.54	55.70 ± 0.60	72.39 ± 0.10

RichERE	Qwen2.5 3B Instruct			Llama 3.1 8B Instruct			Mistral Nemo Instruct 2407		
Strategies	TC	AC	AC|TC	TC	AC	AC|TC	TC	AC	AC|TC
S1: Independent	60.76 ± 0.87	41.81 ± 0.48	68.70 ± 0.21	64.97 ± 0.46	47.05 ± 0.82	72.37 ± 0.73	66.14 ± 0.63	49.48 ± 0.63	74.11 ± 0.40
S2: Forward Transfer	60.76 ± 0.87	41.70 ± 0.42	68.76 ± 0.36	64.97 ± 0.46	46.68 ± 0.79	71.83 ± 0.78	66.14 ± 0.63	**49.68 ± 0.67**	74.44 ± 0.65
S3: Backward Transfer	61.82 ± 0.15	**42.84 ± 0.08**	**69.13 ± 0.32**	**67.79 ± 0.30**	**49.58 ± 0.63**	72.85 ± 0.38	**67.93 ± 0.94**	49.35 ± 0.59	72.91 ± 0.32
S4: Joint Modelling	58.98 ± 0.18	38.92 ± 0.20	66.14 ± 0.07	64.86 ± 0.21	47.97 ± 0.12	**73.55 ± 0.17**	65.96 ± 0.49	49.57 ± 0.09	**74.86 ± 0.26**
S5: Mixed Training	60.65 ± 0.56	41.22 ± 0.58	67.13 ± 0.23	65.17 ± 0.56	46.59 ± 0.55	71.09 ± 0.13	66.35 ± 0.30	48.74 ± 0.07	72.72 ± 0.30
S6.1: Partial Sharing (25%)	61.97 ± 0.64	42.59 ± 0.54	68.31 ± 0.62	65.41 ± 0.30	46.30 ± 0.47	70.27 ± 0.32	66.30 ± 0.93	48.82 ± 0.83	73.25 ± 0.50
S6.1: Partial Sharing (50%)	60.89 ± 0.47	41.76 ± 0.98	68.53 ± 0.96	65.87 ± 0.52	47.22 ± 0.12	71.57 ± 0.34	66.37 ± 0.31	48.91 ± 0.42	73.18 ± 0.24
S6.2: Partial Sharing (75%)	**62.03 ± 0.60**	42.78 ± 0.88	68.65 ± 0.53	65.31 ± 0.41	46.76 ± 0.42	71.45 ± 0.57	66.37 ± 1.03	47.67 ± 1.05	71.70 ± 1.03

Table 4: TC, AC and AC|TC scores on ACE2005 (top) and RichERE (bottom), averaged over three random seeds (± standard error). For each column, the best performance is shown in bold and the second-best is underlined.

Rank	TC ACE2005	TC RichERE	AC ACE2005	AC RichERE	AC|TC ACE2005	AC|TC RichERE
1	Backward Transfer	Backward Transfer	Backward Transfer	Backward Transfer	Forward Transfer	Forward Transfer
2	Partial Sharing (50%)	Partial Sharing (50%) Partial Sharing (75%)	Partial Sharing (25%)	Independent	Independent	Independent Backward Transfer
3	Mixed Training Partial Sharing (25%)	Partial Sharing (25%)	Forward Transfer	Joint Modelling	Partial Sharing (25%)	Joint Modelling
4	Partial Sharing (75%)	Mixed Training	Independent	Forward Transfer Partial Sharing (50%)	Backward Transfer Partial Sharing (75%)	Partial Sharing (50%)
5	Independent Forward Transfer	Independent Forward Transfer	Partial Sharing (75%)	Partial Sharing (75%)	Joint Modelling	Partial Sharing (25%) Partial Sharing (75%)
6	Joint Modelling	Joint Modelling	Mixed Training	Partial Sharing (25%)	Mixed Training	Mixed Training
7			Partial Sharing (50%)	Mixed Training	Partial Sharing (50%)	
8			Joint Modelling			

Table 5: Model-averaged strategy rankings for each metric and benchmark combination. For each model and benchmark, strategies are ranked by score; ranks are then averaged across models to produce a consolidated ordering. A lower rank indicates better average performance.

stronger prior representations of linguistic and event-related semantics, and their greater parameter capacity allows them to accommodate different forms of task interaction more flexibly, learning both tasks effectively regardless of how training is coordinated. This is consistent with findings in related work showing that larger models exhibit greater robustness to task-level design choices in generative IE settings (Ravikumar et al., 2026).

4.1 Event Detection Performance (TC)

We discuss event detection performance across all eight strategies, as measured by the TC metric. Since TC evaluates only the ED component of the pipeline, strategies that differ solely in their EAE adapter—namely S1 and S2—produce identical TC scores by construction.

Backward Transfer is the strongest strategy for TC. S3 ranks first on both ACE2005 and RichERE in terms of performance based on TC, achieving the highest score in four out of six model-benchmark combinations. This consistent advantage suggests that initialising the ED adapter from a pre-trained EAE adapter provides a useful inductive bias. EAE training exposes the model to rich supervision over event participants and their semantic roles, encouraging representations that are broadly sensitive to event-bearing spans, a property that directly benefits event detection. The fact that this advantage persists after task-specific fine-tuning of the ED adapters suggests that the transferred representations provide a meaningful initialisation rather than being simply overwritten.

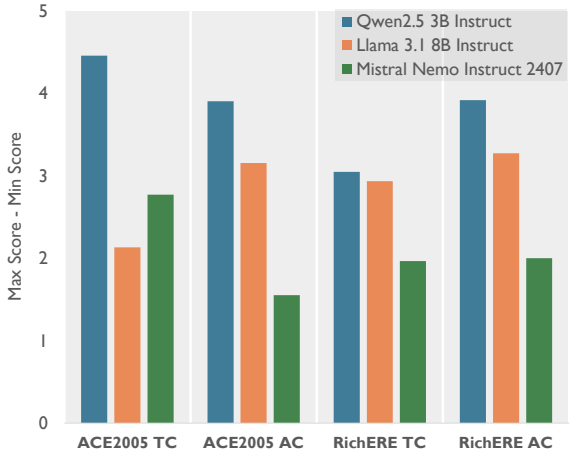


Figure 2: Score gap between the best and worst performing training strategies per model, across metrics and benchmarks.

Independent training leads to poor TC scores. S1 (and by construction S2) ranks second-last on both benchmarks in the model-averaged ranking, outperforming only Joint Modelling. The performance gap between S1 and S3 is particularly informative: the two strategies are identical except for how the ED adapter is initialised, making this a direct controlled comparison that isolates the effect of initialisation from all other experimental factors. The consistent advantage of S3 over S1 across all models and benchmarks confirms that cross-task initialisation provides measurable gains that independent training foregoes.

Joint Modelling is the weakest strategy for TC. S4 ranks last on TC across both benchmarks and all models. Unlike all other strategies, S4 trains on complete event structures and infers in a single pass, forcing a single set of adapters to jointly optimise for trigger identification and argument extraction within a unified output space. This conflation is detrimental to event detection: the competing demands of simultaneously producing triggers, event types, argument spans and roles within a single generation pass degrades the model's ability to precisely identify event-bearing spans.

Mixed Training, by contrast, is competitive for TC. S5 ranks third on ACE2005 and fourth on RichERE, outperforming Joint Modelling in both cases. Although both S4 and S5 use a single set of adapters, they differ fundamentally in both training and inference: S4 trains on complete event structures and infers in a single pass, while S5 trains on interleaved batches of ED and EAE instances and infers in a pipeline. This preservation of task boundaries—at both training and inference time—avoids the conflation that hampers S4. The competitive TC performance of S5 demonstrates that cross-task signals from interleaved training benefit event detection, and that the failure of S4 is attributable specifically to joint generation rather than parameter sharing.

Partial sharing yields strong TC performance. S6.2 (50% sharing) ranks second on both benchmarks, behind only Backward Transfer, making partial sharing the strongest TC strategy aside from S3. On ACE2005, S6.1 (25% sharing) and S6.3 (75% sharing) rank third and fourth respectively, while on RichERE, S6.3 matches S6.2 at rank two and S6.1 ranks third. These results suggest that sharing lower-layer representations between ED and EAE is broadly beneficial for event detection. Lower transformer layers encode common features useful to both subtasks, allowing ED to benefit from the additional supervision signal provided by EAE while retaining task-specific modelling capacity in higher layers.

The consistent advantage of S6.2 over both S6.1 and S6.3 on ACE2005, together with its parity with S6.3 on RichERE, suggests that moderate parameter sharing provides the most effective balance between cross-task transfer and task-specific specialisation. Sharing only 25% of layers yields weaker performance, indicating limited transfer between the subtasks, while increasing sharing to 75% does not provide further gains and can slightly reduce performance on ACE2005. Across both benchmarks, the strongest and most consistent results are obtained with 50% sharing, suggesting that moderate sharing is sufficient to transfer useful event-level representations without excessively constraining task-specific specialisation. This pattern is also reflected in the broader strategy ranking: S6.2 outperforms both fully independent training (S1; 0% sharing) and fully shared mixed training (S5; 100% sharing).

Overall TC paradigm ranking. At the strategy level, S3 and S4 represent the clear performance extremes, ranking first and last respectively. Between these extremes, the hybrid strategies cluster consistently in the upper-middle ranks, establishing partial parameter sharing as the most reliable paradigm for event detection after backward transfer. S5 performs competitively, while S1 and S2 rank in the

lower half, confirming that independent training without cross-task interaction is an ineffective strategy for event detection. Cross-task interaction with preserved task boundaries, whether through initialisation or parameter sharing, is broadly beneficial for TC performance.

4.2 Event Extraction (AC) and Event Argument Extraction (AC|TC) Performance

We discuss AC and AC|TC performance across all eight strategies. AC captures the compounded effect of both ED and EAE (end-to-end EE), while AC|TC isolates argument extraction quality by conditioning on correctly identified triggers.

Backward Transfer remains the strongest strategy for AC. S3 ranks first on AC for both benchmarks in the model-averaged ranking, continuing its strong performance from TC. This is attributable to two complementary factors: S3 achieves the highest TC scores, maximising the number of correctly identified triggers and thus the opportunities for successful argument extraction; and S3's largely strong AC|TC performance indicates that its dedicated EAE adapter, trained independently on argument extraction, extracts arguments effectively given correct triggers. Together, strong event detection and effective argument extraction account for S3's consistent advantage across models and benchmarks.

The disjoint paradigm performs strongly on AC and AC|TC. S1, S2 and S3 collectively dominate both the AC and AC|TC rankings with S2 ranking first for both benchmarks on AC|TC. This consistent superiority of the disjoint paradigm points to a clear finding: dedicated, task-specific EAE adapters are beneficial for argument extraction, providing the parameter capacity to specialise for the distinct demands of identifying event participants and their roles. Within this paradigm, S2 outperforms S1 and S3 on AC|TC across both benchmarks, demonstrating that initialising the EAE adapter from a pre-trained ED adapter improves argument extraction by providing a useful inductive bias without sacrificing dedicated task-specific capacity. This distinguishes forward transfer from other strategies that also introduce cross-task signal but at the cost of shared parameters.

Joint Modelling recovers on AC|TC despite weak TC. S4 ranks last on AC on ACE2005 but third on RichERE, despite ranking last on TC across both benchmarks. The key to this divergence lies in AC|TC: S4 ranks fifth on ACE2005 and third on RichERE, indicating that the unified output space that impairs trigger identification does not equally impair argument extraction. This asymmetry is structurally motivated: while trigger identification is a relatively self-contained objective that is harmed by the competing demands of simultaneous argument generation, arguments are by definition properties of their trigger, with each argument span and role conditioned on a specific trigger identity. Joint generation therefore reflects the natural structure of EAE, and the adapter learns to model trigger-argument relationships directly within a single pass. This coupling partially compensates for the absence of dedicated EAE parameter capacity. On RichERE, this AC|TC strength is sufficient to sustain competitive AC performance despite weak TC. On ACE2005, where argument extraction is substantially harder, the AC|TC performance cannot compensate for the poor TC, and AC collapses accordingly.

Mixed Training is the weakest strategy for AC and AC|TC. S5 ranks sixth on ACE2005 and last on RichERE for AC, and second-last on AC|TC on ACE2005 and last on RichERE, a striking reversal of its competitive TC performance. The contrast with S4 is instructive: despite underperforming S5 on TC, S4 consistently outperforms S5 on AC|TC. Both strategies share a single adapter set without dedicated EAE capacity, but S4's unified output space forces the adapter to jointly model trigger-argument relationships within a single generation pass, which benefits argument extraction given correct triggers. S5, trained on interleaved but separate ED and EAE instances, never jointly models these relationships and therefore lacks this benefit while equally lacking dedicated EAE parameter capacity. S5 thus benefits from neither the trigger-argument coupling of joint generation nor dedicated parameter capacity, accounting for its consistently poor argument extraction performance despite competitive TC.

Partial Sharing performs inconsistently on AC and AC|TC. The hybrid strategies exhibit pronounced benchmark dependence on both metrics. On AC, S6.1 (25% sharing) ranks second on ACE2005 but sixth on RichERE, while S6.2 (50% sharing) ranks seventh on ACE2005 but fourth on RichERE. S6.3 (75% sharing) is comparatively sta-

ble, ranking fifth on AC across both benchmarks. On ACE2005, where argument extraction is substantially harder, S6.1 performs best among hybrid configurations: preserving the greatest task-specific capacity in the upper layers outweighs the benefit of additional cross-task signal when the EAE objective is demanding. On RichERE, S6.2 performs best, suggesting that a more even balance between shared and task-specific capacity is optimal when argument extraction is less demanding. On AC|TC, the hybrid strategies are uniformly weak, with no configuration ranking above third on either benchmark, and their relative ordering shifts inconsistently across benchmarks. Taken together, the hybrid strategies offer no reliable advantage for argument extraction, with performance varying considerably across benchmarks and metrics.

Overall AC and AC|TC paradigm ranking. At the strategy level, S3 ranks first on AC across both benchmarks while S2 ranks first on AC|TC across both benchmarks. At the paradigm level, disjoint strategies perform best on average across both metrics, followed by hybrid and then fully shared training. Unlike TC, where cross-task parameter sharing drives performance improvements, argument extraction benefits most consistently from dedicated task-specific adapter capacity, with paradigm ranking inversely related to the degree of parameter sharing.

4.3 TC vs AC|TC: A Divergent Picture

Comparing TC and AC|TC rankings reveals a systematic shift in strategy performance across the two metrics. S3 ranks first on both TC and AC, but the relative ordering of all other strategies changes considerably. S2, which ranks in the lower half on TC, rises to first on AC|TC across both benchmarks, elevating the disjoint paradigm as the strongest overall for argument extraction. Conversely, the hybrid strategies, competitive on TC, drop substantially on AC|TC. S5, despite competitive performance on TC, falls to second-last on AC|TC on ACE2005 and last on RichERE. S4, the weakest strategy on TC, recovers to third on AC|TC for RichERE and fifth on ACE2005.

This divergence reflects a fundamental disparity between what each subtask demands. Event detection benefits from controlled cross-task interaction, whether through initialisation-based transfer as in S3, or partial parameter sharing as in the hybrid strategies. Argument extraction, by contrast, is more sensitive to dedicated task-specific capacity: the disjoint strategies, which maintain fully independent EAE adapters, consistently outperform shared and hybrid alternatives on AC|TC. S3 satisfies both requirements simultaneously—the ED adapter benefits from EAE initialisation while the independently trained EAE adapter retains full parameter capacity—which accounts for its consistent dominance across both TC and AC. S2 foregoes the benefit of cross-task initialisation for ED but retains full EAE capacity as well as ED-informed initialisation, producing the strongest AC|TC performance overall.

5 Conclusion

We present a systematic study of eight LoRA-based training strategies for generative event extraction, spanning three paradigms: disjoint, fully shared and hybrid parameter allocation. Experiments across three instruction-tuned LLMs and two benchmarks demonstrate that training strategy has a consistent and meaningful effect on extraction performance, with effects more pronounced at smaller model scales. Backward Transfer is the strongest strategy overall on TC and AC, while Forward Transfer is the strongest on AC|TC. More broadly, event detection benefits from cross-task inductive biases, whether through initialisation-based transfer or parameter sharing, while argument extraction is most reliably served by dedicated task-specific adapter capacity. We believe these findings may generalise beyond event extraction to other information extraction tasks that decompose into interdependent subtasks.

Limitations

Our evaluation covers three LLMs ranging from 3B to 12B parameters. While this provides diversity across model family and scale, extending the analysis to larger models would strengthen the generality of our findings. Similarly, while ACE2005 and RichERE differ in scale and schema complexity, evaluation on additional benchmarks spanning broader domains would provide a more comprehensive assessment. Likewise, evaluating on larger datasets would be insightful. We follow recommended practice and use a fixed 2-epoch training schedule across all settings; however, exploring longer training schedules and selecting checkpoints based on validation loss could further refine performance comparisons across strategies. Finally, our

experiments are conducted under LoRA-based fine-tuning. Whether the observed trends hold under full fine-tuning or alternative parameter-efficient methods remains an open question.

References

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023. Exploring the feasibility of chatgpt for event extraction. *Preprint*, arXiv:2303.03836.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. DEGREE: A data-efficient generation-based event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1890–1908, Seattle, United States. Association for Computational Linguistics.

I-Hung Hsu, Kuan-Hao Huang, Shuning Zhang, Wenxin Cheng, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2023. TAGPRIME: A unified framework for relational structure extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12917–12932, Toronto, Canada. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Kuan-Hao Huang, I-Hung Hsu, Tanmay Parekh, Zhiyu Xie, Zixuan Zhang, Prem Natarajan, Kai-Wei Chang, Nanyun Peng, and Heng Ji. 2024. TextEE: Benchmark, reevaluation, reflections, and future challenges in event extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12804–12825, Bangkok, Thailand. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

LDC. 2005. Ace (automatic content extraction) english annotation guidelines for events. Accessed: 2025-02-12.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2795–2806, Online. Association for Computational Linguistics.

Meta and 1 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Mistral AI. 2024. Mistral nemo. Accessed: 2026-03-28.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. *Preprint*, arXiv:2101.05779.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Rishi Ravikumar, Nuhu Ibrahim, and Riza Batista-Navarro. 2026. Lost in formatting: How output formats skew LLM performance on information extraction. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5498–5513, Rabat, Morocco. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *Preprint*, arXiv:1706.05098.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, Denver, Colorado. Association for Computational Linguistics.

Saurabh Srivastava, Sweta Pati, and Ziyu Yao. 2025. Instruction-tuning LLMs for event extraction with annotation guidelines. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13055–13071, Vienna, Austria. Association for Computational Linguistics.

The Comprehensive R Archive Network. *interpret_kendalls_w: Interpret Kendall's Coefficient of Concordance W*. CRAN. Accessed: 2026-05-18.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Wei Xiang and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

A Qualia-Based Audit of Procedural Event Annotations

Kyeongmin Rim and Marc Verhagen and James Pustejovsky
Brandeis University
Waltham, MA, USA
{krim, verhagen, jamesp}@brandeis.edu

Abstract

Procedural event annotations record *what changed* but not the semantic relevance or grounding of the change: whether the annotated entity is the kind of thing whose state matters for the domain. We present Entity Qualia Structure (EQS), a per-entity sortal-type categorization (coarsened from Generative Lexicon's type system to three categories: natural, artifactual, instrument) extracted from existing lexical resources. Applied to the OpenPI food domain, EQS reaches 84.7% coverage of the 518-item entity vocabulary; across 9367 transformation annotations, only 51.1% concern food entities themselves, while 30.2% record state changes of instruments, entities whose sortal type places them outside the food-state task. In a three-way comparison against existing cleanup efforts, EQS uniquely flags 15.6% of annotations that neither human re-annotation (OpenPI-C) nor LLM salience scoring (OpenPI 2.0) catches. Analysis of the AGENTIVE quale reveals that 93% of agentive-positive annotations involve instruments rather than food: entity creation can only be detected when the agentive feature is paired with the associated verb's event semantics.

1 Introduction

Crowdsourcing the annotation of entity state changes in procedural text to anonymous online workers has enabled datasets at large scale; however, in resources such as OpenPI (Tandon et al., 2020), which give crowd annotators free-text (open-vocabulary) input, the resulting annotations do not distinguish semantically central changes from incidental ones. Consider a recipe step *"Bake the cake for 30 minutes."* An annotator might record that the oven became "hot," the timer gone to "0," and the cake went from "raw batter" to "baked." All three are factually correct, but only the last tracks food state; the others describe instrument and timer state. The question is not *what changed*, but whether the entity undergoing change is *the kind of thing whose state matters* for the procedure's domain.

OpenPI's annotation quality has prompted independent cleanup efforts: Wu et al. (2023) produced OpenPI-C via three-stage human re-annotation, filtering $\sim$32% of state changes as not reliably inferrable from the input; Zhang et al. (2024) apply LLM-based clustering and per-step salience scoring (OpenPI 2.0). Both approaches address quality empirically, by re-annotating or filtering with human or LLM judgment. We offer a complementary symbolic angle: defining incidental annotations by the entity's sortal type, grounded in Generative Lexicon theory rather than annotator or model agreement. Our analysis shows that this uniquely catches 15.6% of annotations the empirical methods miss (§4).

Generative Lexicon (GL) theory (Pustejovsky, 1995) provides a formal basis for this distinction. An entity's FORMAL quale determines its sortal type, which we coarsen to natural, artifactual, or instrument for the audit; its AGENTIVE quale records whether the entity has an origin event with an associated agent (typically a maker). The contrast between *bake a potato* and *bake a cake* illustrates why both qualia matter: the same verb and event topology yield different outcomes because potato (a natural kind, no creation origin) is transformed, while cake (an artifact with a baking origin) is created, a phenomenon GL calls *co-composition*, where the event semantics is determined by the verb and its arguments' qualia jointly.

In this paper, we extract these qualia features from existing lexical resources and use them as a symbolic audit of OpenPI annotations. Our contributions are: (1) a cascade method that builds Entity Qualia Structure (EQS) data capturing coarsened GL sortal types from noun-focus language resources, covering 84.7% of the OpenPI food-domain vocabulary with a 32.2% cross-resource disagreement rate as a built-in quality diagnostic;

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 49–57

(2) an audit showing that nearly half of OpenPI food annotations are not about food at all, with instruments alone accounting for 30.2%, a mismatch directly readable from the entity's sortal type; and (3) an analysis of the AGENTIVE quale showing that 93% of agentive-positive annotations involve instruments, confirming that this theoretically motivated feature requires verb-side composition before it becomes discriminative. EQS provides the argument-side input for a compositional account of entity-state semantics; the complementary predicate-side analysis appears in Rim and Pustejovsky (2026).

2 Related Work

2.1 Entity State Annotation and Dataset Quality

The annotation of entity states in procedural text has evolved from tracking textual mentions to capturing implicit argument structures. Early work grounded entity state tracking in Semantic Role Labeling (Palmer et al., 2005) and qualia-based semantic tagging (GLML; Pustejovsky et al., 2009). ProPara (Dalvi et al., 2018) introduced entity tracking in procedural paragraphs with a closed set of state labels (created, destroyed, moved). Subsequent datasets have addressed implicit arguments (RISeC; Jiang et al., 2020), bridging relations under state transformation (RecipeRef; Fang et al., 2022), and entity identity and coreference using GL event models (CUTL; Rim et al., 2023). Kazeminejad et al. (2021) used the VerbNet semantic parser to automatically annotate entity existence and location states on ProPara, illustrating the value of symbolic, lexical-resource-grounded approaches for procedural entity-state work.

OpenPI (Tandon et al., 2020) scaled to open-vocabulary coverage across different procedural domains, but at the cost of the subeventual and ontological constraints found in the earlier literature. The cleanup efforts of Wu et al. (2023) and Zhang et al. (2024), discussed in §1, address the resulting reliability issues with human and LLM judgment respectively; we instead ground our audit in lexical-semantic resources.

2.2 Lexical-Semantic Resources for Events

Following GL's dual-aspect view of event semantics, we develop Entity Qualia Structure (EQS), an entity-side qualia representation that complements predicate-side resources such as VerbNet-GL (Brown et al., 2022). EQS builds on the GL tradition of computational lexicon construction: the SIMPLE ontology (Bel et al., 2000), which standardized GL-native semantic types across 12 European languages; the Brandeis Semantic Ontology (BSO; Pustejovsky et al., 2006; Havasi et al., 2007), an English lexicon informed by SIMPLE, publicly released alongside this work; CoreLex (Buitelaar, 1998), which derives systematic polysemy classes from WordNet (Fellbaum, 1998); and the Principle of Type Ordering (Pustejovsky, 2001), which formally justifies EQS's minimal feature set (sortal type + agentive quale availability) as sufficient for co-composition. The audit operationalizes the two qualia that surface directly in BSO entries: coarsened FORMAL (sortal type) and AGENTIVE (creation availability); TELIC enters the cascade indirectly through the type hierarchy.

2.3 Direct Precedents for Qualia Annotation

Prior work on annotating or extracting GL qualia informs the EQS cascade design. Pustejovsky et al. (2010) introduced the SemEval-2010 GLML task on argument selection and coercion, annotating whether the TELIC or AGENTIVE quale of a noun was activated in verb-argument context; this is the closest methodological ancestor to EQS, though scoped to verbal argument positions rather than discourse-level procedural entities. Yamada and Baldwin (2004) demonstrated automatic acquisition of TELIC and AGENTIVE roles from syntactic patterns and noted that domain-shifting lemmas (e.g., cooking entities appearing in natural-kind and artifact roles across documents) require token- or type-level resolution—the challenge EQS addresses through type-level coarsening. Bouillon et al. (2012) report annotator agreement on qualia annotation in Italian and French complex nominals, finding AGENTIVE relations reliably annotated by trained linguists, which empirically supports AGENTIVE as one of EQS's two operative qualia.

3 EQS: Extraction and Audit Method

OpenPI provides open-vocabulary state annotations across multiple procedural domains; we focus on the food slice for three reasons. First, open-vocabulary annotation is the property that makes annotation quality a research problem worth addressing with symbolic typing, and food is the slice where this property is best exercised by the existing data. Second, food preparation interleaves natu-

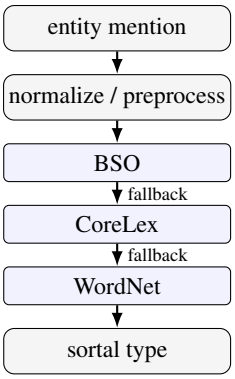


Figure 1: Sortal cascade pipeline. Entity mentions are normalized, then resolved against three lexical resources in priority order (first-resource-wins). The output is a coarsened sortal type that fills FORMAL.TYPE_CAT. Other EQS fields (FORMAL.INDIVIDUATION, AGENTIVE) are populated by direct single-resource lookup, not shown.

ral ingredients, artifactual dishes, and instruments within the same step, exercising exactly the FORMAL contrast that motivates EQS. Third, BSO's type hierarchy is densest in the food domain (its *Nourishment* subtree is the most developed branch), making food the natural first target for cascade evaluation.

3.1 Source Resources and Field Resolution

An EQS record captures two independent qualia features per entity type, grounded in Generative Lexicon theory: FORMAL (sortal type, coarsened to natural, artifactual, or instrument) and AGENTIVE (`yes` if the entity's type has a specific lexicalized creation activity in BSO, unspecified otherwise). These features are extracted automatically from existing lexical resources: the sortal type via the cascade described below, other fields via direct single-resource lookup. Figure 2 shows the AVM schema for two contrastive examples: *cake* (artifactual with a specific creation activity) and *knife* (instrument that is also GL-correctly agentive but whose creation is irrelevant to a recipe context).

The instrument-vs-artifact discrimination is the operative concern that motivates the choice of GL over simpler ontologies. WordNet's `noun.artifact` lumps knives, bowls, and baked goods together; ConceptNet (Speer et al., 2017) has no systematic instrument tag; Wikidata's (Vrandečić and Krötzsch, 2014) Q-types are too granular to coarsen reliably. In BSO, lexicalized TELIC roles are stored per-type, but the existence of a TELIC role is also encoded in the non-terminal

$$\left[\begin{array}{ll} \textit{eqs} & \\ \text{ENTITY} & \textit{cake} \\ \text{FORMAL} & \textit{artifactual} \\ \text{AGENTIVE} & \textit{yes} \end{array}\right]$$

$$\left[\begin{array}{ll} \textit{eqs} & \\ \text{ENTITY} & \textit{knife} \\ \text{FORMAL} & \textit{instrument} \\ \text{AGENTIVE} & \textit{yes} \end{array}\right]$$

Figure 2: EQS schema for two contrastive examples. Cake's agentive=yes (*Bake Activity* in BSO) licenses a creation reading under co-composition with verbs whose subevent matches; knife's agentive=yes (*Create Material Entity*) is GL-correct but operationally inert in recipe context.

node names of the type hierarchy (e.g., *Material Object with InstrumentTelic*). The cascade recovers the instrument category by ancestor walking the hierarchy alone, exploiting these node-name annotations; the per-type lexicalized values are reserved for finer-grained reasoning where ancestor walking is insufficient (e.g., N-N compound resolution; see §3.2).

The FORMAL field is resolved by a cascade of three lexical resources, applied in priority order with first-resource-wins per field. The BSO serves as the primary resource: entity stems are looked up in BSO's ∼50k-entry GL type hierarchy.[1] A set of ancestor-matching rules (hand-crafted for the food domain, external to BSO) walks the type tree to classify each sense (e.g., *Meat* → natural, *Artifactual Food* → artifactual, *Material Object with InstrumentTelic*[2] → instrument). For polysemous stems, all senses are checked; rule priority is set by their order in the domain profile, and the cascade returns the sense whose ancestor chain triggers the earliest-listed rule. CoreLex provides a polysemy-class fallback for entities outside BSO's coverage, mapping lemmas to basic types derived from systematic WordNet polysemy patterns, though it cannot distinguish instruments from general artifacts. WordNet supersenses serve as the last layer, mapping synsets to coarse type categories (`noun.plant` → natural, `noun.artifact`

[1]The cascade consumes the BSO release accompanying this paper at `http://brandeis-llc.github.io/bso`. BSO has been refined incrementally since its original publication (Pustejovsky et al., 2006), kept internal until this release; food-domain classifications are identical to the 2006 release used in earlier development of the cascade.

[2]Specific BSO type labels are subject to revision in future releases; the example rules here document the cascade's rule table at submission. The instrument-vs-artifact discrimination is structural in the type hierarchy and survives label changes.

Field	Value	Count	Source
FORMAL (coverage: 84.7%)	natural	205	BSO→CL→WN
	artifactual	118	
	instrument	116	
AGENTIVE (coverage: 18.0%)	yes	93	BSO core qualia
	None	425	

Table 1: EQS field resolution and FORMAL distribution on OpenPI Food vocabulary (518 entity types; 79 unresolved across five backend-limitation categories). Fields are independent: different resources, different GL layers.

→ artifactual); `noun.food` is skipped as ambiguous between natural and artifactual. Lexicalized food compounds (e.g., *ice cream*, *peanut butter*, *olive oil*) fall in WordNet's ambiguous `noun.food` class and remain unresolved by the cascade; these surface as one of the five backend-limitation categories discussed in §6.

The AGENTIVE field is resolved independently, by BSO core qualia lookup only. It is populated when BSO's type hierarchy annotates a specific creation activity on the entity's type, for example, *Bake Activity* on *Baked Good*, or *Stew Food Activity* on *Stew*. Generic placeholders (e.g., *Prepare Food Activity* on the *Food* supertype) are excluded, as they indicate the type *can* have origins, not that it has a specific lexicalized one. BSO stores the specific activity value, but EQS binarizes it to `yes/none`; the specific activity becomes relevant only at co-composition time, when the verb's subevent structure is available for matching (§5).

Our target corpus is the OpenPI "Food and Entertaining" subset restricted to concrete-goal documents across all three splits: 169 documents (train 150 / dev 12 / test 7), 840 steps, and 9367 transformation annotations. Table 1 summarizes coverage and distribution. Of 518 entity types in this vocabulary, 439 (84.7%) are resolved on FORMAL; the 79 unresolved entities fall into five backend-limitation categories (`noun.food` compounds, productive N-N compounds, brand names and rare lemmas, conjunctions, surface-form variants), discussed in §6.

3.2 Validation and Evaluation

Cross-resource disagreement. FORMAL resolution has no external gold standard at the type level. As an internal sanity check, every lemma is resolved against all three cascade backends unconditionally (bypassing the first-wins shortcut) and the resulting type categorizations are compared across resources; cross-resource disagreement serves as a built-in quality metric. Of 518 entities, 167 (32.2%) show cross-resource disagreement on FORMAL, falling into three patterns: (1) BSO=instrument vs. CL/WN=artifactual (∼85%; CL/WN lack an instrument category; BSO correct); (2) BSO=natural vs. CL=artifactual (polysemy in CoreLex; BSO correct); (3) 18 entities with no BSO coverage where CL and WN disagree, adjudicated against the SIMPLE OWL ontology (Toral and Monachini, 2007). The pattern analysis confirms first-wins resolution is correct in the large majority of cases.

Sample	n	correct	precision
precision-instrument	49	45	0.918
precision-food	40	25	0.625
recall sample (FN spot-check)	41	31	—

Table 2: FORMAL evaluation on a stratified sample. Estimated overall recall 0.575 (FN rate 0.244 extrapolated to the non-instrument population), giving F1 0.707.

Manual evaluation (FORMAL). We complement cross-resource disagreement with a stratified manual spot-check. A trained linguist familiar with GL theory adjudicated three samples: 50 instrument-predicted entities (precision-instrument), 50 food-predicted entities (precision-food), and 50 non-instrument entities mentioned in OpenPI annotations (recall, sampled false-negative-style). Adjudications use cascade evidence (BSO type chain, CoreLex polysemy class, WordNet supersense) for the prototype-level type judgment. We exclude 19 rows whose decision was sourced from outside the cascade backends (e.g., from CoreLex-only or WordNet-only fallback) since they cannot test cascade behavior directly; their adjudications remain valid type judgments and are reported separately as cascade-blind-spot evidence. Effective sample sizes are 49 / 40 / 42. Table 2 reports per-sample precision; the recall sample supports an estimated recall of 0.575 and F1 of 0.707 by extrapolating the false-negative rate to the non-instrument population.

Error analysis (N-N compounds). The dominant error pattern on precision-instrument is noun-noun compounds with a food-noun modifier and a container-noun head: *vodka bottle*, *vanilla extract bottle*, *marshmallow package*, *strawberry custard dish*. All four precision-instrument errors on the 50-entity sample come from this pattern. The cas-

Sub-sample	n	correct	rate
agentive=yes & food	12	11	0.917
agentive=yes & instrument	25	0	0.000
agentive=None (FN check)	25	8	0.320

Table 3: AGENTIVE evaluation on three sub-samples. "Correct" indicates whether the AGENTIVE label correctly reflects *recipe-relevant* creation: for agentive=yes rows, the entity is something the recipe creates (cake, bread); for agentive=None rows, the entity is not created in the recipe (pre-existing tools, ingredients). For instruments, every instance is GL-correct (the instrument *was* manufactured) but not recipe-relevant. 17/25 agentive=None cases are false negatives where BSO misses an applicable creation activity.

cade's head-noun fallback consistently selects the container reading; adjudicators selected the content reading because the modifier is food. This is the canonical N-N compound polysemy problem: morphology alone cannot distinguish [vodka bottle] (the contents) from [glass bottle] (the container). Ye et al. (2025) address this problem with a neural approach: LLM textual enrichment that surfaces qualia-role binding through prompt augmentation. A complementary symbolic route would consult the head noun's TELIC quale (encoded in BSO but not currently invoked by the cascade), matching the modifier's sortal type against the container's functional content type; this is a localized cascade extension rather than a methodological revision.

Manual evaluation (AGENTIVE). We additionally evaluate AGENTIVE on three sub-samples reflecting the field's three operational states (Table 3). The 12 agentive=yes food entities show 91.7% precision: when EQS asserts a specific creation activity for a food entity, the assertion is almost always correct. The 25 agentive=yes instrument entities show 0% operational relevance: every instance is GL-correct (the instrument *was* manufactured) but the creation history is irrelevant in recipe context. The agentive=None sample (25 entities) shows that BSO misses creation activities for 17 of 25 (68%); BSO's qualia coverage is sparser than the operational landscape suggests.

3.3 Auditing OpenPI

We operationalize *instrument* as an artifact whose procedural role is functional-use, not being transformed (state-changed): containers, tools, appliances, surfaces, and measurement gear. The operational test is whether the recipe *produces or transforms* the entity, or whether it *uses* the entity to do work on food; the latter is instrument.

The audit itself is a cross-reference: for each of the 9367 OpenPI food-domain annotations, we look up the entity's EQS FORMAL value and partition annotations by entity category. Entities classified as `instrument` are predicted to be peripheral to food-state tracking—their annotations record tool and container state (bowl weight, knife cleanness, oven temperature) rather than the event's food-level output, though such annotations may be informative for other purposes (e.g., workflow modeling). Entities classified as `natural` or `artifactual` are predicted to carry the food-state signal.

This is deliberately simple: a single symbolic feature (FORMAL) applied without any verb-side analysis, role binding, or co-composition. The contribution is showing how much incidental annotation one entity-level symbolic classification can detect. A richer consistency taxonomy (distinguishing *consistent* annotations (food entity, food-state change) from *lexically-underspecified* ones: food entity in a process event showing state change that the verb's subevent structure does not predict; cf. Generalized Result Role, Jezek and Melloni 2011; Rim et al. 2023) requires pairing the EQS classification with predicate-side subevent structure; Rim and Pustejovsky (2026) provide the complementary verb-side analysis.

4 Results: Incidental-Annotation Analysis

Table 4 presents the primary result. Cross-referencing EQS FORMAL against 9367 OpenPI food-domain annotations reveals that 51.1% involve entities classified as food (natural or artifactual). The largest incidental category is instrument tracking (30.2%): annotations on entities such as *bowl*, *knife*, *spoon*, and *blender*, classified as `instrument` by the cascade, whose identity persists through the event regardless of what attribute changes annotators recorded. The 18.7% unresolved category includes entities not in the EQS vocabulary, falling across five backend-limitation categories (§6). A complementary keyword-based analysis identifies 49% of annotations as low-value via attribute-name matching (location 21.5%, weight 7.6%, cleanness 7.3%); the two methods are complementary, as EQS catches incidental annotations from non-food entities entirely while keyword matching catches incidental annotations *within* food-entity records (e.g., location-only changes on

Category	Trans.	%
Food signal (natural + artifactual)	4786	51.1
Instrument (incidental)	2829	30.2
Unresolved / not in EQS	1752	18.7
Total	9367	100.0

Table 4: OpenPI food annotations by EQS entity category. 51.1% involve actual food entities; 30.2% is incidentally tracked instrument state detected by FORMAL, not surface keywords.

Flagged by	Trans.	% of 9367
EQS only	1458	15.6
OpenPI-C only	2760	29.5
OpenPI 2.0 (local-salience) only	286	3.1
All three	222	2.4
None (kept by all three)	3109	33.2
EQS total	2829	30.2
OpenPI-C total	4350	46.4
OpenPI 2.0 total	1055	11.3

Table 5: Three-way comparison: EQS instrument-flag vs. OpenPI-C re-annotation vs. OpenPI 2.0 per-step low local-salience (≤ 2). EQS uniquely catches 15.6% of annotations the empirical methods miss. We do not use V2's entity clusters or paraphrase expansion because those layers have sub-50% F1 against unreliable gold.

food items).

Comparison with empirical cleanup methods. Table 5 compares EQS's instrument-flag against the two empirical cleanup methods reviewed in §2: OpenPI-C's three-stage human re-annotation (Wu et al., 2023) and OpenPI 2.0's per-step LLM salience scoring (Zhang et al., 2024). EQS uniquely flags 15.6% of annotations that neither OpenPI-C nor OpenPI 2.0's local salience catches: instrument-typed entities whose state changes the empirical methods retain as relevant. Only 2.4% of annotations are flagged by all three methods simultaneously; conversely, 33.2% are retained as signal by every method, providing a high-confidence consensus subset. The three filters are complementary: they ground their decisions on different evidence (entity ontology, annotator agreement, LLM-derived salience) and catch different incidental categories.

5 Discussion

5.1 When Argument Qualia Become Operative

The FORMAL field drives the incidental-annotation analysis in §4, but GL theory predicts that the AGENTIVE quale should provide a finer distinction: whether an entity undergoes *creation* (its origin process is instantiated by the event) or merely *transformation* (its identity is preserved). We test this prediction by cross-referencing the AGENTIVE field against OpenPI annotations.

Of 2626 annotations on AGENTIVE=yes entities, 93.1% involve instruments such as *bowl*, *blender*, *pan*, and *knife* (Table 6). These are GL-correctly agentive: a knife *was* manufactured, and BSO records a *Create Activity* on its type. But a knife's origin process is irrelevant in a recipe context; the recipe does not create knives. Only 12 food entity types (182 annotations, 6.9%) have genuinely operative AGENTIVE qualia: *cookies*, *bread*, *cake*, *crust*, *wine*, *beer*, among others, whose creation process (baking, brewing) may actually be instantiated by the procedural verb.

The AGENTIVE quale answers a type-level question *"does this entity have a lexicalized origin process?"* not a compositional one *"does* this event *instantiate that origin?"* The distinction between a knife's irrelevant manufacturing and a cake's operative baking emerges only when the verb's semantic structure is available for matching: *bake* instantiates cake's AGENTIVE quale (*Bake Activity*); *move* does not. BSO stores the specific activity value (not just yes/none), making this matching feasible in principle, but it requires the predicate side: a co-composition operator that is beyond the scope of the present work.

This finding mirrors a complementary result from the predicate side: Rim and Pustejovsky (2026) show that VerbNet-GL's verb-only prediction achieves only 29.4% overall accuracy for entity identity change, because the verb's subevent structure cannot determine the outcome without argument qualia. Our agentive analysis confirms the converse: argument qualia cannot determine the outcome without the verb. The 76% of process-event steps that show state changes in the OpenPI data despite VN-GL predicting no result state provide further context: EQS's FORMAL can distinguish which of these are legitimate (food entities undergoing implicit transformation) from incidental (instrument state tracking), but the full resolution (predicting *what kind* of change each food entity undergoes, e.g., whether AGENTIVE licenses creation or transformation) requires both sides of the composition.

Category	Trans.	% of ag=yes
Instrument	2444	93.1
Food (natural + artifactual)	182	6.9
Total agentive=yes	2626	100.0

Table 6: Breakdown of AGENTIVE=yes transformations. 93.1% involve instruments, GL-correctly agentive (manufactured), but whose creation is irrelevant to the procedural context. The feature becomes operative only when composed with the verb's subevent structure.

5.2 Repairing the Symbolic Backbone

The audit exposes lexicographic coverage gaps that the symbolic backbone alone cannot resolve. A representative case: BSO has *platter* with senses {Food, Music Artifact} but no Tableware/Dish sense, so the food-domain rule priority picks the Food reading and the contextually-correct tableware reading is unreachable; sibling lemmas (*bowl*, *plate*, *serving bowl*) resolve correctly because a Dish sense *is* present, and absence is silent.

These gaps live in the data layer, not in the A-V extraction method, so the natural place for repair sits outside our pipeline. Manual lexicographic curation is the canonical option, but it scales poorly against a resource the size of BSO. Considering recent advancements in language models, one promising route is LLM-assisted, corpus-rooted reevaluation, where model proposals surface candidate senses or relations and the symbolic backbone serves to validate them against the existing type hierarchy: a division of labor that uses each side for what it does best. The cross-resource disagreement signal we use for validation (§3.2) is a useful place to begin, since it already flags the lemmas where coverage is contested.

5.3 Domain Specificity of the Work

Domain specificity in our pipeline is concentrated at two points, both surfaced through a single `DomainProfile` object in the implementation: the BSO ancestor rules (the paper's explicit food-domain contribution, a small reviewable GL-grounded artifact), and the WordNet `noun.food` supersense skip. The most immediate generalization target is OpenPI's Home and Garden topic, the second-largest concrete-procedure slice within the same corpus (146 concrete documents vs. Food's 169; 5552 concrete transformations vs. Food's 9367) and conceptually adjacent, requiring only `DomainProfile` changes rather than architecture changes. Procedural domains farther from cooking (medical protocols, industrial workflows, scientific procedures) use the same attach point but additionally require sourcing domain-appropriate corpora beyond OpenPI.

A related concern is *within-resource bias*: each backend brings its own design choices that bias classification. Polysemy resolution illustrates this: lemmas like *oil* (BSO senses Fat, Painting, Combustible, Ointment) or *dressing* (Clothing Artifact, Seasoning) are resolved by walking each sense against the BSO ancestor rules, and for food-domain entities this typically lands on the intended sense, but the mechanism is an implicit domain prior rather than contextual disambiguation. Comparable choices in WordNet supersense priority and CoreLex polysemy classes contribute their own biases. These are bias *by design* in the food domain; the same architecture, configured differently, would express different biases in another domain.

A complementary kind of adaptability concerns the task rather than the domain: the framing of an entity as "incidental" is itself task-relative. Instrument annotations are peripheral to food-state tracking but signal for workflow- or tool-modeling tasks, where tracking pan temperature and knife cleanness IS the point. EQS's symbolic typing offers a re-orientable filter rather than a fixed noise/signal partition; swapping the relevance class re-uses the same cascade output.

6 Conclusion

We have presented three contributions. First, EQS: a per-entity qualia representation automatically extracted from noun-level lexical resources, achieving 84.7% coverage on the OpenPI food-domain vocabulary with a 32.2% cross-resource disagreement rate that serves as a built-in quality diagnostic. Second, an incidental-annotation analysis showing that 51.1% of OpenPI food annotations track genuine food-state change, with 30.2% constituting incidental instrument tracking detectable by entity type alone. Third, an agentive analysis demonstrating that argument qualia—even when GL-correctly assigned—are not operative without semantic counterpart: 93% of agentive-positive annotations involve instruments whose creation is irrelevant to the procedural context.

A primary limitation is coverage: the 18.7% unresolved category is not exclusively low-frequency entities, but falls into five backend-limitation cat-

egories, each pointing to a different repair locus. Two sit at the lexical-data layer. `noun.food` compounds (e.g., *ice cream*, *peanut butter*, *olive oil*) fall into WordNet's ambiguous food class that the cascade deliberately skips because it conflates natural and artifactual readings; resolving them needs either a curated lexicalized-compound list or the kind of LLM-assisted sense disambiguation discussed in §5. Brand names and rare lemmas are out-of-vocabulary in all three backends and require either resource extension or surface-form heuristics (capitalization, brand databases).

Two sit at the preprocessing layer. Surface-form variants (plurals, inflections, casing) fall through because BSO is keyed on singular stems, addressable by adding a better lemmatization at lookup time or by extending the stem index with surface variants. Conjunctions (*salt and pepper*, *cream and sugar*) decompose into individually-classifiable components, but conjoined mentions are not currently split before lookup.

Productive N-N compounds (*vodka bottle*, *strawberry custard dish*) sit at the compositional layer and are the theoretically interesting case; their resolution requires reasoning over both modifier and head qualia, as outlined in the N-N error analysis (§3.2).

More broadly, these results suggest that symbolic qualia structure can provide an ontological audit layer that purely human-judgment and LLM-judgment approaches to annotation quality lack. Future work will complete the compositional picture by matching the verb-side semantics against the EQS representation; address the domain-specificity limitation discussed in §5 by scaling the audit to other procedural domains, starting with OpenPI's immediately adjacent Home and Garden slice and extending to medical, industrial, and scientific procedure corpora; and feed the resulting typed entity layer into operational entity-state pipelines that require sortal-type filtering at input.

References

Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Alessandro Lenci, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas, and Antonio Zampolli. 2000. SIMPLE: A general framework for the development of multilingual lexicons. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Pierrette Bouillon, Elisabetta Jezek, Chiara Melloni, and Aurélie Picton. 2012. Annotating qualia relations in Italian and French complex nominals. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*.

Susan Windisch Brown, Julia Bonn, Ghazaleh Kazeminejad, Annie Zaenen, James Pustejovsky, and Martha Palmer. 2022. Semantic Representations for NLP Using VerbNet and the Generative Lexicon. *Frontiers in Artificial Intelligence*, 5.

Paul Buitelaar. 1998. *CoreLex: Systematic Polysemy and Underspecification*. Ph.D. thesis, Brandeis University.

Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. 2018. Tracking state changes in procedural text: a challenge dataset and models for process paragraph comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1595–1604, New Orleans, Louisiana. Association for Computational Linguistics.

Biaoyan Fang, Timothy Baldwin, and Karin Verspoor. 2022. What does it take to bake a cake? the RecipeRef corpus and anaphora resolution in procedural text. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3481–3495, Dublin, Ireland. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Catherine Havasi, Anna Rumshisky, and James Pustejovsky. 2007. An evaluation of the Brandeis Semantic Ontology. In *Proceedings of the Fourth International Workshop on Generative Approaches to the Lexicon (GL2007)*.

Elisabetta Jezek and Chiara Melloni. 2011. Nominals, polysemy, and co-predication. *Journal of Cognitive Science*, 12(1):1–31.

Yiwei Jiang, Klim Zaporojets, Johannes Deleu, Thomas Demeester, and Chris Develder. 2020. Recipe instruction semantics corpus (RISeC): Resolving semantic structure and zero anaphora in recipes. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 821–826, Suzhou, China. Association for Computational Linguistics.

Ghazaleh Kazeminejad, Martha Palmer, Tao Li, and Vivek Srikumar. 2021. Automatic entity state annotation using the VerbNet semantic parser. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 123–132.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

James Pustejovsky. 2001. Type construction and the logic of concepts. In Pierrette Bouillon and Federica Busa, editors, *The Language of Word Meaning*, pages 91–123. Cambridge University Press.

James Pustejovsky, Catherine Havasi, Jessica Littman, Anna Rumshisky, and Marc Verhagen. 2006. Towards a generative lexical resource: The Brandeis semantic ontology. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

James Pustejovsky, Jessica Moszkowicz, Olga Batiukova, and Anna Rumshisky. 2009. GLML: Annotating argument selection and coercion. In *Proceedings of the Eight International Conference on Computational Semantics*, pages 169–180, Tilburg, The Netherlands. Association for Computational Linguistics.

James Pustejovsky, Anna Rumshisky, Alex Plotnick, Elisabetta Jezek, Olga Batiukova, and Valeria Quochi. 2010. SemEval-2010 task 7: Argument selection and coercion. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval)*, pages 27–32.

Kyeongmin Rim and James Pustejovsky. 2026. Subevent structure as a predictor of entity identity change in procedural text. In *Proceedings of the Workshop on Structured Linguistic Data and Evaluation (SLiDE)*.

Kyeongmin Rim, Jingxuan Tu, Bingyang Ye, Marc Verhagen, Eben Holderness, and James Pustejovsky. 2023. The coreference under transformation labeling dataset: Entity tracking in procedural texts using event models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12448–12460, Toronto, Canada. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pages 4444–4451.

Niket Tandon, Keisuke Sakaguchi, Bhavana Dalvi, Dheeraj Rajagopal, Peter Clark, Michal Guerquin, Kyle Richardson, and Eduard Hovy. 2020. A dataset for tracking entities in open domain procedural text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6408–6417, Online. Association for Computational Linguistics.

Antonio Toral and Monica Monachini. 2007. SIMPLE-OWL: A Generative Lexicon ontology for NLP and the semantic web. In *Proceedings of the workshop on GL2007*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Xueqing Wu, Sha Li, and Heng Ji. 2023. OpenPI-C: A better benchmark and stronger baseline for open-vocabulary state tracking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7213–7222, Toronto, Canada. Association for Computational Linguistics.

Ichiro Yamada and Timothy Baldwin. 2004. Automatic discovery of Telic and Agentive roles from corpus data. In *Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation (PACLIC)*.

Bingyang Ye, Jingxuan Tu, and James Pustejovsky. 2025. Enhanced noun-noun compound interpretation through textual enrichment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25896–25911, Suzhou, China. Association for Computational Linguistics.

Li Zhang, Hainiu Xu, Abhinav Kommula, Chris Callison-Burch, and Niket Tandon. 2024. OpenPI2.0: An improved dataset for entity tracking in texts. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 166–178, St. Julian's, Malta. Association for Computational Linguistics.

Benchmarking Models for Low-Resource Nepali Event Extraction with Trigger Phrase Identification and Event Classification

Sujal Maharjan[1]*, Astha Shrestha[1]*, Lakshmojee Koduru[2], Sweta Poudel[3], Shuvam Shiwakoti[4], Rabin Thapa[1], Kritesh Rauniyar[5], Surendrabikram Thapa[4]
[1]IIMS College, Kathmandu, Nepal, [2] Google
[3]Kathmandu Engineering College, Tribhuvan University, Kathmandu, Nepal
[4]Virginia Tech, USA, [5]Macquarie University, Australia
sujalmaharjan007@gmail.com, aasthashrestha688@gmail.com

Abstract

Research on Event Extraction (EE) in South Asian languages is crucial for understanding information dissemination and enabling automated news analysis in morphologically rich, low-resource settings. To address the scarcity of high-quality, publicly available datasets, we present Nepali Event Extraction (NepEE), a manually annotated corpus comprising 10,226 Devanagari sentences. The dataset includes annotations for trigger identification and event type classification, achieving high inter-annotator agreement with Fleiss' $\kappa = 0.812$ for trigger identification and $\kappa = 0.855$ for event classification. Our dataset was developed through a rigorous iterative three-phase protocol involving five expert native speakers to ensure linguistic precision. We conduct benchmarking across a broad spectrum of approaches, including classical feature-based models, five fine-tuned Transformer encoders, and contemporary instruction-tuned Large Language Models (LLMs) using zero-shot and fixed few-shot prompting. Our analysis shows that Indic-specialized Transformers achieve superior classification performance, while traditional methods and few-shot prompting struggle with the challenges of exact span extraction in morphologically complex contexts. Furthermore, we quantify performance differences between sentence-level and span-level tasks, establishing strong baselines for future research. The findings and released NepEE dataset provide a valuable resource for advancing event understanding in low-resource languages (LRLs). The dataset, code, and experimental resources are publicly available at GitHub/SUJAL390/EEUCA-ACL-2026.

1 Introduction

Event Extraction (EE) is an important task in Information Extraction (IE), moving beyond entity recognition toward identifying structured event information from unstructured text (Hürriyetoğlu et al., 2025; Xiang and Wang, 2019; Li et al., 2022). The task is commonly divided into two interrelated subtasks: Trigger Identification, which involves anchoring an occurrence to its most salient lexical unit, and Event Type Classification, which maps that anchor to a specific node in a predefined semantic taxonomy. While the field has witnessed a paradigm shift toward high-performance neural architectures facilitated by mature benchmarks such as the Automatic Content Extraction (ACE) 2005 and the Event and Relation Extraction (ERE) corpora (Doddington et al., 2004; Walker et al., 2006), these advances have largely bypassed low-resource languages (LRLs) like Nepali. This digital divide creates a significant bottleneck for the deployment of context-aware systems in the South Asian region, where timely IE is a prerequisite for applications ranging from automated news synthesis to real-time disaster response monitoring (Grishman, 2019).

Nepali, an Indo-Aryan language spoken by approximately 30 million individuals, exhibits a complex set of linguistic peculiarities that challenge traditional extraction frameworks based on Western European languages. Nepali, a morphologically rich SOV language, employs intricate agglutinative structures and compound verb clusters, e.g., नियुक्त भए (*was appointed*), to denote actions. Unlike English, where a trigger is often a distinct lexical unit, Nepali triggers are frequently nominalized or split, where a verbal noun such as सम्झौता (*agreement*) carries the primary semantic load of the event. Furthermore, the extensive use of honorifics and auxiliary inflections necessitates granular character-level span detection to

*The authors contributed equally to this work and are designated as joint first authors. The author order follows alphabetical order by last name.

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 58–71

avoid the inclusion of extraneous morphological markers. Recent research in the region, including the Nepali Language Understanding Evaluation (NLUE) benchmark (Nyachhyon et al., 2025), has successfully pioneered foundational tasks like Named Entity Recognition (NER) and Part-of-Speech (POS) tagging; however, event-level semantic parsing remains underexplored.

To bridge this infrastructure gap, we present a novel, high-quality, human-annotated dataset specifically curated for Nepali trigger identification and event type classification. Recognizing the inherent subjectivity and linguistic nuance involved in semantic labeling, we implemented a rigorous annotation protocol involving five native Nepali speakers with advanced expertise in linguistics and media analysis. To ensure the scientific validity and reproducibility of the corpus, we conducted an exhaustive inter-annotator agreement analysis, yielding a Fleiss' κ of 0.812 for trigger identification and 0.855 for event type classification. According to the diagnostic benchmarks established by Landis and Koch (1977), these coefficients indicate near-perfect agreement, validating our annotation guidelines as a robust and scalable framework for capturing the nuances of Devanagari event semantics. Our schema covers eight diverse event categories, including *Disaster and Accidents*, *Political*, and *Economic Event*, providing a representative cross-section of the contemporary Nepali media landscape.

Beyond the introduction of the corpus, this paper establishes a comprehensive computational baseline by evaluating the proposed dataset across four distinct modeling paradigms to delineate the current performance ceiling. We contrast classical machine learning frameworks, such as feature-engineered Support Vector Machines (SVM) and Random Forests, with state-of-the-art (SOTA) Transformer-based architectures and Large Language Models (LLMs). This evaluation includes massive multilingual encoders such as mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020), alongside regionally optimized models like IndicBERTv2 (Doddapaneni et al., 2023) and the specialized nepaliBERT (Ghimire, 2022). Furthermore, we benchmark the generative capabilities of leading LLMs including Qwen-2.5 (Qwen et al., 2025), Gemma-3 (Team et al., 2025), Phi-4 (Abdin et al., 2024), and Llama-3.1 (Grattafiori et al., 2024) under both zero-shot and fixed few-shot prompting configurations. Our results not only provide a rigorous performance benchmark but also offer a diagnostic analysis of the hard cases in Nepali EE, such as indirect phrasing and nominalized triggers, that continue to challenge modern Natural Language Understanding (NLU). Through this work, we provide the foundational data and empirical framework necessary to support future research on South Asian NLU, fostering a more inclusive and linguistically diverse global AI ecosystem.

2 Related Work

To contextualize the challenges addressed by NepEE, we review prior research spanning event extraction benchmarks, multilingual transformer architectures, and Nepali natural language understanding.

2.1 Foundational Benchmarks and the Evolution of Event Extraction

The trajectory of EE as a distinct sub-discipline of IE has been fundamentally shaped by the availability of high-quality, human-annotated corpora. Early foundational efforts were anchored by the ACE 2005 program (Doddington et al., 2004) and the subsequent ERE datasets (Walker et al., 2006), which established the canonical two-stage paradigm: Trigger Identification and Argument Role Labeling. While these benchmarks facilitated the transition from pattern-matching heuristics to statistical and neural models, their linguistic foundations are deeply rooted in Western European syntax.

Recent scholarship has highlighted the inadequacy of these schemas when applied to languages with distinct typological features. As noted by Grishman (2019), the reliance on distinct, monolexemic triggers in English does not easily port to languages where event anchors are distributed across complex morphological clusters. To address the limitations of the small-scale ACE corpora, the community introduced massive general-domain datasets such as MAVEN (Wang et al., 2020) and RAMS (Ebner et al., 2020), which expanded the taxonomic depth of events to thousands of categories. However, these datasets continue to exhibit a significant high-resource bias. Our work bridges this gap by introducing a standardized semantic benchmark for Nepali, focusing on the structural complexities of trigger identification and event type classification.

2.2 The Cross-Lingual Gap in Transformer Architectures

The advent of pre-trained Transformer architectures has redefined the state-of-the-art for sequence labeling and semantic parsing. Multilingual encoders, such as mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020), have demonstrated remarkable cross-lingual transfer capabilities by mapping diverse languages into a shared embedding space. While foundational research in domain adaptation suggests that learning shared representations is an optimal strategy for transfer learning (Glorot et al., 2011; Bengio, 2012), recent empirical studies on Low-Resource Languages (LRLs) reveal a more nuanced reality. Specifically, massive multilingual models often exhibit a "curse of multilinguality," where the representation quality for any single language may be diluted as the number of supported languages increases under a fixed parameter budget. However, this limitation is not an absolute law; recent work suggests that increasing model capacity, utilizing Mixture-of-Experts (MoE), or leveraging high-quality instruction-tuning data (such as Bactrian-X (Li et al., 2023)) can effectively mitigate these bottlenecks. Despite the rise of Large Language Models (LLMs), fine-tuned encoders remain indispensable for specific extraction tasks, as zero-shot prompting of generalist models can still underperform compared to dedicated, task-specific baselines (Chen et al., 2024).

For morphologically rich scripts, research indicates that monolingual pre-training or specialized vocabulary adaptation significantly outperforms zero-shot cross-lingual transfer (Pfeiffer et al., 2020). In the South Asian context, the development of Nepali-BERT Ghimire (2022) represented a significant milestone, providing a model trained on native Devanagari corpora that captures the unique distributional semantics of Nepali more effectively than generalized multilingual counterparts. Our benchmarking framework extends this line of inquiry by evaluating whether these monolingual advantages translate to high-level semantic tasks like sentence-level trigger identification, particularly in the presence of complex agglutinative inflections and honorific markers that are often absent in multilingual training sets.

2.3 Event Extraction in Indic and South Asian Languages

Research into EE for the Indo-Aryan language family has gained momentum, yet remains fragmented. Studies in Hindi have leveraged deep neural architectures for event-centric extraction tasks (Sahoo et al., 2020). However, Nepali presents a unique typological profile characterized by its specific handling of nominalized triggers and compound verb clusters. Unlike Hindi, where certain auxiliary structures are more standardized, Nepali exhibits a higher degree of verbal agglutination where the semantic core of an event may be embedded within a multi-token cluster (e.g., नियुक्त भए) (*was appointed*).

Furthermore, existing IE efforts in the region often conflate Nepali with other Devanagari-based languages, failing to account for its distinct verb-final (SOV) constraints and its extensive use of verbal nouns as action anchors. Recent work on the IndicNLPSuite (Kakwani et al., 2020) provided foundational resources for many South Asian languages, yet high-level tasks like event-centric NLU for Nepali were not included in the original benchmarks. By introducing a human-annotated EE corpus with near-perfect inter-annotator agreement ($\kappa > 0.8$), we provide one of the first large-scale efforts to formalize event-level semantics for Nepali, distinguishing it from general Indic-language extraction models.

2.4 The Consolidation of Nepali Natural Language Understanding

Historically, Nepali NLP was confined to foundational tasks such as rule-based (POS) tagging, Named Entity Recognition (NER) like Everest-NER (Niraula and Chapagain, 2022), and basic sentiment Analysis (Bal, 2004). The landscape underwent a rapid modernization with the introduction of the Nep-gLUE and NLUE benchmark (Timilsina et al., 2022; Nyachhyon et al., 2025). These benchmarks provided standardized datasets for NER, question answering, and document classification, thereby creating a performance baseline for the language.

Despite these advances, a significant gap remains in the domain of high-level semantic extraction (Rauniyar et al., 2023; Thapa et al., 2023). Current Nepali NLU benchmarks primarily focus on entity-level or document-level tasks, leaving the intermediate layer of sentence-level event se-

mantics unaddressed. As Nyachhyon et al. (2025) argue, the development of sophisticated NLU for Nepali requires moving beyond foundational syntax toward structured knowledge extraction. We present a novel gold-standard dataset for Nepali EE, covering both trigger identification and event type classification. Our work establishes a benchmark for this task by evaluating a range of models, including classical machine learning algorithms, multilingual transformers, domain-specific Nepali transformers, and LLMs (zero-shot and fixed few-shot prompting). This framework provides a rigorous diagnostic for how different architectures handle the unique semantic and structural challenges of the Devanagari script.

3 Dataset

In this section, we describe our data collection process and the iterative annotation schema developed for the Nepali EE task.

3.1 Data Collection

The dataset was constructed from a publicly released corpus of 65,000 Nepali sentences (Paudyal, 2017). From this base corpus, we curated sentences through a controlled selection procedure designed to ensure event salience and annotation suitability.

Sentences containing explicit eventive expressions, including verbal predicates and nominalized forms, were prioritized. Each sentence was manually reviewed by five trained native Nepali speakers to verify semantic completeness and contextual interpretability prior to annotation. During selection, we continuously monitored class frequencies and applied corrective sampling to maintain balanced representation across the eight event categories. The final dataset consists of 10,226 sentences with a stable class distribution, as shown in Table 2.

3.2 Annotation Process

To ensure high-quality annotations, we engaged five experienced native Nepali speakers possessing a deep understanding of local linguistic structures and media discourse. Annotators were provided with comprehensive guidelines, complete with illustrative examples, for the two primary tasks: trigger identification and event type classification. To maximize inter-annotator consistency and resolve linguistic ambiguities, we implemented a structured, iterative three-phase annotation schema (illustrated in Figure 1). This protocol consisted of an initial dry run, an instruction revision phase, and a final conflict resolution phase.

- **Initial Dry Run**: We initiated the annotation process with a dry run of 40 sample sentences. This phase was crucial in gauging the effectiveness of the guidelines. Initially, annotators faced confusion in identifying the minimal span for compound verb clusters. For instance, in the phrase अनुदान दिएको छ (*has provided a grant*), some annotators selected the entire phrase, while others selected only the core nominalized trigger अनुदान (*grant*). These edge cases were logged for subsequent guideline refinement.

- **Instruction Revision Phase**: Building upon insights from the dry run, the annotation process entered a second phase where 100 additional sentences were annotated. During this phase, annotators were provided with refined instructions, which were adjusted based on the feedback from the initial dry run. This step aimed to enhance the clarity and precision of annotations, particularly in identifying split predicates and character-level boundaries for triggers.

- **Conflict Resolution**: In the final stage, annotators engaged in a collaborative discussion to address discrepancies that arose while annotating 100 sentences after the revision of instructions. This consensus-building process allowed for a thorough review of annotations and a shared understanding of the final guidelines. The resolution of occasional ambiguities was achieved through regular meetings and consultations with experts in annotation. The resolution of ambiguities ensured consistency and accuracy of annotations, enhancing the overall quality of the Nepali Event Extraction (NepEE) dataset.

3.3 Annotation Guidelines

To ensure annotation consistency and linguistic consistency, we devised detailed annotation guidelines to assist the annotators. Given a sentence, it was annotated for two primary interdependent tasks: trigger identification and event type classification.

Table 1: Comparative Summary of Benchmarking and Event Extraction Tasks. EN = English, ZH = Chinese, AR = Arabic, NE = Nepali, Trigger ID = Trigger Identification, Event Class. = Event Type Classification,

Work	Task	Datasets	Data Size	Language
Doddington et al. (2004)	Entity, Relation, Event Ext.	ACE	∼1.1M words	EN, ZH, AR
Wang et al. (2020)	Event Detection (ED)	MAVEN	118,732 instances	EN
Ebner et al. (2020)	Multi-Sentence Arg. Linking	RAMS	9,124 events	EN
Kakwani et al. (2020)	Multilingual NLU Bench.	IndicGLUE	2,473,708 instances	11 Indic
Timilsina et al. (2022)	Nepali NLU Benchmarking	Nep-gLUE	286,941 annotations	NE
Nyachhyon et al. (2025)	NLU Benchmarking (12 tasks)	NLUE	∼341K instances	NE
Ours	**Trigger ID & Event Class.**	**NepEE**	**10,226 sentences**	**NE**

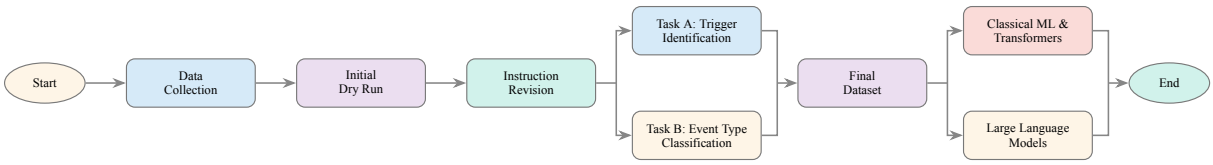


Figure 1: Overview of the end-to-end pipeline for the NepEE dataset.

Table 2: Distribution of class labels for the Event Type Classification task.

Task Label	#Samples	%
Political Event	1,419	13.8
Other Event	1,397	13.7
Economic Event	1,371	13.4
Sports Event	1,289	12.6
Entertainment Event	1,238	12.1
Health Event	1,227	12.0
Disaster and Accidents	1,174	11.5
Education Event	1,111	10.9
Total	**10,226**	**100.0**

A. Trigger Identification: This task identifies the minimal lexical span that signals an event. In a morphologically rich language like Nepali, triggers are frequently complex. Annotators were guided by the following criteria:

- Verbal Triggers: Identifying the minimal span capturing the action, including auxiliary verbs or compound forms e.g., नियुक्त भए (*was appointed*) when they constitute a single semantic unit.
- Nominalized Triggers: Accepting verbal nouns such as सम्झौता (*agreement*) or घोषणा (*announcement*) when they function as the primary action of the sentence.

B. Event Type Classification: Upon identifying a trigger, annotators assigned one of eight predefined event types based on the semantic context (Table 3).

Table 3: Event type classification schema used in annotation. "Disaster & Acc." denotes Disaster and Accidents.

Event Type	Description
Political	Governance-related occurrences such as appointments, resignations, or elections.
Economic	Business activities, trade agreements, and fiscal policy developments.
Sports	Matches, tournaments, victories, or athletic achievements.
Entertainment	Film premieres, festivals, award ceremonies, and artistic performances.
Health	Disease outbreaks, medical updates, or public health campaigns.
Disaster & Acc.	Natural disasters, fires, or transportation accidents.
Education	Academic results, institutional announcements, or education reforms.
Other	Clear events not covered by the above categories.

C. Special Cases: Annotators were provided with instructions for linguistic outliers. For Multiple Events in a single sentence, the primary trigger was chosen. For Idioms or indirect phrasing, annotators focused on clearly realized actions to avoid over-annotation. This methodical approach ensured the reliability and comprehensiveness of the NepEE corpus.

4 Data Analysis

This section provides a detailed analysis of the dataset.

Table 4: Fleiss' Kappa across annotation phases.

Phase	Annotators	κ_{Trig}	κ_{Type}
Pilot Phase	$\alpha_1, \alpha_2, \alpha_3$	0.621	0.764
	$\alpha_2, \alpha_3, \alpha_4$	0.645	0.781
	$\alpha_3, \alpha_4, \alpha_5$	0.638	0.775
Final Phase	$\alpha_1, \alpha_2, \alpha_3$	0.805	0.849
	$\alpha_2, \alpha_3, \alpha_4$	0.818	0.855
	$\alpha_3, \alpha_4, \alpha_5$	0.814	0.861
Aggregated	**Final avg.**	**0.812**	**0.855**

4.1 Inter-annotator Agreement

The reliability of a human-annotated semantic resource is fundamentally predicated on the degree of consensus achieved among independent judges (Fleiss, 1971; Falotico and Quatto, 2015). To quantify this metric for the NepEE corpus, we employed Fleiss' Kappa (κ) to measure agreement across our five native speakers for both the span-level and category-level tasks. Our final analysis yielded an overall agreement of $\kappa = 0.812$ for trigger identification and $\kappa = 0.855$ for event type classification.

A comparative analysis of the agreement coefficients across the longitudinal phases of the project (Table 4) underscores the efficacy of our iterative instruction refinement. The initial pilot phase revealed a moderate agreement for trigger spans, primarily due to the morphological ambiguity inherent in Devanagari verb-particle clusters. However, the subsequent instruction revision phase, which formalized the boundaries for split-predicates and nominalized action anchors, resulted in a substantial performance uplift. According to the interpretative benchmarks of Landis and Koch (1977), the terminal scores indicate near-perfect agreement, ensuring that the resulting dataset serves as a high-quality benchmark for assessing the semantic granularity of modern Nepali NLU systems.

4.2 Linguistic Analysis: Keywords and Topic Salience

To characterize the thematic and lexical properties of the NepEE corpus, we implemented a robust computational pipeline designed to address the morphological complexity and orthographic variations of the Nepali language. Our methodology integrates Unicode NFC normalization to ensure consistent character composition, Devanagari-specific regex filtering, and a custom-curated stopword removal process targeting high-frequency functional noise. Leveraging a class-based TF-IDF (c-TF-IDF) framework, a supervised adaptation of the procedure popularized by the BERTopic framework (Grootendorst, 2022), we quantified the salience of tokens across the eight event categories to identify the distinctive lexical features associated with each class.

The c-TF-IDF results (see Appendix for lexical distribution) demonstrate high semantic density and categorical distinctiveness. For instance, the *Economic Event* class is characterized by domain-specific anchors such as सेयर (*share*) and नेप्से (*NEPSE*), while *Disaster and Accidents* exhibits a strong correlation with vehicular and environmental lexemes like बस (*bus*) and पहिरो (*landslide*). The emergence of these highly relevant keywords validates the gold-standard manual annotations, confirming that the dataset successfully captures the underlying thematic distributions of the Nepali news corpus. It is also worth noting how domain-specific context influences trigger semantics. For example, in the *Education Event* category, विजयी (*victorious*) frequently acts as a trigger for student union elections or academic competitions, while सञ्चालन (*operation*) typically anchors events related to the opening or running of educational institutions.

Furthermore, we conducted a trigger word ambiguity analysis using a cross-tabulation matrix to visualize trigger-class overlap (Figure A2). This analysis quantifies trigger polysemy, the phenomenon where a single lexeme sparks different event types depending on the sentential context. A primary example of semantic overlap is observed in the trigger मृत्यु (*death*), which appears frequently in both *Disaster and Accidents* ($n = 85$) and *Health Event* ($n = 27$). Similarly, while the trigger सार्वजनिक (*public*) shows a strong bias toward *Entertainment Event* ($n = 77$), its distribution across multiple domains reflects its status as a high-utility functional anchor in Nepali media discourse. This dual methodology of keyword salience and trigger ambiguity analysis establishes a rigorous linguistic baseline, highlighting the challenges of contextual disambiguation inherent in automated Nepali EE.

5 Experimental Results and Analysis

To ensure a rigorous and reproducible evaluation, our experimental design treats supervised and generative paradigms differently. For all supervised models, including discriminative classical classi-

Table 5: Model performance for Trigger Identification. **Bold** indicates the best model and F1 score within a category; † indicates the overall highest performer.

Model	Precision	Recall	Macro F1
Classical Models			
CRF	0.6149	0.1916	0.2719
LogisticRegression	0.5569	0.2609	0.3400
RandomForest	0.6074	0.2104	0.2985
PassiveAggressive	0.4208	0.3394	**0.3506**
Transformer-based			
IndicBERTv2	0.7302	0.6908	0.7093
XLM-RoBERTa	0.7445	0.6748	0.7058
NepaliBERT	0.6927	0.6083	0.6419
MuRIL	0.3044	0.3333	0.3182
mBERT†	0.7331	0.6955	**0.7132**
LLM (Zero-shot prompting)			
Qwen2.5	0.3095	0.2495	**0.2651**
Gemma-3	0.2258	0.2128	0.2074
Phi-4	0.1494	0.4299	0.1929
Llama-3.1	0.1018	0.1959	0.1149
LLM (Few-shot prompting)			
Qwen2.5	0.3035	0.3156	**0.2889**
Gemma-3	0.2679	0.3157	0.2722
Phi-4	0.2307	0.4999	0.2878
Llama-3.1	0.1886	0.2131	0.1820

Table 6: Model performance for Event Type classification. **Bold** indicates the best model and F1 score within a category; † indicates the overall highest performer.

Model	Precision	Recall	Macro F1
Classical Models			
SVM	0.7998	0.7768	**0.7856**
RandomForest	0.7420	0.7267	0.7320
LogisticRegression	0.7864	0.7779	0.7813
MultinomialNB	0.7762	0.7664	0.7700
Transformer-based			
IndicBERTv2†	0.8520	0.8576	**0.8536**
XLM-RoBERTa	0.8276	0.8363	0.8298
NepaliBERT	0.8235	0.8262	0.8245
MuRIL	0.8002	0.8078	0.7923
mBERT	0.7915	0.7950	0.7925
LLM (Zero-shot prompting)			
Qwen2.5	0.7015	0.6891	0.6852
Gemma-3	0.7769	0.6701	**0.6977**
Phi-4	0.7353	0.6691	0.6898
Llama-3.1	0.7004	0.5111	0.5259
LLM (Few-shot prompting)			
Qwen2.5	0.7171	0.7124	0.7096
Gemma-3	0.7655	0.7474	**0.7538**
Phi-4	0.7318	0.6707	0.6848
Llama-3.1	0.7084	0.6574	0.6386

fiers and fine-tuned transformer encoders, we utilize a standardized 80/10/10 split (seed = 42) to train, validate, and test the models strictly on unseen data. Conversely, because the instruction-tuned Large Language Models (LLMs) underwent no parameter updates or fine-tuning, there is no risk of parameter leakage. Therefore, to obtain the most statistically robust and comprehensive measure of their zero-shot and few-shot inference capabilities, the LLMs were evaluated across the entire dataset of 10,226 sentences.

5.1 Experimental Methodology

Linguistic Pre-processing Engine To address the morphological richness and orthographic variations of Nepali, we developed a specialized linguistic processor. The engine performs NFKC Unicode normalization and handles zero-width joiners. A core component of our pipeline is an Orthographic Consonantal Root Extractor, which generates a structural representation by stripping all vowel signs (known as *matras* in Nepali) from the token. This method mitigates inflectional variance and was used as a fuzzy-matching fallback to align triggers during data labelling for classical models. Furthermore, we integrate a TnT POS Tagger (Brants, 2000) trained on the Nepali portion of the Indian languages corpus to provide shallow morpho-syntactic signals derived from POS tags and suffix heuristics for our discriminative baselines.

Discriminative Baselines For trigger identification, we framed the task as a token-wise classification problem using the BIO (Beginning, Inside, Outside) tagging formulation (Ramshaw and Marcus, 1995). We evaluated Conditional Random Fields (CRF) (Lafferty et al., 2001) alongside Passive-Aggressive (Crammer et al., 2006), Random Forest, and Logistic Regression classifiers. These models were vectorized via a contextual feature window (w_{i-1}, w_i, w_{i+1}) incorporating extracted consonantal roots and POS tags. For event type classification, we established lexical baselines using SVM (Cortes and Vapnik, 1995), Random Forest, Logistic Regression, and Multinomial Naive Bayes. All classical models were optimized through the Optuna framework (Akiba et al., 2019) across 60 trials per model.

Supervised Transformer Fine-tuning We fine-tuned five SOTA encoder architectures: IndicBERTv2 (IndicBERTv2-MLM-Sam-TLM),

MuRIL (muril-base-cased), XLM-RoBERTa (xlm-roberta-base), NepaliBERT, and mBERT (bert-base-multilingual-cased). Trigger identification was implemented via token classification layers using subword-to-character offset_mapping. Unlike the classical models, transformers for trigger identification used strict literal matching for trigger alignment to ensure character-perfect span offsets. All models were trained for 6 epochs with a learning rate of 2×10^{-5} and a batch size of 32 on Kaggle's T4 GPUs.

Generative LLM Evaluation We assessed the zero-shot and few-shot prompting ($k = 2$) capabilities of four leading models (see Appendix for prompt templates): Qwen-2.5 (Qwen2.5-7B-Instruct), Gemma-3 (gemma-3-4b-it), Phi-4 (Phi-4-mini-instruct), and Llama-3.1 (Llama-3.1-8B-Instruct). Because these models operate strictly in inference mode, they were evaluated across the full corpus to maximize statistical confidence. Performance for trigger identification was measured via token-level overlap F1 to accommodate the generative nature of the models. To ensure evaluation rigor, a heuristic-based label mapper normalized generative outputs to our predefined event categories. All LLM inference was performed on Modal.com using NVIDIA L4 GPUs, 16GB RAM and the vLLM engine (Kwon et al., 2023) for optimized throughput.

5.2 Performance Analysis and Insights

5.2.1 Task A: Trigger Identification

Span-level extraction (Table 5) proved substantially more complex than classification. Supervised mBERT achieved the highest F1-macro (0.7132). Within the classical paradigm, the token-wise Passive-Aggressive classifier ($F1 = 0.3506$) outperformed the CRF ($F1 = 0.2719$), indicating that high-dimensional local contextual signals are highly discriminative for Nepali triggers.

5.2.2 Task B: Event Type Classification

As summarized in Table 6, the supervised IndicBERTv2 model achieves the overall highest macro F1-score (0.8536). A significant finding is the robustness of optimized classical models; the SVM baseline ($F1 = 0.7856$) outperformed several zero-shot LLM configurations, suggesting that specialized morphological features effectively capture event-thematic distributions in the Nepali news domain. Among LLMs, Gemma-3 demonstrated the strongest few-shot performance ($F1 = 0.7538$).

5.3 Error Analysis on Trigger Identification

Trigger Identification for LLMs was evaluated using a strict token-level overlap metric. Under this framework, generative models heavily underperformed, ranging from 0.1149 (Llama-3.1-8B) to 0.2651 (Qwen2.5-7B) in zero-shot settings, peaking at 0.2889 (Qwen2.5-7B) with few-shot prompting. Because strict boundary-based metrics severely penalize boundary inflation alongside correct spans, they obscure the models' actual semantic comprehension. We established a classification taxonomy (Tables A2 and A3) to dissect these false negatives.

The primary driver of precision failure is contextual over-extraction. Instead of isolating single event triggers, models default to summarization, inflating token denominators. Given the event text "गत असार २४ को भोटेकोशीको बाढीले रसुवागढी बन्द भएपछि चीनसँगको आयात व्यापारको अन्तिम विकल्प तातोपानी नाका मात्र हो।" (*After the Bhotekoshi flood on Asar 24 closed Rasuwagadhi checkpoint, the only remaining option for import trade with China is the Tatopani border point.*) (truth: बाढीले) (*due to the flood*), Llama-3.1 extracts an entire argument summary: "भोटेकोशी बाढी रसुवागढी आयात व्यापार तातोपानी" (*Bhotekoshi River, flood, Rasuwagadhi checkpoint, import, trade, Tatopani border point*). This generative habit drives Llama's severe in-sentence span mismatch rate (46.3% zero-shot prompting; 65.8% few-shot prompting).

Furthermore, morphologically rich Nepali text induces systematic boundary misalignment. Models frequently capture inflectional suffixes and auxiliary verbs alongside the root trigger. For the nominal trigger वृद्धि (*increase*), Gemma-3 expands to वृद्धि भएको (*has increased*), while Phi-4-mini generates वृद्धि भएको देखिएको छ (*has been observed to have increased*). These mismatches account for 4.3% (Llama) to 12.7% (Phi-4-mini) of few-shot errors.

Ultimately, these low F1-scores are partly influenced by the strict boundary-based evaluation protocol. While exact-match metrics penalize span expansion, qualitative analysis indicates that generative models often identify semantically relevant event regions but fail to isolate minimal trigger spans. This suggests that span boundary precision,

rather than semantic localization, remains the primary challenge for generative models in Nepali trigger identification.

6 Conclusion

In this paper, we presented NepEE, a manually annotated Nepali EE dataset comprising 10,226 Devanagari sentences with span-level trigger annotations and eight event categories. The dataset establishes a comprehensive benchmark for trigger identification and event type classification in Nepali. Through systematic evaluation across classical machine learning models, transformer-based encoders, and instruction-tuned LLMs, we provide strong baselines and highlight key linguistic challenges including morphological variation, nominalized triggers, and predicate ambiguity. High inter-annotator agreement further supports the reliability of the annotations. The proposed dataset lays a critical foundation for structured IE in Nepali and aims to stimulate broader research on inclusive and multilingual NLP systems.

7 Limitations

Despite its contributions, this work has several limitations. The dataset is derived primarily from a single data source, which may not reflect informal or conversational Nepali and may limit cross-domain generalization. The annotation schema focuses solely on trigger spans and event-type labels and does not include argument roles such as participants, time, and location, which restricts full event-structure modeling. While NepEE currently focuses on these foundational steps, it serves as the first milestone in a broader roadmap. Future iterations of the corpus will extend the schema to include full argument role labeling (e.g., agents, locations, temporal markers) and event coreference, bringing Nepali NLU closer to comprehensive, ACE-style event extraction pipelines.

Furthermore, event boundaries in Nepali can be linguistically complex due to compounding, light verb constructions, and context-dependent trigger interpretation, which makes precise span identification inherently challenging. While careful guidelines were followed, certain edge cases require semantic judgment that may not always be uniformly resolved.

Third, our evaluation of LLMs for trigger identification relies on strict exact-match token overlap metric, which inherently penalizes generative models that produce semantically correct but morphologically misaligned spans (e.g., including auxiliary verbs). Future work should incorporate partial-match F1 metrics and explore advanced prompting strategies (such as Chain-of-Thought reasoning, schema-constrained generation, or Parameter-Efficient Fine-Tuning) to better harness generative capabilities for exact boundary extraction.

In addition, the benchmarks presented in this study are intended to establish competitive baselines rather than define upper performance limits. Continued progress may be achieved through larger-scale data collection, domain diversification, and exploration of more specialized architectures tailored to morphologically rich languages.

8 Ethical Considerations

The foundational sentences for this work originate from Sanjaal Corps' open-source Nepali collection distributed under the Apache-2.0 license by Sanjaal Corps. As the source material is openly licensed for research use, explicit individual consent was not required. Protecting the integrity of the Sanjaal Corps source material was a priority during our processing phase. The dataset does not introduce additional personally identifiable information beyond what is present in the original corpus.

For the annotation process, trained native Nepali speakers were engaged and annotators received a fair wage that matched current local pay scales for linguistic work. Annotators were provided with detailed guidelines and the workflow began with pilot rounds aimed at tightening inter-annotator consistency across complex categories. Given that certain sentences involve topics such as conflict, crime, or public health events, annotators were informed in advance about the nature of the content. Participation was voluntary, and annotators retained the right to withdraw from the task at any stage. Supervisory support was available to address concerns during the annotation process.

As with any curated corpus, the dataset may reflect biases present in the underlying source material. While high inter-annotator agreement supports the internal consistency of the annotations, it does not eliminate potential societal or distributional bias. Researchers are therefore encouraged to exercise caution when deploying models trained on this dataset in sensitive applications.

Finally, we encourage environmentally respon-

sible research practices. Efficient model training, transparent reporting of computational resources, and the use of carbon footprint estimation tools are recommended to reduce the environmental impact of large-scale experimentation.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Bal Krishna Bal. 2004. Structure of nepali grammar. *PAN Localization, Madan Puraskar Pustakalaya, Kathmandu, Nepal*, pages 332–396.

Yoshua Bengio. 2012. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36. JMLR Workshop and Conference Proceedings.

Thorsten Brants. 2000. Tnt–a statistical part-of-speech tagger. In *Sixth Applied Natural Language Processing Conference*, pages 224–231.

Ruirui Chen, Chengwei Qin, Weifeng Jiang, and Dongkyu Choi. 2024. Is a large language model a good annotator for event extraction? In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17772–17780.

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.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7(Mar):551–585.

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.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, Ralph M Weischedel, and 1 others. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8057–8077.

Rosa Falotico and Piero Quatto. 2015. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49(2):463–470.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Rajan Ghimire. 2022. NepaliBERT. https://huggingface.co/Rajan/NepaliBERT. Accessed: 2023-02-25.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 513–520.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ralph Grishman. 2019. Twenty-five years of information extraction. *Natural Language Engineering*, 25(6):677–692.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Ali Hürriyetoğlu, Surendrabikram Thapa, Hristo Tanev, and Surabhi Adhikari. 2025. Findings and insights from the 8th workshop on challenges and applications of automated extraction of socio-political events from text. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts*, pages 1–5.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul NC, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite:

Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 4948–4961.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Qian Li, Jianxin Li, Jiawei Sheng, Shiyao Cui, Jia Wu, Yiming Hei, Hao Peng, Shu Guo, Lihong Wang, Amin Beheshti, and 1 others. 2022. A survey on deep learning event extraction: Approaches and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 35(5):6301–6321.

Nobal Niraula and Jeevan Chapagain. 2022. Named entity recognition for nepali: data sets and algorithms. In *The International FLAIRS Conference Proceedings*, volume 35.

Jinu Nyachhyon, Mridul Sharma, Prajwal Thapa, and Bal Krishna Bal. 2025. Consolidating and developing benchmarking datasets for the nepali natural language understanding tasks. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1906–1925.

Kushal Paudyal. 2017. NepaliDataSets: Publicly released Nepali datasets of Sanjaal Corps. https://github.com/sanjaalcorps/NepaliDataSets. Accessed: 2025-04-17.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7654–7673.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third workshop on very large corpora*.

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.

Sovan Kumar Sahoo, Saumajit Saha, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A platform for event extraction in hindi. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2241–2250.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Surendrabikram Thapa, Kritesh Rauniyar, Shuvam Shiwakoti, Sweta Poudel, Usman Naseem, and Mehwish Nasim. 2023. Nehate: Large-scale annotated data shedding light on hate speech in nepali local election discourse. *Frontiers in Artificial Intelligence and Applications*, 372:2346–2353.

Sulav Timilsina, Milan Gautam, and Binod Bhattarai. 2022. Nepberta: Nepali language model trained in a large corpus. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 273–284.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *(No Title)*.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. Maven: A massive general domain event detection dataset. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1652–1671.

Wei Xiang and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

A Appendix

A.1 Prompts

Table A1 presents the lexical distribution of major event categories in the NepEE dataset, including the most frequent trigger words and salient

category-specific keywords ranked using c-TF-IDF. Figure A1 illustrates the zero-shot and few-shot prompting strategies employed for trigger identification and event classification tasks.

A.2 Trigger word ambiguity analysis

Trigger words in event extraction often exhibit varying degrees of semantic ambiguity, where the same lexical item may be associated with multiple event categories depending on contextual usage. To better characterize this challenge in the NepEE dataset, we analyze the distribution of frequently occurring trigger words across different event types. Figure A2 presents a heatmap illustrating the association between selected Nepali trigger words and event categories. The visualization reveals that while some triggers are strongly aligned with a single category, others appear across multiple event types, indicating substantial contextual overlap. For example, certain trigger words commonly associated with political or entertainment events also occur in educational or economic contexts. Such ambiguity highlights the need for context-aware modeling approaches that go beyond isolated trigger identification and incorporate broader semantic and syntactic cues.

A.3 Error analysis

Tables A2 and A3 present the distribution of generative failure modes for zero-shot and few-shot trigger identification, respectively. Across most models, the dominant source of error is *In-Sentence Mismatch*, where the predicted trigger originates from the input sentence but does not correspond to the annotated gold trigger. This suggests that models are often capable of identifying event-relevant lexical spans, yet struggle to precisely localize the correct trigger expression.

Hallucination errors are particularly prominent for Llama-3.1-8B and Phi-4-mini in the zero-shot setting, indicating a tendency to generate unsupported or fabricated trigger words. In contrast, Qwen2.5-7B demonstrates comparatively lower hallucination rates and achieves the highest exact-match performance among the evaluated models. Few-shot prompting generally reduces hallucination and under-extraction errors, especially for Llama-3.1-8B, but also increases morphological mismatch rates in several cases, suggesting that demonstrations encourage semantically related but morphologically inconsistent outputs.

Additionally, Phi-4-mini exhibits substantially higher over-extraction behavior across both settings, frequently generating longer phrases or multiple tokens instead of concise trigger spans. Overall, the findings highlight that trigger identification in Nepali remains challenging not only because of semantic ambiguity, but also due to morphological variation and the generative tendencies of LLMs. This behavior may also reflect inconsistencies between semantically plausible trigger expressions and the single annotated trigger span present in the dataset. In several cases, models generate alternative lexical forms that are contextually appropriate but differ from the gold annotation due to synonym usage or inflectional variation. These findings suggest that future Nepali event extraction systems may benefit from more flexible evaluation schemes and annotation strategies.

Table A1: Lexical distribution and keyword salience across event categories. Trigger frequency (Freq) is shown in parentheses; keywords are ranked by c-TF-IDF.

Event	Top Triggers (Freq)	Significant Keywords (Ranked by salience)
Disaster & Acc.	मृत्यु (*death*, 85), दुर्घटना (*accident*, 71)	मृत्यु (*death*), दुर्घटना (*accident*), बस (*bus*), बाढी (*flood*), पहिरो (*landslide*), नदी (*river*)
Political	छलफल (*discussion*, 36), निर्णय (*decision*, 29)	निर्वाचन (*election*), पार्टी (*party*), संविधान (*constitution*), निर्वाचित (*elected*), मत (*vote*), सरकार (*government*)
Economic	लगानी (*investment*, 22), खर्च (*expense*, 20)	सेयर (*share*), लगानी (*investment*), बैंक (*bank*), कारोबार (*transaction*), मूल्य (*price*), भुक्तानी (*payment*), नेप्से (*NEPSE*)
Sports	पराजित (*defeated*, 48), गोल (*goal*, 27)	रन (*run*), खेल (*game*), विकेट (*wicket*), गोल (*goal*), क्रिकेट (*cricket*), लिग (*league*)
Entertainment	सार्वजनिक (*released*, 77), प्रदर्शन (*screening*, 21)	चलचित्र (*movie*), गीत (*song*), फिल्म (*film*), नाटक (*drama*), अवार्ड (*award*), संगीत (*music*)
Health	उपचार (*treatment*, 28), मृत्यु (*death*, 27)	रोग (*disease*), क्यान्सर (*cancer*), औषधि (*medicine*), अस्पताल (*hospital*), संक्रमण (*infection*), खोप (*vaccine*)
Education	विजयी (*victorious*, 14), सञ्चालन (*operation*, 11)	परीक्षा (*exam*), भर्ना (*admission*), विद्यालय (*school*), विश्वविद्यालय (*university*), शैक्षिक (*academic*)

Prompt: As a domain expert and native Nepali annotator, extract the event trigger word(s) from the text. A trigger is the specific word or phrase indicating that an event has occurred. Return only the trigger word(s).

Prompt: As a domain expert and native Nepali annotator, classify the sentence into one of these categories:
Political Event, Other Event, Economic Event, Sports Event, Entertainment Event, Health Event, Disaster and Accidents, Education Event.
Return only one category name.

Prompt: As a domain expert and native Nepali annotator, extract the event trigger word(s) from the text. A trigger is the specific word or phrase indicating that an event has occurred. Return only the Nepali trigger word(s).

P1	P2
Sentence: काठमाडौंमा ट्रक ठोकिए। *Trucks collided in Kathmandu.* **Answer:** ठोकिए (*collided*)	**Sentence:** नेपाल र भारतबीच व्यापार सम्झौता भयो। *A trade agreement was made between Nepal and India.* **Answer:** सम्झौता भयो (*An agreement was made.*)

Prompt: As a domain expert and native Nepali annotator, classify the sentence into one of these categories:
Political Event, Other Event, Economic Event, Sports Event, Entertainment Event, Health Event, Disaster and Accidents, Education Event.
Return only one category name.

P1	P2
Sentence: बाढीले गाउँ बगायो। *The village was swept away by the flood.* **Answer:** Disaster and Accidents	**Sentence:** नेपाल र भारतबीच व्यापार सम्झौता भयो। *A trade agreement was made between Nepal and India.* **Answer:** Economic Event

Figure A1: Zero-shot and few-shot prompts used for evaluation.

Table A2: Generative failure mode distribution for zero-shot trigger identification. Values represent the percentage of total dataset predictions ($N = 10,226$) falling into each taxonomy category.

Model	Exact Match	Over-extract	Under-extract	Morphological Mismatch	In-Sentence Mismatch	Hallucination	Abstention
Qwen2.5-7B	16.3%	0.5%	15.2%	2.0%	64.5%	1.3%	0.1%
Gemma-3-4B	12.2%	1.5%	9.1%	4.8%	61.5%	10.5%	0.5%
Llama-3.1-8B	2.8%	10.6%	3.6%	2.4%	46.3%	34.3%	0.0%
Phi-4-mini	3.8%	31.7%	1.7%	5.6%	21.1%	30.1%	6.1%

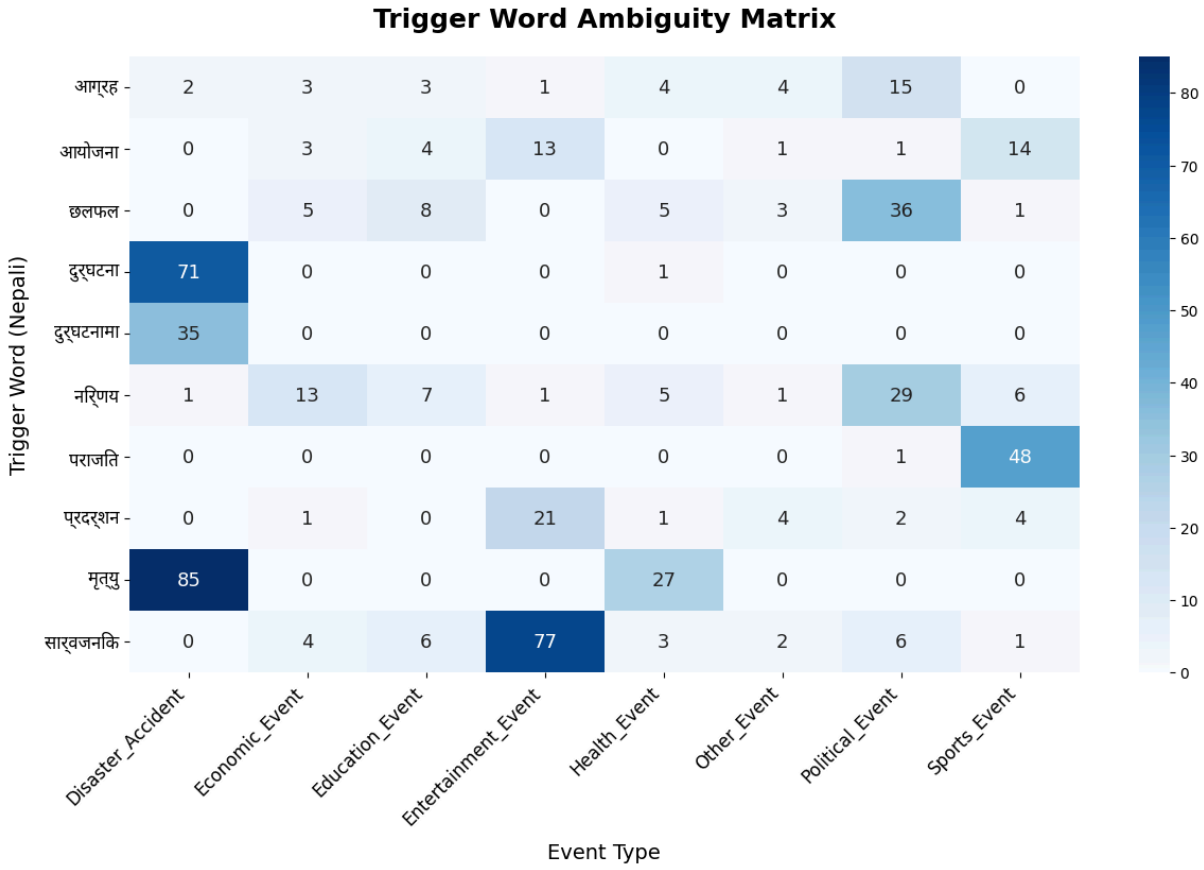


Figure A2: Heatmap of trigger-word ambiguity across event types in the NepEE dataset. Due to font rendering limitations during figure generation, some Nepali Unicode characters may appear slightly distorted in the visualization.

Table A3: Generative failure mode distribution for few-shot trigger identification. Values represent the percentage of total dataset predictions ($N = 10,226$) falling into each taxonomy category.

Model	Exact Match	Over-extract	Under-extract	Morphological Mismatch	In-Sentence Mismatch	Hallucination	Abstention
Qwen2.5-7B	14.5%	3.1%	10.1%	9.0%	61.1%	2.0%	0.1%
Gemma-3-4B	12.5%	4.4%	6.1%	11.8%	51.9%	13.3%	0.1%
Llama-3.1-8B	9.6%	4.4%	6.3%	4.3%	65.8%	9.6%	0.0%
Phi-4-mini	6.3%	30.4%	1.3%	12.7%	26.9%	19.5%	3.0%

A Self-Reflective LLM-based Architecture for Semi-Open Event Extraction

Hristo Tanev , Michel De Bollivier , Bertrand De Longueville
Joint Research Centre, European Commission
{hristo.tanev, michel.de-bollivier, bertrand.de-longueville}@ec.europa.eu

Abstract

We present a multi-agent *reflective* architecture for event extraction based on generative large language models (LLMs). Our architecture is the first of its kind to perform **Semi-Open Event Extraction (SOEE)**, a hybrid framework that combines a fixed set of event-template fields with dynamically generated attributes produced using self-reflection. A further contribution of the system is its operationalization of reflection as an internal question-generation and answering process. It is defined as the generation of questions about missing or implicit event information and finding their answers within the system itself. We model event extraction as an iterative dialogue between a **reflective** LLM-based agent, which generates questions to uncover missing event information and a set of **expert** agents, which provide domain-aware answers to these questions. The expert agents also generate the initial event template using a generative LLM. Across all evaluation experiments on articles from the health domain, **MAREA** demonstrates strong core-field extraction and effective reflective template expansion, with its three question-generation strategies producing useful additional event attributes. The observed errors were mainly due to imprecise prompt interpretation or inaccurate interpretation of the source article by the LLM, rather than to hallucination or an intrinsic inability to retrieve the requested information.

1 Introduction

LLMs are increasingly adopted within the event extraction community (Li et al., 2025), where recent work demonstrates that prompting and generative approaches can produce structured outputs in zero- and few-shot settings, creating new prompt-driven extraction paradigms. Moreover, the successful application of LLM in NLP has motivated approaches that move beyond single-pass prompting toward more iterative and collaborative forms of processing.

The integration of LLMs into multi-agent systems (MAS) represents a further evolution in addressing complex NLP tasks, including event extraction (Zhang et al., 2026), (Guo et al., 2026), (Wang and Huang, 2024), question answering (Zong et al., 2024), summarization (Kim and Kim, 2025), (Celikyilmaz et al., 2018), fact checking (Lin et al., 2025), and scientific text generation. Within these architectures, multiple agents—each associated with specialized reasoning roles—collaborate through the exchange of intermediate representations, hypothesis generation and evaluation, and other structured steps in the analytical process. Recent studies, such as (Wang and Huang, 2024), suggest that such distributed reasoning mechanisms enable multi-agent systems to achieve deeper and more robust analyses compared to single-agent LLM models.

In our work, we also address **self-reflection** as a specific type of collaborative reasoning inside a MAS. During such reasoning, a system generates questions about missing or implicit information (in our case event-related) and answers them within the MAS itself. This system aspect is related to a growing body of work on formulating event extraction as question answering (Lu et al., 2025, 2023; Hong and Liu, 2024).

In this paper, we propose a multi-agent architecture specifically designed for **Semi-Open Event Extraction (SOEE)**.

In SOEE, a sub-set of event fields — such as event type, date, and location, specified by the system user — remains fixed in order to preserve cross-domain comparability. The remaining fields are defined dynamically at runtime, depending on the content of the article, the nature and type of the event, and other event-specific attributes that may be important but are not covered by the predefined schema. In our architecture, these flexible

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 72–82

attributes are inferred through a reflective reasoning process in which the LLM generates questions and then answers them. This process is implemented as an iterative dialogue: an initial event template is first constructed and then progressively refined and expanded through question generation and answering.

Our method is organized as a three-layer architecture: (1) an expert layer, which is responsible for constructing the initial event representation and answering the questions generated during reflection; (2) a reflective layer, which is responsible for generating natural-language questions aimed at identifying missing event information and expanding or refining the current event template; (3) a coordination layer, usually containing one manager agent who coordinates the activities of the other agents.

To demonstrate the feasibility of our proposed method, we developed a prototype called **MAREA** , **M**ulti-**A**gent **R**eflective architecture for **E**vent **A**nalysis.

The remainder of this paper is organized as follows. Section 2 introduces the concept of SOEE. Section 3 reviews related research on multi-agent reasoning, LLM-based, and open event extraction. Section 4 presents the design of the **MAREA** system, detailing the interactions between the coordination, reflective, and expert layers. Section 5 describes the evaluation of **MAREA** on a dataset of health-related news reports and discusses the observed improvements in template completeness and attribute discovery. Section 6 discusses the limitations of our approach. Section 7 presents an overview of the achieved results and future directions.

2 Semi-Open Event Extraction (SOEE)

A central challenge in event extraction is to balance the structural consistency of event templates with the richness of the information they encode.Traditional closed-domain extraction systems rely on ontologies and taxonomies, such as ACE or ERE (Song et al., 2015) and event templates with fixed structures such as the ones presented in (Grishman et al., 2002). Such an approach ensures consistency in the event template fields, but limits generalization to unseen event types. In contrast, open event extraction (OEE) approaches (Deng et al., 2022) offer domain flexibility but may potentially generate heterogeneous or inconsistent representations. To address this trade-off, we introduce the concept of SOEE, a hybrid form of event extraction that combines a fixed core schema with dynamically extensible attributes.

In SOEE, a sub-set of event fields — such as `event_type`, `date`, `location`, and several others — remains fixed to preserve cross-domain comparability, while other fields are contextually defined by the system at runtime. These flexible attributes are inferred through a reflective reasoning process, introduced in the previous section.

3 Related work

The research most closely related to our work falls into two main research lines: (1) the application of LLMs, agents, and multi-agent systems to event extraction; and (2) approaches to Open Event Extraction.

Recently, LLMs have proven to be effective tools for event detection and extraction (Meng et al., 2024; Tanev et al., 2025b; Wang et al., 2025). Beyond single-LLM approaches, recent work has explored multi-agent LLM architectures for event extraction to improve extraction quality. These include debate-based event template refinement (Wang and Huang, 2024), multi-agent generation-and-extraction for document-level event argument extraction (Zhang et al., 2026), and programming-style agent decomposition for zero-shot event extraction (Guo et al., 2026). All the works mentioned so far, however, assume a fixed event template. Our approach, on the other hand, is the first to experiment with a semi-open event schema, using a multi-agent LLM architecture.

Another line of research relevant to our work is Open Event Extraction (OEE), which does not assume a fixed event template or pre-defined event types. Instead, it typically extracts less structured event information from text without targeting specific event classes. The extracted information is often represented as tuples, such as (time, location, keyword set), or pairs, such as an event name and a date. OEE systems may also extract event triggers aimed at broad classes of events (Tong et al., 2020). Works such as (Deng et al., 2022) and (Wang et al., 2019) show that OEE can provide effective solutions for event annotation and extraction. In our implementation of SOEE, similarly to OEE, we enrich the event template with new, context-dependent event arguments, however our approach relies on a predefined core schema, which

increases its usefulness. Many event extraction related applications, like automatic event database filling, assume the presence of standard arguments and a fixed core event structure.

4 System Architecture

As already explained in the introduction, the **MAREA** event extraction architecture is made up of three layers: Expert, Reflection, and Coordination. The input of the event extraction process is a news article, while the output is a set of event templates (Aone and Ramos-Santacruz, 2000), structurally describing the events from the article.

- **Expert Layer**: This layer consists of two types of agents responsible for constructing the initial event representation and answering the questions generated during reflection:
    1. **A Template-Proposing Agent** generates the initial event template. It formulates the LLM prompts that specify the extraction task, define the initial template structure, and, when needed, provide few-shot examples consisting of source texts and their corresponding event annotations.
    2. **Answering Agent** responds to the questions generated by the Reflective Layer. It formulates prompts that guide the LLM toward locating and extracting the requested information from the source text. This agent also manages the interaction with the LLM and, in some cases, it may resolve specific sub-tasks without relying on the LLM.
- **Reflective Layer**: This layer is responsible for expanding the event template with contextually relevant out-of-schema fields and fill in their values. It is the core component that enables the semi-open character of the architecture, allowing the system to move beyond a fixed schema and introduce new, context-dependent event attributes. In MAREA, this layer contains two agents:
    1. A generic question formulating agenh uses thr,e different strategies, presented in 4.1, to generate natural-language questions aimed at identifying missing event information and expanding or refining the event template, created by the template-proposing agent.
    2. A spatial reasoning agent, which identifies the location names inside the input article using spaCy (Vasiliev, 2020) and then asks the LLM to identify the semantic role of each location.
- **Coordinating layer with a manager agent**: The manager agent in this layer coordinates the activities of the other agents, controls the execution flow of the extraction process, and ensures effective interaction with the user or with the software system in which the SOEE module is embedded.

The basic sequence of text processing stages is outlined below.

1. A news article is passed to the manager agent which forwards it to the template-proposing and to the reflective agents.
2. The template-proposing agent creates an initial set of event templates, one for each event within the article, by prompting the LLM using few-shot learning settings:
   The templates are created by prompting the LLM to produce event templates with a structure defined in the prompt. Several input and output examples are also provided in it. This prompt has the following structure:
   "You are an information extraction assistant, specialized in health-related events, such as [event_type_list]. Given the following news article, extract all distinct health-related events mentioned: [news_article]. Consider as an example this news article as sample input [sample_news_article]! Consider the following extracted event templates as a sample output: [sample_event_templates_in_JSON]".
3. After the initial templates are generated, they are passed to the reflective agent layer:
    (a) The question formulating agent generates a set of questions, e.g. "What measures are being taken to prevent the spread of COVID-19?", aimed at finding new event attributes and their values.
    (b) The spatial reasoning agent discovers the places mentioned inside the news article and their semantic functions for each extracted event ("place of infection", "event place", "hospital location", "responding

authority location", etc.) Each semantic role for a location constitutes a new candidate field for the event template.

4. The answering agent forwards the questions generated by the reflective agents to the LLM, using prompts that instruct the model to provide concise and precise answers. Then, it parses the LLM answer into a JSON structure.

In summary, **MAREA** first generates an initial event template through few-shot learning, followed by a reflection phase in which clarifying questions are formulated to identify potential event attributes beyond the predefined schema, thereby improving template completeness.

4.1 Question formulation strategies

Our system uses three question formulation or *reflective* strategies, that are implemented in the question formulation agent from the reflective layer:

1. Mapping event text to a predefined set of questions: We have trained a BERT model, given a sentence to map it to a sub-set of 29 frequent questions and corresponding fields from the domain of health events. Example of such a questions-field pair is ("Where did the infection occur?", "infection-place"). Due to lack of space, we cannot give the details of the BERT model training here. We used this BERT model to map all the sentences from the LLM-generated event summary to question-field sets and then we group them into one question set for each event.

2. Generic prompt-based question generation: We created a few-shot prompt for asking the LLM itself to suggest questions for expanding the event template, see Table 4. In this prompt, we give as parameters the article text and the event. The prompt asks the LLM to suggest questions which can discover new information and new event fields. The few-shot prompt itself was optimized using the MIPROv2 prompt optimization algorithm (Opsahl-Ong et al., 2024), implemented in the DSPy prompt programming and optimization framework. We used a small training set containing 5 tuples having the structure (article; event; relevant additional question set) to run the MIPROv2 otimizer with the **LLaMA-3.1-70b-instruct** model.

3. Keyword-based question generation: In this strategy, we first prompt the LLM to identify a set of keywords and key phrases for the event. Then, for each keyword or phrase, we prompt the LLM using the question: "How does keyword/phrase relate to the event?" The prompt also asks the LLM to propose a new template field name to accommodate the answer.

4.2 Application of MAREA to Health-Related Event Extraction

Within the **MAREA** architecture, only the template-proposing agent in the Expert layer is specialized for the health domain. Its domain knowledge comprises definitions of the core event fields and a sample news article annotated with two example events, used to support few-shot prompting. All remaining agents—including the question formulation and the spatial-reasoning components—are intentionally designed to remain domain-independent, enabling reuse across domains without architectural modification.

4.2.1 Event Template

As we mentioned, we adopt a SOEE strategy, in which a sub-set of the event template fields is fixed, following established approaches in health-related event extraction (Piskorski et al., 2023; Linge et al., 2012). This design preserves structural consistency while allowing the system to dynamically extend the template in the reflective reasoning phase.

For our experiments, we define a set of health-related event types grounded in prior research (Tanev et al., 2025b; Piskorski et al., 2023):

Outbreak: Sudden rise in disease cases; *Product_recall*: Withdrawal of unsafe medical or food products; *Study*: Research activities related to diseases, treatments, or vaccines; *Drug_approval*: Regulatory approval of drugs or vaccines; *Health_policy_change*: Changes to health-related laws or guidelines; *Disease_statistics_update*: Updated incidence or mortality data; *Biological_threat*: Bioterrorism incidents or emerging high-risk pathogens; *Pandemic_response*: Measures addressing large-scale epidemics; *Other_health*: Miscellaneous health-related events.

The fixed portion of the event template consists of the core fields shown in Table 1.

These event types and fixed template fields are explicitly provided to the LLM during prompting and serve as a structural basis for the initial event

Field name	Description
event_type	Category of the health-related event (e.g., outbreak, policy change, biological threat).
actors	People, organizations, or groups directly involved in or affected by the event.
description	Concise natural-language summary of the event.
event_text	Longer textual description preserving the main details and context of the event.
disease	Name of the disease or health condition associated with the event, if applicable.
biological_agent	Pathogen or biological agent responsible for the disease (e.g., virus, bacterium).
symptoms	Reported symptoms associated with the disease or health event.
where_place	Specific place or locality where the event occurred or was reported.
where_country	Country in which the event took place.
when_start	Date or time period marking the beginning of the event.
when_end	Date or time period marking the end of the event, if specified.
number_of_cases	Reported number of confirmed or suspected cases associated with the event.
number_of_deaths	Reported number of fatalities related to the event.
main_reason	Primary cause, trigger, or motivating factor underlying the event.
measures_taken	Actions, interventions, or policies implemented in response to the event.

Table 1: Core fields of the semi-open health event template and their semantics.

template generation.

5 Experiments and Evaluation

To evaluate the proposed approach, we compiled a corpus of health-related news articles from three sources: The first sub-setconsists of 300 articles collected from the *Europe Media Monitor* (Steinberger et al., 2013) over several months in 2020. The second sub-setincludes approximately 120 articles gathered from the Fox News Health RSS feed [1] over several weeks in September 2025. The third source is a random sample of 60 articles from March 2026, retrieved from the Medical Express RSS (section Infection Diseases) [2]. All articles were processed using **MAREA** , backed by the **LLaMA 3.1–70B–Instruct** large language model.

For the final evaluation, we selected 65 articles in random from the 451 articles corpus. **MAREA** extracted 70 events from these 65 articles.

[1] https://moxie.foxnews.com/google-publisher/health.xml

[2] https://medicalxpress.com/rss-feed/breaking/infectious-diseases-news

One expert evaluator has manually inspected the output of the **MAREA** system. The annotator has labeled all extracted field values as correct or not and additionally has indicated the missing values from each extracted event template. Additionally, the evaluator has searched for events which were mentioned but not extracted by the system.

5.1 Event detection

The evaluation showed that event detection achieved 100% precision and recall on the 65-article test set. All 70 events extracted by**MAREA**were judged to be relevant, and the expert annotator did not identify any additional events that were missed by the system. Event identity was determined by considering the event summary generated by the LLM, encoded in the ‘description‘ field, together with the other event attributes. Under this matching criterion, all system-generated events corresponded to manually identified events, yielding perfect event-level detection performance on this dataset.

5.2 Core fields extraction

Table 5 shows the precision, recall and F1 score for the accuracy of the core fields extraction, that is the fields predefined by the health event schema, defined in Table 1.

Performance (F1) is very high (over 0.85) for important event fields such as *event type*, *actors*, *disease*, *where-country*, and *measures taken*. The dataset did not contain enough data for evaluating the *number of deaths* field, since most of the articles were about clinical studies and disease statistics update, reporting only number of new cases. Fatalities were reported in only one of the 65 articles from the test set. Another significant field, *number of cases* was evaluated excluding the events of type STUDY for which the number of cases was not relevant. Instead, articles about these events typically report annual disease rates for the studied diseases, which in this case is ambiguous. Therefore, in order to ensure clarity in the evaluation process and to avoid ambiguity, we excluded the events classified as STUDY from our evaluation.

Notably, *event type* obtained a very high accuracy of 90%, demonstrating that our approach our approach achieves very strong performance in event classification.

Performance is lower, but remains satisfactory, for key spatial and temporal attributes: *where-country* and the event starting date in the *where-start* field. The *where-place* field for which the**MAREA**system has a relatively low performance, was intended to store the populated place name. Although the LLaMA model was instructed to do so, it has extracted instead locations such as names of universities and in some cases even countries. Even if these values can formally be considered to be partially correct, their level of granularity was different from the required one. Consequently, we measured low levels of precision and recall for this field. Another problematic field was found to be *when end* field, containing the end date of the event. The recall of extracting this attribute was notably low, as well as the precision. We have inspected the errors and found that this information is not always obviously stated and may even require temporal inference.

Altogether, the errors in our core fields extraction stem from incorrect interpretation of the prompt by the LLaMA model. For example, the **where place** mistakes are places which are correct as event locations, but do not correspond to the required geographic granularity. Similarly, the **biological agent** field often captured organisms, which were not disease vectors, as required by the prompt.

Taken together, these results indicate that MAREA, when powered by **LLaMA-3.1-70B-Instruct**, holds considerable promise for event extraction from real-world news data. Although the system shows lower performance on some core event attributes, largely due to incorrect prompt interpretation, our analysis suggests that improvements in post-processing and prompt design could substantially enhance overall extraction performance. It is also noteworthy that the event attribute extraction accuracy is comparable to that reported in previous work on a dataset from the same genre, namely medical news (Tanev et al., 2025a). Although the test corpora differ, our achieved accuracy is considerably higher than the baseline performance reported in that study. Thus, our work further supports the conclusion of the aforementioned study that LLMs can reliably perform event extraction and classification.

5.3 Additional fields added to the template

In order to assess the reflective capacity of our architecture and the question- and field-formulating strategies, described in section 4.1, we have randomly selected **31** event templates from 31 different articles from the test set. For these templates the reflective agent layer (both the question formulating and the spatial reasoning agents) has generated **147** additional event fields in total.

We have classified these new fields and their values into 4 relevance categories: (a) **Irrelevant**: Irrelevant information or incorrect field names or values; (b) **Low relevance**: formally correct field name and value; information is new, but it is not strongly connected to the event; (c) **Medium relevance**: relevant field name and a correct value; the field-value pair brings new information, but field name could be improved; (d) **High relevance**: relevant field name and a correct value, bringing new information, relevant to the event; (e) **Duplicate**: the field and value are correct, but they duplicate information already present in the template. Table 3 represents the distribution of the 147 event attribute-value pairs across these relevance categories. It is important also to note that the values of the fields are correct except few, classified as **Irrelevant**. Considering this, the evaluation of the attribute (field name) - value pairs is driven by the field name relevance.

It is noteworthy that 56% of the new generated fields have high or medium relevance (the last two rows in the Table 3). Highly relevant fields are completely correct as name and value, highly relevant to the event, and contain new information w.r.t. the other part of the template. These fields are 36% of all additionally-generated event fields. They are complemented by the medium-relevance ones, which are still correct and introduce relevant and innovative information, although their names could be improved.

If we exclude the 24% duplicates from the evaluation, the percentage of the medium and high relevance template fields becomes **77%**. Some interesting template-expanding questions and suggested new fields, generated by the question formulation agent are presented in Table 4.

5.4 Evaluation of the BERT question-generation module

The BERT-based question-proposing module suggests template-expansion questions in parallel with the prompt and keyword -based question-generation strategies. The module relies on a BERT model that maps each sentence to a set of 29 predefined questions, as described in Section 4.1.

We evaluated partially the accuracy of the BERT question-proposing module as follows: A proposed question and field was considered correct, if its answer (field value) could be found in the corresponding article; otherwise, it was considered incorrect. For this evaluation, we selected 68 articles from our corpus of 451 news articles. These articles were not used in the other evaluation experiments and all concerned infectious-disease outbreaks. This topic was chosen because outbreak-related articles typically contain information relevant to many of the 29 predefined questions used by the BERT module.

We ran the**MAREA**system on this dataset, extracting 68 events, one from each article, and recorded the questions suggested by the BERT module. Then, for each event–article pair, we manually identified which of the 29 predefined questions are answerable from the article text. Finally, we measured the overlap between the correctly suggested questions and the manually identified answerable questions. This allowed us to compute precision, recall, and F1-score for each event, as well as the overall micro-averaged precision, recall, and F1-score.

Table 2 summarizes the results. The precision, namely the proportion of answerable questions among all questions generated by BERT, is above 0.70. However, recall, which measures how well the module covers all relevant answerable questions, is considerably lower, at approximately 0.43.

The fact that more than half of the relevant questions are missed is compensated for by the other question-generation strategies. At the same time, the macro precision of 0.75 suggests that, when the module does propose a question, it is often relevant and answerable, providing evidence for the usefulness of this question-proposing strategy as a complementary component.

5.5 Missing information

Since the recall is not addressed in our evaluation, due to lack of annotated data, we have asked one evaluator to estimate how many facts are missing from each template from a sub-set of randomly selected **20 templates**. The evaluator considered the information inside each template and the corresponding article. He counted the number of missing facts from each template, where each fact could be positioned in one event field. Out of 20 event templates, the evaluator has found **8 missing facts**, which shows that on average our current implementation of **MAREA** in the health domain has a probability of missing a fact of **0.4**. In our test settings, this number can be plausibly interpreted as "in 40% of the cases our system may miss one important fact". This means, however, that most of the information is successfully captured in the output template, since the generated event templates in this test set have approximately 13 fields on average.

5.6 Evaluation overview

Overall,**MAREA**demonstrates strong core-field extraction and useful semi-open template expansion through reflective question generation, while the main remaining challenges concern prompt interpretation, field-name normalization, and incomplete question coverage.

6 Limitations

This study has several limitations. First, although**MAREA**performs well on many core event fields, some errors result from prompt misinterpretation by the underlying LLM, especially for fields requiring fine-grained spatial and temporal interpretation. Second, the dynamically generated fields are not always equally informative: some

Averaging	TP	FP	FN	Precision	Recall	F1
Macro	–	–	–	0.7468	0.4364	0.4713
Micro	283	110	363	0.7201	0.4381	0.5448

Table 2: Evaluation results for BERT suggested questions

Macro scores are computed as averages over records. Undefined per-record F1 scores were treated as 0. Micro scores are computed from pooled TP, FP, and FN across all records.

are duplicate, low-relevance, or require better field name. Finally, some evaluations were conducted on small manually inspected samples, and all evaluation experiments were carried out by a single annotator, which introduces a subjective bias and prevents us from estimating inter-annotator agreement. Broader testing across domains, event types, LLM models, and multiple annotators is therefore needed.

7 Conclusions

In this paper, we introduced **MAREA** , a reflective multi-agent architecture designed to perform SOEE. The proposed system models event extraction as a combination of an LLM-based few-shot learning template generation and an internal question–answer process in which specialized agents collaborate to enrich event representations. By combining these two approaches, **MAREA** is an attempt to address the trade-off between structural consistency and informational completeness in event extraction.

The experimental results in the health domain show that **MAREA** achieves consistently good performance across a wide range of core event fields, including event type, actors, disease information, countries, and public health measures. At the same time, the reflective agent successfully identifies and populates additional fields that provide complementary semantic context. The majority of the newly introduced fields are relevant and significant for the event template completeness. The probability to miss a fact for this approach was found to be 0.4, still this experiment was done on a very small scale. This number, however means that this approach tends to capture the larger part of the information in the generated event templates.

Future work will focus on three main directions. First, we plan to improve both the prompting strategy and the post-processing of generated templates. We will further optimize the prompts used by the template-proposing and answering agents, possibly using automatic prompt optimization methods such as MIPROv2. This may reduce errors caused by prompt misinterpretation. We also plan to introduce post-processing procedures for normalizing and filtering generated templates. These may include merging information from semantically similar fields, removing duplicate or low-informative attributes, and applying NLP-based filters to enforce field-specific constraints. For example, for the ‘where-place‘ field, named-entity recognition could be used to retain only entities identified as locations.

Second, we plan to conduct more extensive evaluations of the SOEE approach. This includes testing on benchmark data, when available for health-related news, providing second annotator evaluations, and experimenting with different LLM backends, such as various GPT models, open-weight models, and smaller locally installable LLMs. We also plan to carry out qualitative analyses of the system's behavior, including checking whether similar events across different articles, or different runs on the same article, lead to consistent outputs. In addition, we will investigate the causes of errors, such as prompt misinterpretation, inaccurate text interpretation, or possible LLM hallucinations, with particular attention to the reflective layer.

Third, we plan to exploit the most accurately extracted fields for the creation of synthetic or silver-standard event annotations. Fields such as ‘event type‘, ‘description‘, ‘actors‘, ‘where-country‘, ‘disease‘, ‘measures taken‘, and others which achieved high extraction accuracy in our evaluation, could be used as reliable anchors for automatically constructing larger weakly supervised corpora. Such silver-standard datasets could support further training, prompt optimization, and evaluation of event extraction systems in domains where manually annotated data are scarce.

Finally, producing new versions of **MAREA** for domains like security or disaster relief is an exciting research direction.

Field correctness and informativeness	Relative quantity
Irrelevant	0.05
Low	0.12
Medium	0.20
High	0.36
Duplicate	0.24

Table 3: Relevance of the fields added by reflection layer

Question	Suggested Field
What are the factors that could contribute to outbreak resurgence?	resurgence_factors
What is the expected timeline for implementing the new economic stimulus efforts?	implementation_timeline
What is the trend of new cases?	case_trend
What is the current severity of the outbreak?	outbreak_severity
Where did Robert O'Brien contract the viral infection?	infection_location
Who is criticizing the government's handling of the pandemic?	criticism_source
What journal published the findings of the study?	publication_venue
How does "flu Vaccine" relates to the event?	vaccine_type
What is the purpose of the early testing of the new vaccine candidate?	vaccine_purpose

Table 4: Questions and suggested template fields by reflection agent.

Field	Precision	Recall	F1-score
event_type	0.901	0.901	0.901
description	0.933	0.933	0.933
event_text	0.796	0.796	0.796
disease	0.922	0.870	0.895
biological-agent	0.600	0.600	0.600
symptoms	0.789	0.714	0.750
main-reason	0.864	0.704	0.776
actors	0.931	0.931	0.931
where-place	0.538	0.583	0.560
where-country	1.000	0.972	0.986
when-start	0.833	0.714	0.769
when-end	0.500	0.429	0.462
number_of_cases*	0.714	0.714	0.714
measures taken	0.960	0.923	0.941

Table 5: Core fields extraction accuracy.
*Computed after excluding STUDY events, for which number_of_cases is not applicable.

References

Chinatsu Aone and Mila Ramos-Santacruz. 2000. Rees: a large-scale relation and event extraction system. In *Sixth applied natural language processing conference*, pages 76–83.

Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. 2018. Deep communicating agents for abstractive summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long Papers)*, pages 1662–1675, New Orleans, Louisiana. Association for Computational Linguistics.

Haolin Deng, Yanan Zhang, Yangfan Zhang, Wangyang Ying, Changlong Yu, Jun Gao, Wei Wang, Xiaoling Bai, Nan Yang, Jin Ma, and 1 others. 2022. 2event: Benchmarking open event extraction with a large-scale chinese title dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6511–6524.

Ralph Grishman, Silja Huttunen, and Roman Yangarber. 2002. Information extraction for enhanced access

to disease outbreak reports. *Journal of biomedical informatics*, 35(4):236–246.

Quanjiang Guo, Sijie Wang, Jinchuan Zhang, Ben Zhang, Zhao Kang, Ling Tian, and Ke Yan. 2026. Extracting events like code: A multi-agent programming framework for zero-shot event extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30880–30887.

Zijin Hong and Jian Liu. 2024. Towards better question generation in QA-based event extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9025–9038, Bangkok, Thailand. Association for Computational Linguistics.

Hyuntak Kim and Byung-Hak Kim. 2025. NexusSum: Hierarchical LLM agents for long-form narrative summarization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10120–10157, Vienna, Austria. Association for Computational Linguistics.

Bobo Li, Xudong Han, Jiang Liu, Yuzhe Ding, Liqiang Jing, Zhaoqi Zhang, Jinheng Li, Xinya Du, Fei Li, Meishan Zhang, and 1 others. 2025. Event extraction in large language model:a holistic survey of method, modality, and future. *arXiv preprint arXiv:2512.19537*.

Hongzhan Lin, Yang Deng, Yuxuan Gu, Wenxuan Zhang, Jing Ma, See Kiong Ng, and Tat-Seng Chua. 2025. Fact-audit: An adaptive multi-agent framework for dynamic fact-checking evaluation of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 360–381.

Jens P Linge, Marco Verile, Hristo Tanev, Vanni Zavarella, Flavio Fuart, and Erik van der Goot. 2012. Media monitoring of public health threats with medisys. *C. WILLIAM, CWR. WEB-STER, D. BALAHUR, et al*, pages 17–31.

Di Lu, Shihao Ran, Joel Tetreault, and Alejandro Jaimes. 2023. Event extraction as question generation and answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1666–1688, Toronto, Canada. Association for Computational Linguistics.

Meng Lu, Yuzhang Xie, Zhenyu Bi, Shuxiang Cao, and Xuan Wang. 2025. Crossagentie: Cross-type and cross-task multi-agent llm collaboration for zero-shot information extraction. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics.

Zihao Meng, Tao Liu, Heng Zhang, Kai Feng, and Peng Zhao. 2024. Cean: Contrastive event aggregation network with llm-based augmentation for event extraction. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 321–333.

Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. Optimizing instructions and demonstrations for multi-stage language model programs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9340–9366.

Jakub Piskorski, Nicolas Stefanovitch, Brian Doherty, Jens P Linge, Sopho Kharazi, Jas Mantero, Guillaume Jacquet, Alessio Spadaro, Giulia Teodori, and 1 others. 2023. Multi-label infectious disease news event corpus. In *Text2Story@ ECIR*, pages 171–183.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ere: Annotation of entities, relations, and events. In *Proceedings of the 3rd workshop on EVENTS: Definition, detection, coreference, and representation*, pages 89–98.

Ralf Steinberger, Bruno Pouliquen, and Erik Van der Goot. 2013. An introduction to the europe media monitor family of applications. *Information Access in a Multilingual World*.

Hristo Tanev, Nicolas Stefanovitch, Tomáš Harmatha, and Diana F. Sousa. 2025a. Exploring the performance of large language models for event detection and extraction in the health domain. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 1237–1247, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Hristo Tanev, Nicolas Stefanovitch, Tomáš Harmatha, and Diana F Sousa. 2025b. Exploring the performance of large language models for event detection and extraction in the health domain. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*.

Meihan Tong, Bin Xu, Shuai Wang, Yixin Cao, Lei Hou, Juanzi Li, and Jun Xie. 2020. Improving event detection via open-domain event trigger knowledge. Association for Computational Linguistics.

Yuli Vasiliev. 2020. *Natural Language Processing with Python and spaCy: A Practical Introduction*. No Starch Press.

Rui Wang, Jiaoli Liu, Yu Yan, Liwei Zang, Huimin Wang, and Jianyi Liu. 2025. Document-level event extraction framework based on prompt learning. In *International Conference on Computer Application and Information Security (ICCAIS 2024)*, volume 13562, pages 996–1001. SPIE.

Rui Wang, Deyu Zhou, and Yulan He. 2019. Open event extraction from online text using a generative adversarial network. *arXiv preprint arXiv:1908.09246*.

Sijia Wang and Lifu Huang. 2024. Debate as optimization: Adaptive conformal prediction and diverse retrieval for event extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16422–16435, Miami, Florida, USA. Association for Computational Linguistics.

Guangjun Zhang, Hu Zhang, Yazhou Han, Yue Fan, Yuhang Shao, Hongye Tan, and Ru Li. 2026. Learning to generate and extract: A multi-agent collaboration framework for zero-shot document-level event arguments extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34665–34673.

Chang Zong, Yuchen Yan, Weiming Lu, Jian Shao, Yongfeng Huang, Heng Chang, and Yueting Zhuang. 2024. Triad: A framework leveraging a multi-role LLM-based agent to solve knowledge base question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1698–1710, Miami, Florida, USA. Association for Computational Linguistics.

GENOME: A New Geopolitical Event Methodology and Dataset using Large Language Models

Alessandro Dell'Orto* **Jesse Kommandeur***
The Hague Centre for Strategic Studies
alessandrodell'orto@hcss.nl jessekommandeur@hcss.nl

Abstract

Quantitative research in international relations relies heavily on structured event data, yet existing automated datasets lack up-to-date coverage of both conflictual and cooperative interactions. We introduce GENOME (Geopolitical Event News Observatory, Mapping, and Extraction), an automatically extracted dataset that implements PLOVER's 16 event types and extends its Actor–Recipient schema with a Third Party role to capture multilateral relations from newswire data. GENOME's pipeline comprises event extraction, ontology-based classification, entity normalization, and deduplication, leveraging GPT models with one-shot prompting and enforced structured outputs. We compare GENOME against POLECAT dataset over a five-month overlap period across event volume, temporal dynamics, and geographical coverage. Results show that while the two datasets align closely on conflict event types, GENOME captures a more balanced distribution of cooperative events, particularly verbal interactions nearly absent in POLECAT. GENOME also demonstrates improved temporal precision by attributing events to their inferred date of occurrence rather than publication date, and effective deduplication of highly covered events.

1 Introduction

Collecting, coding and analysing geopolitical events has been a pivotal challenge for quantitative studies in International Relations since the 1960s (McClelland, 1961; Yonamine, 2016). Several datasets have been compiled over the following decades, where the main trade-off has been the balance between cost and reliability of manually annotated datasets or the size and rapidity of automated solutions. Most widely diffused manually coded ones are either historic and no longer maintained, exclusively conflict-focused or limited to specific domains, regions or diplomatic processes (Raleigh et al., 2023; Olsen et al., 2024).

On the other hand, automatically extracted datasets are based on event ontologies, rulebooks that provide a schema for representing events in a structured, machine-readable way. Numerous ontologies have been developed since the late 1970s (McClelland, 1978; Azar, 1980; Bond et al., 1994; King and Lowe, 2003b), giving rise to a range of automatically coded datasets (King and Lowe, 2003a; O'Brien, 2010; Jenkins et al., 2012; Leetaru and Schrodt, 2013; Salam et al., 2020). Most notably, CAMEO (Gerner et al., 2002) served as the foundation for ICEWS, GDELT, and Phoenix, and was succeeded by the PLOVER ontology, redesigned for greater simplicity, flexibility, and compatibility with machine learning approaches (Open Event Data Alliance, 2024). In 2023, POLECAT was introduced as an automatically coded dataset with global coverage of both conflict and cooperation political events (Halterman et al., 2023a), based on PLOVER, using the NGEC coder which employs transformer-based models to extract events from multilingual news sources (Halterman et al., 2023b). Around August 2024, POLECAT has stopped receiving updates, leaving a gap for up-to-date automatically extracted events covering both conflictual and cooperative interactions. Another relevant issue is that these ontologies lack publicly available gold datasets to test Event Extraction (EE) tasks on them, forcing scholars to adapt their methods to other event structures where such gold standards exist, such as ACE05 (Halterman et al., 2023b; Gao et al., 2023; Brandt and Sianan, 2025).

In recent years, Large Language Models (LLMs) have captured significant attention for socio-political EE, though zero-shot approaches often fall short for rigorous codebook-based measurement (Cai and O'Connor, 2023; Chen et al., 2024; Halterman and Keith, 2025). Earlier work on zero-shot ranking for socio-political texts found that re-

* Equal contribution.

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 83–95

trieval quality degrades as the target label becomes more general, and that declarative label formulations outperform dictionary definitions (Akdemir and Hürriyetoğlu, 2022). Key challenges include hallucination, output inconsistency, deviation from ontology schemas, and degraded performance on non-Western sources and low-resource languages (Thapa et al., 2025), as well as mismatches between the nuanced, evolving nature of real-world events and the rigid dyadic structures of current ontologies (Brandt and Sianan, 2025). Existing LLM-based EE pipelines and datasets have all been either conflict or region-focused, like the horn of Africa (Bai et al., 2025; Meher and Brandt, 2025; Semnani et al., 2025). However, tracking low-intensity, cooperative, and verbal interactions is valuable for early-warning and conflict prediction, as escalation is best understood as a sequential process in which preceding lower-intensity interactions carry meaningful predictive signal (Halkia et al., 2020; Beardsley et al., 2024).

Given the current gap left by POLECAT in automatically extracted geopolitical events from news data that tackle both conflictual and cooperative, verbal and material inter-state interactions, we introduce GENOME (Geopolitical Event News Observatory, Mapping, and Extraction). The name reflects the dataset's ambition: just as genomics maps the fundamental building blocks of biological life, GENOME aims to map the building blocks of geopolitical interaction through underlying patterns of conflict and cooperation between states. GENOME implements PLOVER's 16 event types and extends its Actor-Recipient schema to capture multilateral relations from newswire data that more closely mirror real-world geopolitical interactions. It incorporates up-to-date newswire articles and a two-phase extraction-classification pipeline leveraging GPT models, along with a series of entity normalization and deduplication techniques.

While GENOME, like most other ontology-based datasets, lacks a manually annotated gold standard, we evaluate it by comparing a 5-month overlapping sample of approximately 28,000 events against POLECAT: even with different input sources, both datasets show similar behaviour in conflict events and temporal dynamics, while differing significantly in how they track cooperation, especially verbal. We also show that GENOME better aligns events with their actual occurrence dates (rather than news publication dates), accurately resolves international entities such as the IMF and NATO, and reduces redundancy among heavily reported events. The dataset and analysis scripts are publicly released.[1] The remainder of this paper describes GENOME's design and ontology changes (Section 2), details our methodology (Section 3), presents our experimental setup and results (Sections 4–5), and discusses findings and limitations (Section 6).

2 Concept and Design

2.1 Sources

GENOME receives as input a corpus of English-language newswire articles collected from a commercial source, primarily covering international affairs and including a publication date. The corpus currently spans from January 2024 to January 2026 and contains approximately 148,000 articles from around 2,400 unique sources, of which roughly 30,000 articles fall within the February–June 2024 period used for comparison (Sections 4–5).

2.2 Ontology Changes

GENOME events are based on a simplified and slightly modified version of the PLOVER ontology. They are *simplified* because the system only focuses on the "what" of the "What–How–Why" logic from PLOVER. It does this by classifying events by their Event Type, using the same 16 types grouped into four categories based on their Conflict/Cooperation and Material/Verbal nature. Moreover, it follows the same temporal and reality logic (focusing on things that have already happened and are not just being discussed), explicit Actor logic (where Actors must be explicitly named while the Recipient role is optional), and compound logic (where events that involve multiple Actors for a single role generate a single entry).

They are *modified* because we extend PLOVER's Actor–Recipient framework by introducing the overarching concept of an Agent, defined as an individual, organization, country, or social group that is capable of taking action (speaking, moving, fighting) or having a political action directed at them. Within this framework, we identify three Agent roles:

- **Actor:** The Agent that initiates or performs the action described in the event.

[1] https://github.com/HCSS-Data-Lab/Submission-GENOME

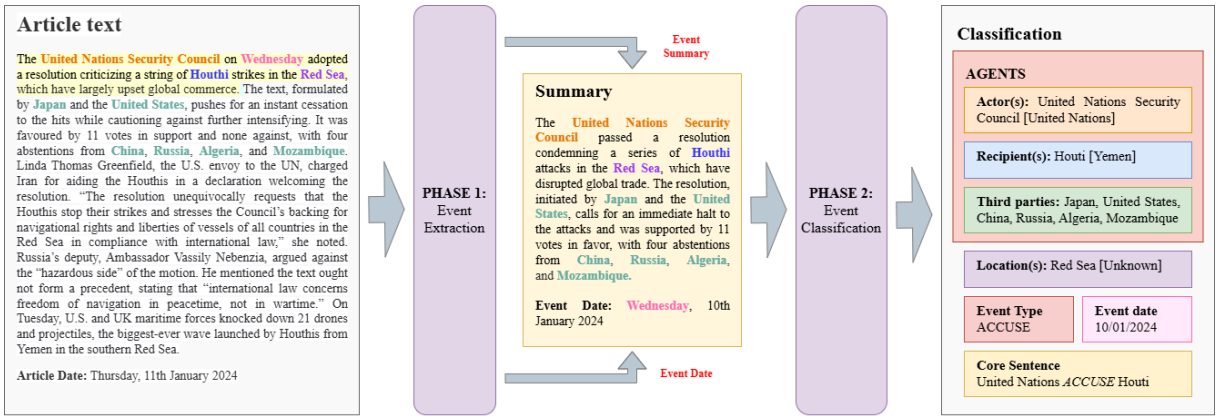


Figure 1: Simplified overview of the event extraction, classification, and normalization modules of the GENOME pipeline.

- **Recipient:** The Agent that is the direct target of the action. This role is optional, as not all events have a clearly defined recipient.
- **Third Party:** An Agent that provides contextual information relevant to the event but is neither the initiator (Actor) nor the target (Recipient). For example, in "A accused B of arming C," A is the Actor and B is the Recipient, while C is a Third Party. The implied action "B arming C" is not treated as a separate event, as it violates the reality logic.

We introduce the Third Party role for two reasons. First, in pilot experiments, we observed that LLMs tend to assign a role to every identified Agent, so it is useful to provide a designated category for context-bearing entities. Second, similar to how POLECAT further classifies Actors and Recipients, Third Parties can also be categorized based on their role (e.g., "facilitator" or "reporter") and nature (e.g., "government" or "military"). This creates a pathway from dyadic actor–recipient representations to richer multi-entity event structures (e.g., hypergraph-style representations) that more explicitly preserve temporal and relational structure (Ahrabian et al., 2025).

3 Methodology

3.1 Pipeline and Model Choice

Our pipeline consists of four modules executed sequentially: (1) extracting event summaries from articles' raw text and inferring the event date; (2) classifying events by applying our version of the PLOVER ontology; (3) normalizing ambiguous names and resolving entities; and (4) deduplicating multiple occurrences of the same event. Figure 1 visualizes a simplified version of the first three modules. For both the extraction and classification phases, our strategy relies on one-shot prompting combined with enforced structured outputs using Pydantic models. During our testing, OpenAI models proved to be the most robust choice, offering the best balance across performance, cost, and native support for both structured output and fine-tuning capabilities.

3.2 Event Extraction

The extraction task is primarily a summarization task, which does not require the advanced reasoning capabilities of frontier LLMs. We deliberately separate event extraction from classification to reduce the complexity of individual tasks, allowing us to utilize more cost-effective, lower-parameter models. Moreover, distilling the raw text into a structured, concise event summary standardizes the input for the classification task, significantly reducing input size and variability (You et al., 2025).

To reduce input tokens and isolate the core news lead, article texts are truncated at the sentence boundary nearest the 200-word mark. This approach is highly effective, but requires the input text to follow the standard newswire format, where the main news is placed at the beginning of the article. `gpt-4o-mini` was chosen as the most cost-effective option for this kind of task.

We inject the article's date in long british format and pre-cleaned text as a prompt to extract event objects. Each object comprises three components. First, an inferred event date (e.g., deriving "14th January 2024" from an article dated "Monday, 15th January 2024" that reports an event "on Sunday"); second, an event description containing all funda-

mental and contextual elements—a structure that mimics the concise event summaries of the ACLED dataset (Raleigh et al., 2023); and third, a source quote representing the verbatim text used to construct the object. The prompt explicitly requests each event description to be fully self-contained and understandable without reference to external information, to articulate the core action of the event in a concise form, and to complement it with the main relevant contextual information. Lastly, the model is tasked to extract the core event of the article, and to only report multiple events if a single "core" event cannot be unequivocally identified. This ensures that extracted events are the actual subject of the newswire article, rather than contextual or historical events mentioned in passing.

3.3 Event Classification

Event objects are expanded by a second LLM call, which classifies them based on our extension of the PLOVER ontology described in Section 2.2. This task, particularly correctly identifying the core event action and assigning each involved party to its Agent role, is more cognitively complex. However, this complexity is reduced by the concise and consistently structured summaries produced in the previous step, which yield a highly consistent format for both input and output.

To reduce cost, we leverage this repetitiveness by fine-tuning `gpt-4.1-nano` on 775 event descriptions classified with `gpt-5.1`, with an 80/20 train/test split, using OpenAI's supervised fine-tuning tool. The process yielded a training loss of 0.114 and a validation loss of 0.122, demonstrating successful model convergence and strong generalization to unseen data with minimal overfitting.

The classification prompt includes the definition of Agent and of Agent roles as they are defined in Section 2.2, along with examples to distinguish recognized entities that are agents (Actor, Recipient, Third Party) from those that are not (e.g., locations such as "the Middle East", or concepts such as "Climate Change"). To facilitate reasoning on roles, we ask the model to report the main action as "A *verb* B", where A and B are the Actor or Recipient and the verb is the event type from the PLOVER ontology (Figure 1). Finally, the model maps all entities to their corresponding country, which serves as the basis for the normalization step.

3.4 Entity Normalization

Entity normalization proceeds in two stages. First, extracted names of countries and international bodies are matched via fuzzy search against a manually curated reference list of 248 countries and 40 organizations, along with a series of synonyms. Unmatched entries are resolved manually. This process successfully resolved all inferred country names, reducing the count of unique countries and international organizations by 81.7%.

Second, agent names are grouped by their normalized country or organization. These names are converted into vector embeddings using a Sentence-Transformer model and clustered based on semantic similarity using FAISS. The most frequent entry in each cluster is selected as the canonical normalized name. Consequently, the volume of unique agent names decreased by 35.4% to 32,624 entries.

3.5 Deduplication

To consolidate duplicates extracted from multiple sources, we implement a multi-criteria deduplication pipeline. Candidate duplicate events are first filtered by temporal proximity within a 2-day window, then scored using a weighted composite of four similarity components: semantic similarity of event summary embeddings using `all-MiniLM-L6-v2` (0.45), actor and recipient name overlap via Jaccard similarity (0.25), event type match (0.20), and location overlap (0.10). Pairs exceeding a composite threshold of 0.80 are linked as duplicates. Duplicate clusters are resolved via connected components on the resulting similarity graph, merging all contributing article identifiers.

3.6 Operational Cost

Across the full two-year corpus (roughly 6,000 newswire articles ingested per month), total OpenAI API spend was approximately $55, comprising $19 for extraction with `gpt-4o-mini`, $33 for classification with the fine-tuned `gpt-4.1-nano`, and a one-off $4 fine-tuning job, all using the batch API at a 50% discount on standard rates. This corresponds to about $2.30 per month, or on the order of $0.0004 per processed article. Because fine-tuning is performed once on a silver-standard sample, its cost is amortized over all subsequent extractions and does not scale with corpus size.

4 Experimental Setup

Evaluating GENOME presents inherent challenges common to automated event extraction: no manually annotated gold standard exists for the PLOVER ontology, and direct record-level alignment with POLECAT is infeasible due to differences in input data and GENOME's extended Actor–Recipient schema introduced in Section 2.2. Consequently, the next best alternative and the most robust evaluation feasible under these constraints is to assess how similarly GENOME behaves with respect to POLECAT and known real-world geopolitical events. We therefore adopt a systematic indirect comparison against POLECAT of multiple versions of the GENOME dataset and a cross-comparison between those versions, complemented by a case study to validate the datasets against known occurrences.

4.1 Dataset Variants

We evaluate our pipeline against the publicly accessible version of POLECAT (Scarborough et al., 2023). All quantitative comparisons are restricted to the five-month overlap period between the datasets (February to June 2024). We compare POLECAT against three GENOME configurations:

- **GENOME (article_date)**: The output of our pipeline after the normalization step (Section 3.4), timestamping events using the *publication date* of the source news article.
- **GENOME (event_date)**: The same table as GENOME (article_date), but incorporating the *event date* extracted by the model during the extraction phase (Section 3.2).
- **GENOME (dedup)**: The output of the pipeline after the deduplication step (Section 3.5), timestamping events with the same extracted *event date* as GENOME (event_date).

While POLECAT includes an analogous extracted event date, we present evidence in Sections 5 and 6 suggesting that its dates more closely reflect publication timing rather than actual occurrences. Lastly, in subsequent sections, when the specific date choice is irrelevant to an evaluation metric, we refer to the non-deduplicated versions collectively as GENOME, and the deduplicated version as GENOME (dedup).

4.2 Approach and Metrics

We structure our evaluation around four complementary objectives. First, we assess event volume and balance, comparing the total number of events recorded by each dataset and their distribution across the four PLOVER quad categories (verbal/material $\times$ conflict/cooperation) and across PLOVER's 16 individual event types. This reveals whether the two systems, despite different input sources, agree on how geopolitical activity is categorized under a shared ontology.

Second, we examine temporal dynamics. Using daily event counts, we compute summary statistics including means, standard deviations, and the coefficient of variation (CV) to quantify volatility. We compare weekday and weekend means through a weekend/weekday ratio (W/D) to assess sensitivity to media publication cycles. We also analyse the distribution of the lag between article publication date and inferred event date, and we measure the Pearson correlation between daily event volume and deduplication rate (defined as the proportion of events removed on a given day) to assess whether deduplication preferentially targets high-volume days. To formally quantify cross-dataset agreement, we compute Pearson correlations between POLECAT and each GENOME variant on daily event counts, both overall and stratified by PLOVER quad category.

Third, we evaluate geographical coverage and entity structure. We compare the sets of unique country dyads across datasets. For both POLECAT and GENOME, we treat a country dyad as the set of events in which both countries appear among the Actor or Recipient Agent roles. We examine the overlap and composition of shared and exclusive dyads, particularly in terms of event types and international organization resolution. At the country level, we quantify the volume gap and identify where surpluses concentrate. We also report the average number of actors, recipients, and third-party countries per event to explore structural differences.

Fourth, we conduct dyad-level case studies against known real-world occurrences. The primary case study examines Israeli–Iranian dyadic interactions between March 25 and April 25, 2024. This period contains three major, heavily covered events: the Israeli airstrike on the Iranian embassy complex in Damascus (April 1), the Iranian retaliatory strikes on Israel (April 13), and the Israeli

attack on Iranian military sites (April 19). With the same criteria we isolate the 14 dyads with at least 100 events in both POLECAT and GENOME, for which we compute the same daily-count correlation by quad category. The first case study serves as a qualitative check on temporal precision and deduplication effects against known real-world occurrences, while the 14-dyad correlation sweep tests whether aggregate agreement persists once geographic mix is controlled for.

5 Results

5.1 Event Volume and Balance

Volume: Figure 2 presents the distribution of event types across datasets. POLECAT records substantially higher event volume than GENOME (157,328 vs. 28,530 events), with a ratio of approximately 5.5:1, despite drawing from far fewer sources (242 vs. 1,036). GENOME (dedup) reduces the number of events to 16,795 (60% of the GENOME initial volume), bringing the ratio with POLECAT above 9:1 (full counts in Table 4 in the Appendix).

Macro proportions: POLECAT is markedly more conflict-heavy (77.7% conflict vs. 22.3% cooperation), whereas GENOME presents a more balanced distribution (56.3% conflict vs. 43.7% cooperation), with the deduplicated version showing an even more balanced split. Furthermore, the composition of these interactions differs significantly: POLECAT's cooperation is primarily material (19.4%) rather than verbal (2.9%), while GENOME displays the inverse, heavily favouring verbal cooperation (34.2%) over material actions (8.5%).

Event-Type proportions: Conflict event types are distributed similarly across both datasets. Cooperation, however, diverges sharply. In verbal cooperation, POLECAT is almost entirely dominated by CONCEDE (94.1%), with no AGREE or SUPPORT events recorded, while GENOME distributes most of its verbal cooperation across CONSULT (66.6%) and AGREE (24.9%). Material cooperation also differs structurally, though both datasets agree on AID as the dominant type.

5.2 Temporal Dynamics

Daily flow: POLECAT exhibits the highest volatility (CV 40.9%), driven by a sharp weekend drop (W/D 0.321) that pulls its mean well below the median (Table 1). GENOME (dedup) shows the most stable flow (CV 30.1%) but a wider weekday–weekend gap than the non-deduplicated versions, while GENOME (event_date) is notably more volatile than GENOME (article_date).

Metric	POLECAT	GENOME (art.)	GENOME (evt.)	GENOME (dedup)
Mean	1048.85	190.20	190.20	111.97
Median	1235.0	188.0	184.5	116.0
Std	429.34	62.93	75.80	33.67
CV (%)	40.9	33.1	39.9	30.1
Wkday	1302.24	206.40	206.55	124.28
Wkend	418.33	149.88	149.51	81.33
W/D	0.321	0.726	0.724	0.654

Table 1: Events per day summary statistics (Feb–Jun 2024). Wkday = mean weekday count, Wkend = mean weekend count, W/D = weekend/weekday ratio.

Δ (days)	Count	% of Total
< 0	1,247	4.4
$= 0$	14,933	52.3
1	8,946	31.4
2	1,440	5.0
3–5	1,133	4.0
6–10	529	1.9
> 10	302	1.1

Table 2: Distribution of event date vs. article date lag (Δ) for GENOME prior to deduplication (Feb–Jun 2024). Total events: 28,530.

Event vs. Article Date: GENOME's inferred event dates follow expected behaviour: most events (52.3%) fall on the article publication date, with shares declining steadily as lag increases, and only 4.4% show a negative difference; we attribute this to hallucination, time-zone mismatches, or scheduled future events (e.g., "The meeting will start on Thursday") (Table 2). This is clear in the Israel–Iran case study (Figure 3), where GENOME (event_date) produces sharp spikes on the exact dates of the three major events, while both POLECAT and GENOME (article_date) show a visible, similar delay.

Deduplication effects: The daily volume of non-deduplicated GENOME events and the deduplication rate are significantly correlated ($r = 0.57$, $p < 0.001$), consistent with high-volume days containing more duplicate reports. The Israel–Iran case study (Figure 3) shows the same pattern.

Aggregate correlation: Looking beyond the Iran–Israel case, daily Pearson correlations between POLECAT and the three GENOME variants over the full 150-day window are $+0.40$ (article_date), $+0.30$ (event_date), and $+0.58$ (dedup),

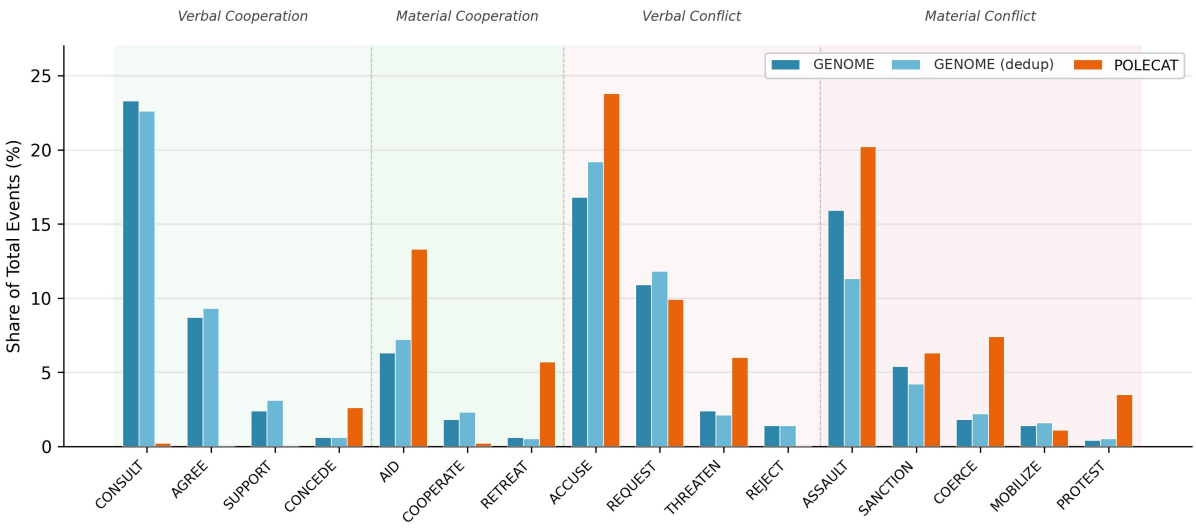


Figure 2: Distribution of event types by share of total events (%) for GENOME, GENOME (dedup), and POLECAT (Feb–Jun 2024). Background shading distinguishes the four PLOVER quad categories. Conflict event types show close alignment between datasets, while cooperation—especially verbal—diverges sharply.

Quad category	article	event	dedup
Overall	+0.40	+0.30	+0.58
Verbal coop.	+0.41	+0.37	+0.39
Material coop.	+0.25	+0.28	+0.43
Verbal conf.	+0.41	+0.33	+0.51
Material conf.	+0.09	−0.00	+0.20

Table 3: Daily-count Pearson r between POLECAT and each GENOME variant by PLOVER quad category (Feb–Jun 2024).

with deduplication strengthening agreement in all four PLOVER quads (Table 3). Despite their similar event-type shares, material conflict shows the weakest aggregate daily-count agreement across all three variants (Pearson $r = +0.09$ article_date, $r = -0.00$ event_date, $r = +0.20$ dedup).

Dyad-level correlation: Restricting to the 14 dyads with at least 100 events in both GENOME and POLECAT, material conflict correlation rises sharply: positive in 13 of 14 dyads with article_date r ranging from $+0.12$ to $+0.81$ (Iran–Israel). The sole exception, Ukraine–United States, is also the only dyad where cooperation dominates (verbal cooperation $r = +0.60$). Across all quad categories, daily correlation with POLECAT is highest under article_date in 8 of 14 dyads, confirming POLECAT's bias toward publication cycles (full per-dyad breakdown in Appendix Table 5).

Day-of-week analysis: The overall weekend/weekday ratio remains stable across GENOME date versions (Table 1), but switching to event dates shifts the composition: weekend material conflict rises by 8.2%, while verbal and cooperative events decline. Across all GENOME versions, material events are less sensitive to day-of-week effects than verbal ones, which drop visibly on weekends (Figure 4). POLECAT, by contrast, shows a uniform drop across all categories.

5.3 Geographical Coverage and Entity Structure

Dyad-level comparison: POLECAT has far more unique country dyads (4,471 vs. GENOME's 1,691), but including GENOME's third-party countries narrows this gap substantially (4,191). Only 2,384 dyads are shared, yet these cover the top dyads by volume (e.g., Israel–Iran, Russia–Ukraine, USA–China), with higher mean events per dyad in the shared group—suggesting convergence on the most salient relationships.

Exclusive dyads and entity resolution: The datasets' exclusive dyads differ structurally: nearly half of GENOME's belong to verbal cooperation (under 2% for POLECAT). GENOME also resolves specific international bodies (NATO, IMF, G7) that POLECAT aggregates under a generic "International Organization" label. Conversely, POLECAT's exclusive dyads include state-to-state pairs such as Chile–Venezuela and Colombia–India, indicating gaps in GENOME's non-Western coverage.

Country-level distribution: For most countries, POLECAT records substantially more events than GENOME: the five with the largest gap (India, Russia, Israel, Ukraine, Brazil) alone account for a dif-

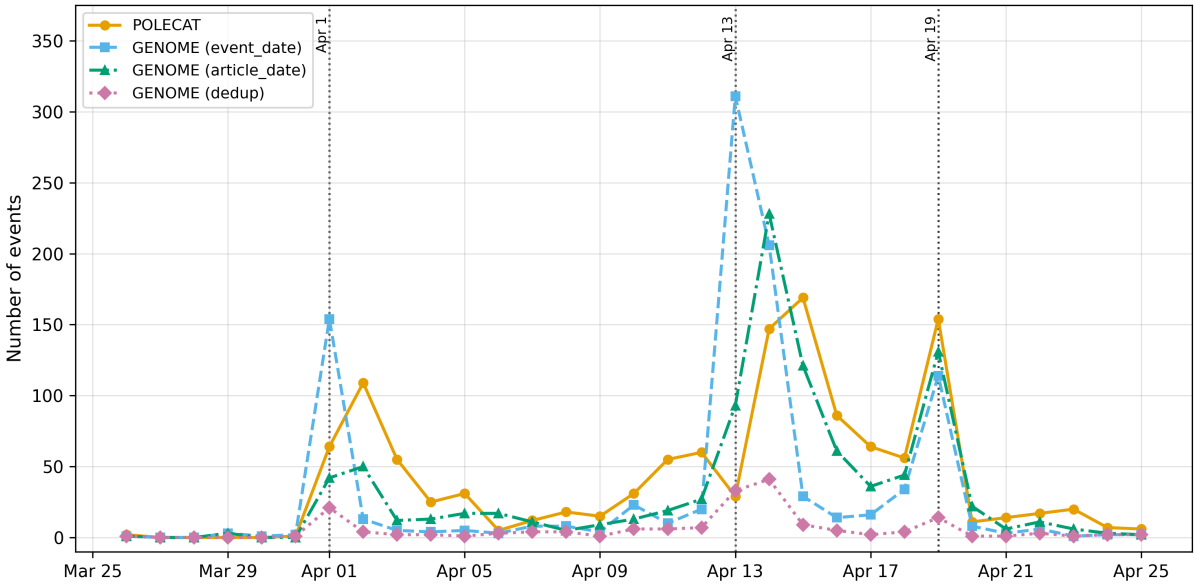


Figure 3: Daily event counts for the Israel–Iran case study (Mar 25 – Apr 25, 2024).

ference of 42,000 events. This disparity is also pronounced in Latin America, the Caucasus, and South and Southeast Asia. GENOME exceeds POLECAT only for a narrow cluster, most notably Palestine and the United Nations.

Event structure: GENOME incorporates more countries per event: 1.34 actors, 0.88 recipients, and 1.09 third parties on average, compared to POLECAT's 1.05 actors and 0.55 recipients.

6 Discussion

GENOME and POLECAT can be understood as similar products with different underlying mechanics, which is reflected in differences in event volume, event composition, temporal attribution, and entity representation. GENOME produces a substantially lower event volume, largely because it typically yields one event per article while encoding richer agent structures. In contrast, POLECAT extracts more events per article and covers a broader range of sources, including some non-Western regions. It is also likely that POLECAT introduces more duplicates of the same underlying event, either due to more redundant inputs or a higher extraction rate within individual articles.

Despite strong alignment between the two datasets on conflict event types, reflecting their shared ontology, a notable divergence emerges in cooperative events. POLECAT contains almost no CONSULT events and shows a generally low volume of verbal cooperation. This pattern suggests that routine diplomatic interactions, such as meetings and official statements, are filtered out at the input stage rather than misclassified. GENOME, by contrast, captures a more balanced distribution of cooperative events, which is particularly relevant for early warning applications where lower intensity interactions may carry predictive signal. Despite operating on different input corpora, the two datasets converge on the same major events. Daily-count correlations with POLECAT are weakest in aggregate, with material conflict in particular appearing nearly uncorrelated. However, once restricted to dyads where both systems record substantial activity, material conflict correlation turns sharply positive in nearly all cases. The aggregate gap therefore reflects differing geographic coverage rather than mismatched classifications.

GENOME also appears better able to attribute events to their actual date of occurrence rather than the publication date of the reporting source. This distinction is reflected in the increased volatility observed when moving from GENOME (article_date) to GENOME (event_date): publication dates introduce artificial smoothing due to reporting delays, whereas event dates cluster reports onto the day the event occurred, producing sharper temporal signals around known geopolitical developments. This effect is especially visible in the Israel–Iran case study (Figure 3). POLECAT, in contrast, shows a uniform decline in events over weekends, indicating stronger sensitivity to media publication cycles. This bias is further confirmed at the dyad level, where POLECAT's daily counts align most closely

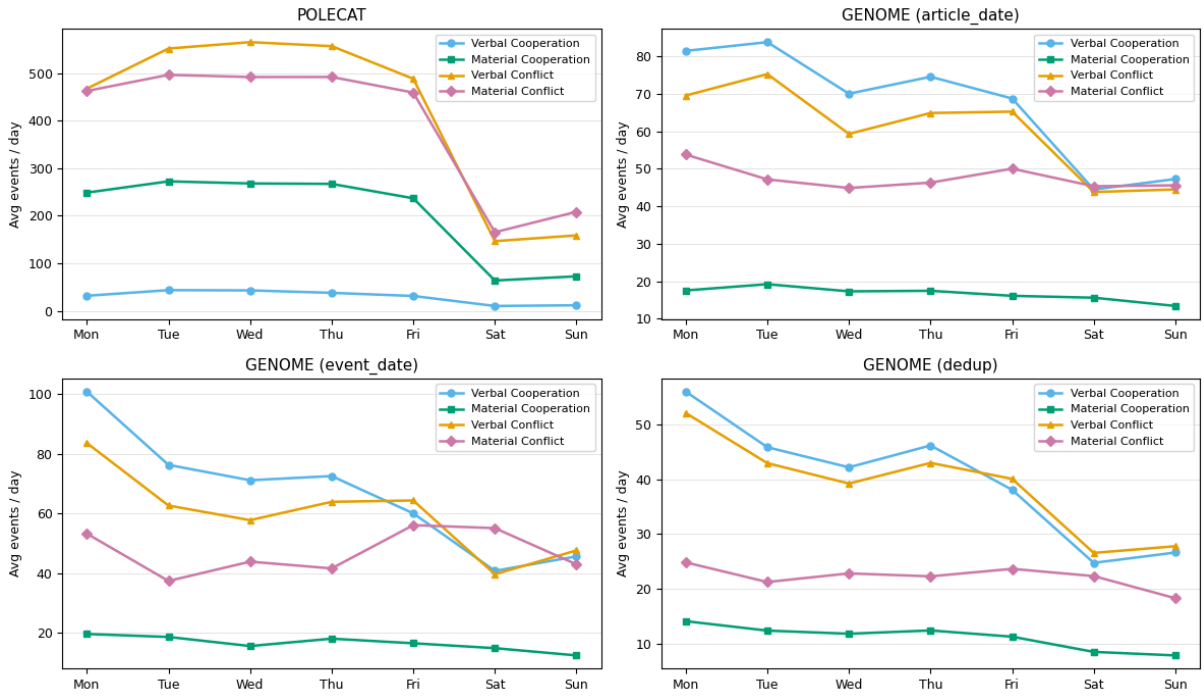


Figure 4: Day-of-week event volume by category across datasets.

with GENOME (article_date) in the majority of shared dyads. GENOME also exhibits a weekend reduction, but this is concentrated in verbal events, consistent with reduced diplomatic activity and public statements, while material events remain relatively stable. This suggests that GENOME's pipeline partially mitigates media cycle bias. Its deduplication step further reduces redundancy in highly reported events, although its accuracy still requires formal evaluation.

Finally, GENOME demonstrates more granular entity resolution. It distinguishes a range of international organisations, including NATO, IMF, WTO, BRICS, and the ICC, whereas POLECAT largely restricts such entities to major supranational bodies like the European Union and the United Nations, grouping others under a generic "International Organization" category. This allows GENOME to capture supranational interactions in greater detail. In addition, the Third Party role further supports the shift toward multi-entity event representations described in Section 2.2.

7 Conclusion

As of today, GENOME provides a continuously updated, automatically extracted dataset of geopolitical events covering both conflictual and cooperative, verbal and material inter-state interactions. Its pipeline demonstrates that LLM-based extraction can produce structured event data that aligns with established ontologies and captures dynamics consistent with real-world patterns. However, while POLECAT embodies a more granular detail level (e.g., geolocalization, Wikidata normalization, context-mode entries), GENOME currently represents a foundation to be built upon. Future work will focus on reintroducing these features, enriching the dataset with finer-grained classification of events and agent attributes, and expanding coverage through more diverse, multilingual input sources. Additionally, developing a manually annotated gold standard for the PLOVER ontology remains a priority, both for GENOME and for the broader socio-political event data community, as it would enable systematic evaluation of extraction quality across competing systems.

Limitations

Several limitations should be acknowledged. First, GENOME relies exclusively on English-language newswire sources from a single commercial provider, which introduces both structural and linguistic bias toward Western media perspectives and English-speaking regions. This likely explains the coverage gaps in Latin America, the Caucasus, and South and Southeast Asia identified in the geographical comparison with POLECAT. Relying on Western-centric, English-language media

inherently restricts the observability of localized or low-intensity geopolitical interactions in the Global South.

Second, GENOME lacks a manually annotated gold standard for evaluation, a limitation shared by the broader automated geopolitical event extraction community. While the indirect comparison with POLECAT and known real-world occurrences provides useful indicators of systemic validity, the lack of record-level ground truth prevents a formal calculation of standard precision, recall, and F1 metrics for the extraction and classification modules. Without such ground truth, it remains difficult to determine whether divergences from POLECAT, particularly in cooperative event types, reflect genuine improvements in coverage or systematic biases introduced by the LLM-based pipeline. Developing such a dataset remains a priority to enable systematic evaluation across competing systems.

Third, to minimize hallucination and standardize inputs, the extraction module is deliberately prompted to identify the single "core" event of an article, discarding secondary, historical, or contextual interactions unless strictly necessary. While this design choice successfully reduces noise and redundancy, it inherently caps the recall of the system, potentially missing valid but subordinate events embedded later in the text. Additionally, article texts are truncated at approximately 200 words under the assumption of a standard newswire inverted-pyramid structure. Articles that deviate from this format, such as feature pieces, opinion columns, or non-Western journalistic conventions, may yield incomplete or inaccurate extractions. The 4.4% of events with negative date lags further suggests that the date inference mechanism is susceptible to hallucination, time-zone ambiguities, and references to scheduled future events.

Fourth, while the multi-criteria deduplication pipeline and entity resolution steps demonstrably reduce the redundancy of highly covered events and clean the dataset's network structure, their accuracy has not been formally evaluated against human judgments. The chosen similarity thresholds (e.g., the 0.80 composite score) were optimized iteratively on pilot data but may require domain-specific tuning to prevent the over-merging of distinct but similar events occurring in close temporal proximity. Similarly, the 35.4% reduction in unique agent names through embedding-based clustering risks merging distinct entities that share similar names or descriptions.

Fifth, the current implementation utilizes a simplified version of the PLOVER ontology. It focuses on the 16 root event types and introduces a novel Third Party role, but it temporarily omits finer-grained sub-types, exact geolocalization, Wikidata normalization, and contextual-mode tagging that POLECAT provides, limiting the granularity of downstream analyses.

Finally, the pipeline depends on proprietary, closed-source LLMs (the OpenAI GPT family), including `gpt-4o-mini` for extraction and a fine-tuned `gpt-4.1-nano` for classification. While cost-effective and highly performant for structured outputs, this reliance introduces concerns regarding reproducibility over time, as underlying model weights and behaviors may be updated by the provider without notice. The fine-tuning process further relies on silver-standard labels produced by `gpt-5.1` rather than human annotations, potentially propagating systematic errors into the classification module.

References

Kaiqian Ahrabian, Ethan Boxer, and Jay Pujara. 2025. Toward better temporal structures for geopolitical events forecasting. In *Proceedings of the 4th Workshop on Generation, Evaluation & Metrics (GEM 2024)*, pages 588–602. Association for Computational Linguistics.

Kiymet Akdemir and Ali Hürriyetoğlu. 2022. Zero-shot ranking socio-political texts with transformer language models to reduce close reading time. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022)*, pages 124–132. Association for Computational Linguistics.

Edward E. Azar. 1980. The conflict and peace data bank (COPDAB) project. *Journal of Conflict Resolution*, 24(1):143–152.

Rui Bai, Dong Lu, Sheng Ran, Emily M. Olson, Himank Lamba, Allison Cahill, Joel Tetreault, and Alejandro Jaimes. 2025. CEHA: A dataset of conflict events in the horn of africa. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*. Association for Computational Linguistics.

Kyle Beardsley, Patrick James, Jonathan Wilkenfeld, and David Quinn. 2024. What is escalation? Measuring crisis dynamics in international relations with human and LLM generated event data. ArXiv:2402.03340.

Doug Bond, Brian Bennett, and William Vogele. 1994. Data development and interaction events analysis using KEDS/PANDA: An interim report. Technical report, International Studies Association, Washington, DC.

Patrick T. Brandt and Marlo Sianan. 2025. Measurement of event data from text. *Frontiers in Political Science*, 6.

Erica Cai and Brendan O'Connor. 2023. A monte carlo language model pipeline for zero-shot sociopolitical event extraction. ArXiv:2305.15051.

Ruoxi Chen, Chengzhi Qin, Wenxing Jiang, and Donna Choi. 2024. Is a large language model a good annotator for event extraction? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17772–17780.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023. Exploring the feasibility of ChatGPT for event extraction. *Preprint*, arXiv:2303.03836.

Deborah J. Gerner, Philip A. Schrodt, Omür Yilmaz, and Rajaa Abu-Jabr. 2002. Conflict and mediation event observations (CAMEO): A new event data framework for the analysis of foreign policy interactions. Paper presented at the International Studies Association, New Orleans.

Myrsini Halkia, Stefano Ferri, Michail Papazoglou, and Marlies Kampen Schellens. 2020. Dynamic global conflict risk index. Technical report, Publications Office of the European Union. JRC Technical Report.

Andrew Halterman, Benjamin E. Bagozzi, Andreas Beger, Philip A. Schrodt, and Grace I. Scarborough. 2023a. PLOVER and POLECAT: A new political event ontology and dataset. SocArXiv.

Andrew Halterman and Katherine A. Keith. 2025. Codebook LLMs: Evaluating LLMs as measurement tools for political science concepts. *Political Analysis*.

Andrew Halterman, Philip A. Schrodt, Andreas Beger, Benjamin E. Bagozzi, and Grace I. Scarborough. 2023b. Creating custom event data without dictionaries: A bag-of-tricks. ArXiv:2304.01331.

J. Craig Jenkins, Charles L. Taylor, Marianne Abbott, Thomas V. Maher, and Lindsey Peterson. 2012. The world handbook of political indicators IV. Technical report, Mershon Center for International Security Studies, The Ohio State University.

Gary King and Will Lowe. 2003a. 10 million international dyadic events. IQSS Dataverse Network.

Gary King and Will Lowe. 2003b. An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization*, 57(3):617–642.

Kalev Leetaru and Philip A. Schrodt. 2013. GDELT: Global data on events, location, and tone, 1979–2012. In *ISA Annual Convention*, volume 2, pages 1–49.

Charles A. McClelland. 1961. The acute international crisis. *World Politics*, 14(1):182–204.

Charles A. McClelland. 1978. World event/interaction survey, 1966–1978. Technical Report 5211, ICPSR.

Shrey Meher and Patrick T. Brandt. 2025. ConflLlama: Domain-specific adaptation of large language models for conflict event classification. *Research & Politics*, 12(3).

Sean P. O'Brien. 2010. Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review*, 12(1):87–104.

Helene Brinken Olsen, Étienne Simon, Erik Velldal, and Lilja Øvrelid. 2024. Socio-political events of conflict and unrest: A survey of available datasets. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 40–53. Association for Computational Linguistics.

Open Event Data Alliance. 2024. PLOVER: Political language ontology for verifiable event records — event, actor and data interchange specification (draft version 2.0).

Clionadh Raleigh, Roudabeh Kishi, and Andrew Linke. 2023. Political instability patterns are obscured by conflict dataset scope conditions, sources, and coding choices. *Humanities and Social Sciences Communications*, 10:74.

Shehzad Salam, Patrick Brandt, Vito D'Orazio, Jennifer Holmes, Javier Osorio, and Latifur Khan. 2020. An online structured political event dataset based on CAMEO ontology. SocArXiv.

Grace I. Scarborough, Benjamin E. Bagozzi, Andreas Beger, John Berrie, Andrew Halterman, Philip A. Schrodt, and Jevon Spivey. 2023. POLECAT weekly data.

Sina J. Semnani, Peng Zhang, Wenxuan Zhai, Hao Li, Nicholas Beauchamp, Tyler Billing, Koko Kishi, Michaela Li, and Monica Lam. 2025. LEMONADE: A large multilingual expert-annotated abstractive event dataset for the real world. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25813–25852. Association for Computational Linguistics.

Surendrabikram Thapa, Suresh Adhikari, Hristo Tanev, and Ali Hürriyetoğlu. 2025. Challenges and applications of automated extraction of socio-political events at the age of large language models. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025*, pages 6–19.

Jay Yonamine. 2016. A guide to event data: Past, present, and future. *All Azimuth: A Journal of Foreign Policy and Peace*.

Huiling You, Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2025. Event-based evaluation of abstractive news summarization. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 504–510. Association for Computational Linguistics.

Appendix

A Event Type Counts

See Table 4.

B Dyad-level Pearson correlations

See Table 5.

Table 4: Event counts and shares by type for GENOME, GENOME (dedup), and POLECAT (Feb–Jun 2024).

	GENOME		Dedup		POLECAT	
Type	**N**	**%**	**N**	**%**	**N**	**%**
Verbal cooperation						
CONSULT	6,646	23.3	3,794	22.6	267	0.2
AGREE	2,486	8.7	1,565	9.3	0	0.0
SUPPORT	686	2.4	513	3.1	0	0.0
CONCEDE	168	0.6	105	0.6	4,227	2.7
Material cooperation						
AID	1,787	6.3	1,207	7.2	21,175	13.5
COOPERATE	525	1.8	382	2.3	319	0.2
RETREAT	166	0.6	88	0.5	9,101	5.8
Verbal conflict						
ACCUSE	4,786	16.8	3,232	19.2	37,580	23.9
REQUEST	3,121	10.9	1,989	11.8	15,738	10.0
THREATEN	685	2.4	351	2.1	9,484	6.0
REJECT	387	1.4	238	1.4	29	0.0
Material conflict						
ASSAULT	4,526	15.9	1,903	11.3	30,254	19.2
SANCTION	1,554	5.4	701	4.2	10,093	6.4
COERCE	506	1.8	374	2.2	11,858	7.5
MOBILIZE	401	1.4	268	1.6	1,746	1.1
PROTEST	100	0.4	85	0.5	5,457	3.5
Total	**28,530**	**100**	**16,795**	**100**	**157,328**	**100**

Table 5: Daily-count Pearson r between POLECAT and each GENOME variant (art = article_date, evt = event_date, dd = dedup) for the 14 dyads with at least 100 events in both datasets, broken down by PLOVER quad category (Feb–Jun 2024). Dyads ordered by combined volume. "n/a" indicates that one or both sides had zero events in the corresponding quad.

	Overall			V-Coop			M-Coop			V-Conf			M-Conf		
Dyad	art	evt	dd	art	evt	dd	art	evt	dd	art	evt	dd	art	evt	dd
Russia–Ukraine	+0.24	+0.20	+0.15	+0.12	+0.29	−0.04	+0.07	+0.14	+0.15	+0.24	+0.30	+0.18	+0.21	+0.11	+0.11
Iran–Israel	+0.86	+0.46	+0.58	+0.32	+0.36	−0.02	+0.00	+0.18	+0.21	+0.91	+0.62	+0.58	+0.81	+0.42	+0.47
Israel–United States	+0.38	+0.26	+0.33	+0.03	+0.09	+0.08	+0.13	+0.11	+0.03	+0.30	+0.18	+0.25	+0.18	+0.17	+0.31
Russia–United States	+0.61	+0.56	+0.47	−0.04	−0.07	−0.07	+0.15	+0.08	+0.09	+0.46	+0.41	+0.47	+0.37	+0.31	+0.08
Israel–Syria	+0.20	+0.23	+0.11	n/a	n/a	n/a	+0.06	+0.06	+0.06	+0.06	+0.06	+0.06	+0.24	+0.26	+0.10
China–United States	+0.49	+0.50	+0.45	−0.07	−0.06	−0.07	−0.07	+0.08	+0.11	+0.31	+0.36	+0.39	+0.24	+0.17	+0.23
Israel–Lebanon	+0.43	+0.35	+0.25	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	+0.35	+0.23	+0.14	+0.40	+0.36	+0.28
Iran–United States	+0.65	+0.65	+0.55	n/a	n/a	n/a	+0.03	+0.22	+0.20	+0.56	+0.48	+0.51	+0.53	+0.61	+0.36
Ukraine–United States	+0.46	+0.29	+0.07	+0.60	+0.49	−0.06	+0.32	+0.26	+0.18	−0.05	+0.04	−0.00	−0.04	−0.04	−0.04
United States–Yemen	+0.16	+0.08	+0.32	n/a	n/a	n/a	−0.02	−0.02	−0.01	+0.10	+0.03	+0.03	+0.18	+0.10	+0.35
China–Philippines	+0.59	+0.33	+0.35	+0.28	+0.28	+0.37	−0.03	−0.03	−0.03	+0.51	+0.45	+0.38	+0.25	+0.13	+0.07
North Korea–South Korea	+0.70	+0.49	+0.56	+0.26	+0.16	−0.04	−0.04	−0.04	−0.04	+0.51	+0.36	+0.41	+0.49	+0.47	+0.36
Russia–United Kingdom	+0.64	+0.70	+0.44	−0.03	−0.03	−0.03	n/a	n/a	n/a	+0.31	+0.27	+0.29	+0.74	+0.78	+0.51
United Kingdom–Yemen	+0.11	+0.14	+0.42	n/a	n/a	n/a	−0.01	−0.01	−0.01	−0.04	−0.04	−0.04	+0.12	+0.14	+0.47

FNLP412@EEUCA 2026: Understanding Toxic Behavioral Intent in Gaming Chat Logs using Transfer Learning and Synthetic Data Augmentation

Mihai Radu Rădulescu
University of Bucharest
mihai-radu.radulescu@s.unibuc.ro

Abstract

Our paper explores several machine learning methods for detecting toxic language in gaming-related chat utterances. We start with the GameTox dataset, perform some data preprocessing and augment the minority classes with LLM-generated synthetic data. We then set a baseline using a classic Logistic Regression model and continue to explore several approaches to surpassing it, by leveraging the leading multilingual transformer models (XLM-RoBERTa and DeBERTa-V3) to classify our test data. We achieve a top result of 0.6725 Macro-F1 (2nd place on shared task leaderboard) using a MDeBERTa-V3 model which we pretrained on the Jigsaw dataset for 1 epoch and then fine-tuned on our GameTox data for 5 epochs.

1 Introduction

Toxic behavior and messaging in online multiplayer games is widespread. This paper explores designing and training several machine learning models to help implement robust detection methods to ensure user safety. We have written this as part of a shared task (Thapa et al., 2026) within the EEUCA workshop (Hürriyetoğlu et al., 2026) (ACL 2026). Our results show that, of several approaches considered, the best performing one is a DeBERTa-V3 model which we pretrained on a more generic toxicity dataset (Kivlichan et al., 2020), achieving a Macro-F1 score of 0.6725 (2nd place on the final leaderboard).

2 Related Work

Since the EEUCA shared task is still ongoing, there is no established SOTA or baseline regarding the GameTox dataset. In the original GameTox paper (Naseem et al., 2025), the authors identify Joint BERT (Chen et al., 2019) as the most promising model, with a 0.89 accuracy score. On a similar gaming chat dataset, CONDA (Weld et al., 2021), the winning paper (Jia et al., 2024) describes the BRAR model, which implements attention residuals, slot filling and label forcing. ToxBuster (Yang et al., 2023) is a model which reaches an accuracy of 0.82 by creating context from groups of sequential chat messages (not possible for our dataset). ToXCL (Hoang et al., 2024) is a framework for toxicity detection using Contrastive Loss to address dataset class imbalance issues with good results. However, its applicability to gaming chat is not as performant.

3 Method

3.1 Dataset

The shared task is based on GameTox (Naseem et al., 2025), a dataset of chat utterances collected from the game "World of Tanks". The dataset contains messages labeled on a scale from 0 to 5 (an annotation schema shared with Crisishatemm (Bhandari et al., 2023)), with the labels indicating the following mapping:

- 0 - Non-toxic[4]
- 1 - Insults and flaming[5]
- 2 - Other offensive texts[6]
- 3 - Hate and harassment[7]
- 4 - Threats[8]
- 5 - Extremism[9]

The messages are generally very short, with an average length of 13.84 characters and a 75th percentile of 18 characters (Figures 2 and 3). This brevity presents additional challenges for effective training and detection. Another noteworthy characteristic is the multilingual aspect of the dataset: messages are mostly in English, but also in Russian, German, Polish, French, and other languages, making it essential to use models capable of handling multiple languages.

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 96–103

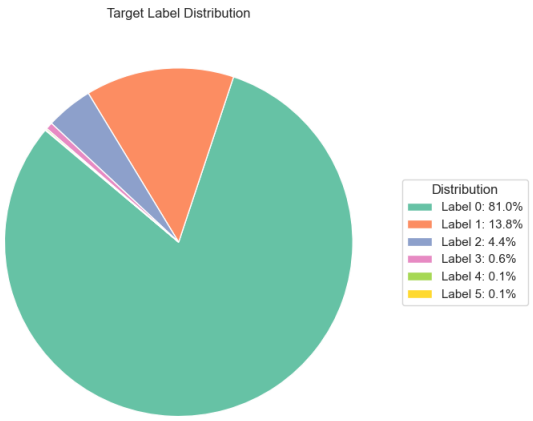


Figure 1: Label distribution

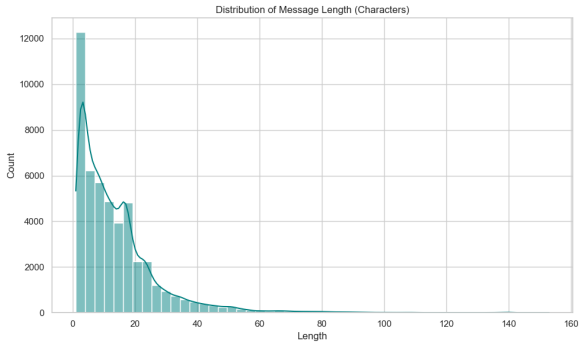


Figure 2: Message length

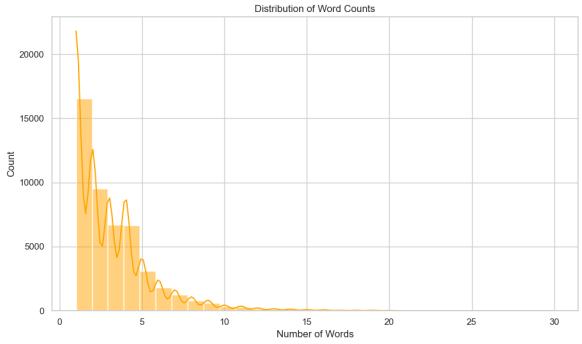


Figure 3: Word count

The dataset is highly imbalanced, with the majority of messages being labeled as non-toxic (81%), down to 0.06% labeled 5 - only 27 samples - Figure 1. To address this imbalance, we employed several preprocessing techniques to clean and normalize the text data, as detailed below.

3.2 Preprocessing

After exploring the dataset, we put together the following steps for preprocessing the message data:

1. **URL removal**: we replaced all URLs with a [URL] tag, to avoid URL content influencing our training.
2. **Lowercase**: we converted all text to lowercase to ensure consistency.
3. **User mentions**: we replaced all @user mentions/tagging with a common [USER] tag.
4. **Repetition normalization**: we reduced characters repeated more than twice. This keeps the emphasis, but removes redundancy. (e.g. 'looooool' -> 'lool')
5. **Slang map**: we applied a mapping for common slang, abbreviations, typos and obfuscations, most of which we manually extracted from the messages in classes 3-5.
6. **Augmentation**: we generated 50 synthetic data samples for classes 4-5, where our original dataset had very few examples. We used Gemini 3 Pro with the prompt in Appendix A to generate these samples.

3.3 Models

3.3.1 M1: Basic baseline

Since the shared task does not provide a baseline, we implemented a simple Logistic Regression model using TF-IDF (Spärck Jones, 1972) features. This model serves as a reference point for evaluating the performance of more complex models in our next attempts.

Our baseline Macro-F1 score, using LR, was 0.5046 on the validation set and 0.4665 on the test set.

3.3.2 M2: XLMR (0.3B parameters)

As a next step, we fine-tuned the XLM-RoBERTa-base transformer model (Conneau et al., 2019) on the GameTox dataset. We chose XLMR because it is a multilingual model pre-trained on 100 languages, making it suitable for our task which includes messages in multiple languages. We used the HuggingFace Transformers library for implementation, and trained the model for 5 epochs with

a learning rate of 2e-5 and a batch size of 16. Prior to training, we performed a 5-fold stratified split of the training data (ensuring class distribution is maintained in each fold), and used 4 folds as training data and 1 fold as validation data, over 5 iterations. We then averaged the results across the folds to obtain a robust estimate of model performance.
Our fine-tuned XLMR model achieved a Macro-F1 score of 0.5297 on the validation set, and 0.5752 on the test set. At this point, even while this was the leading model in the shared task leaderboard, we aimed to push performance even higher.

3.3.3 M3: XLMR pretrained on Jigsaw

To improve our score, we next adopted a transfer learning approach. We first pre-trained the XLM-RoBERTa-base model on the Jigsaw Multilingual Toxic Comment Classification dataset (Kivlichan et al., 2020), which contains a large number of toxic comments in multiple languages. This pre-training step helps the model learn general features of toxic language, which can then be fine-tuned on the more specific GameTox dataset. After pre-training on Jigsaw for 1 epoch, we fine-tuned the model on GameTox for 5 epochs using the same hyperparameters as before. Since Jigsaw is a much larger dataset, its pre-training should improve our model's sensitivity to detect toxicity in written form. However, Jigsaw is not specifically focused on gaming chat, so we transferred this knowledge to the specifics of our GameTox dataset in the subsequent training step.
The result was a Macro-F1 score of 0.4755 on the test set, which was lower than our previous XLMR model. This indicates that while transfer learning can be beneficial, it may not always lead to improved performance, especially if the pre-training dataset is not closely aligned with the target task.

3.3.4 M4: MDeBERTA-V3 (0.27B parameters) pretrained on Jigsaw

Since the first transfer learning attempt attained lower performance than expected, our next approach involved switching to a different transformer architecture: Microsoft DeBERTa-V3 (He et al., 2021), which has shown strong performance on various NLP tasks, and is still multilingual. We followed the same transfer learning approach as before, pre-training the DeBERTa-V3-base model on the Jigsaw dataset for 1 epoch, and then fine-tuning it on GameTox for 5 epochs. We used a learning rate of 2e-5 and a batch size of 16. Specifically for the pretraining step, we increased the maximum input length to 256 tokens (from the default 64), since many Jigsaw comments are longer and would be otherwise truncated. This allows the model to capture more context from longer comments, which could be beneficial for toxicity detection.
Our model achieved a Macro-F1 score of 0.6258 on the validation set, and 0.6725 on the test set. It is our best-scoring model so far, and remained our top contender, but we tried a few other approaches to see if we can improve on our performance.

3.3.5 M5: Ensemble prediction

Since our models are showing different strengths and weaknesses (higher precision vs higher recall), we decided to combine their predictions using an ensemble approach. We used a weighted average of the predicted probabilities of our **M3** and **M4** models, with weights determined based on their validation performance (F1 score).
This resulted in a Macro-F1 of 0.6165 - not bad, but lower than the simple predictions of **M4**.

3.3.6 M6: Pseudo label augmentation of M4

Our last improvement idea involved another type of data augmentation, where we used our leading model as a "teacher" model to train a fresh "student" model. We again used the DeBERTa-V3-base model as the student, and trained it to mimic the predictions of the **M4** model on the training data. This involved merging the original training data (triplicated, to keep it dominant) with the pseudo-labeled data generated by the teacher model.
We then trained the student model on this merged dataset for 3 epochs, obtaining a Macro-F1 score of 0.5927 on the test set, and 0.6114 after threshold optimization (see 3.3.8).

3.3.7 M7: Pseudo label augmentation of M5

The final model we experimented with was trained similarly to **M6**, but on the predictions generated by **M5** instead of **M4**. The result was a Macro-F1 score of 0.6311, still below our **M4** model.

3.3.8 Threshold Optimization

At inference time, by default, calculating a prediction would mean picking the class label with the highest predicted probability. Given the imbalanced nature of our dataset, we decided to optimize the decision thresholds for each class to improve recall on the less represented classes. We performed a grid search over [0.10, 0.95] threshold values

for each class and determined the optimal thresholds to maximize the F1 score. Next, during the prediction generation, we implemented a severity waterfall, and checked each predicted probability against each class threshold, in descending order (5 to 0). If the probability exceeded a threshold, we chose this class as our final prediction and skipped to the next sample.
Some models reacted favorably to the threshold tuning, while others did not. Where not explicitly specified, we chose the results with the higher Macro-F1 score of the two.

4 Future Work

Strong class imbalance, such as found within our dataset, has been shown to be well handled by Focal Loss (Lin et al., 2018) and Adversarial Weight Perturbation (Wu et al., 2020) techniques.
Another avenue worth exploring is using Joint Intent Classification and Slot Filling (Chen et al., 2019) - we are ignoring slots in our implementation at this point and haven't explored this further.
On the data augmentation front, there are several approaches which could yield interesting results. One is to go beyond synthetic data generation using LLMs, and explore backtranslation (Sugiyama and Yoshinaga, 2019) or other synthetic generation methods.
Similar to how we used the Jigsaw dataset, one could search for another dataset to use for pretraining the transformers during the first epoch - perhaps the CONDA dataset (Weld et al., 2021) or other toxicity-focused datasets.
Model distillation between architectures (e.g. using XLMR as a teacher and DeBERTa as student, or using a monolingual model as the student for a multilingual teacher) was also not explored due to time constraints, and might be a promising avenue.
Finally, there are other multilingual transformer architectures, such as BERT or other sizes of XLMR/MDeBERTa, which could yield different results.

5 Conclusion

Of all the various model and parameter combinations we experimented with, our best result was obtained using model **M4** (3.3.4) - a Macro-F1 score of 0.6725 in the shared task competition, which placed us at position #2 in the leaderboard at the end of the test phase.
A comparative table detailing the performance obtained with the various implemented models can be seen in Table 1.

Our present research on the topic of toxicity in online gaming chat shows that DeBERTa-V3 is a capable multilingual model for this task, and pretraining it on a toxicity-related dataset (even if not gaming focused) surpasses other implementations using straightforward training or other transformer models.

Limitations

The transformers we have used for this task, such as XLM-RoBERTa and DeBERTa-V3, are relatively large models (approx. 0.3B parameters) that require significant computational resources for both training and inference. On a MacBook M4 Pro, training an epoch can take one hour, and the chances for improving hyperparameters or experimenting with model modifications are limited by time. Additionally, while these models are pretrained on multiple languages, their performance may vary significantly across different languages present in the dataset - this is a limitation we did not explore in this project. When we pretrained the XLMR model (M3 3.3.3), we used the same max length of 64 as for the GameTox dataset. However, Jigsaw text is much more ample, thus a max length of 256 would have been more appropriate, ensuring less content is truncated. Larger models (such as DeBERTa-V3-large or XLM-RoBERTa-XL) could potentially yield better results, but their training would not have been feasible within our resource constraints.

Ethical Considerations

The datasets processed in this research contain highly offensive, hateful, and disturbing language, including threats, hate speech, severe profanity and extremism. Our models are discriminative (classifiers) rather than generative; they do not generate new toxic text, but serve only to flag existing toxicity. While this technology is designed to foster safer online communities, we recognize the dual-use potential of automated moderation tools. If misapplied, such systems could be used for censorship or surveillance. We emphasize that our models are intended to assist human moderators by flagging potential violations, not to act as autonomous decision makers without human oversight.

	M1	M2	M3	M4	M5	M6	M7
F1 Macro	0.4665	0.5752	0.4755	0.6725	0.6165	0.6311	0.6114
Accuracy	0.8804	0.8739	0.8448	0.8992	0.8964	0.8794	0.8954
Precision	0.4561	0.526	0.4162	0.6636	0.6135	0.5783	0.5695
Recall	0.5029	0.6521	0.6588	0.6846	0.6237	0.7194	0.6844

Table 1: Training results

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.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *Preprint*, arXiv:1902.10909.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Nhat M. Hoang, Xuan Long Do, Duc Anh Do, Duc Anh Vu, and Luu Anh Tuan. 2024. Toxcl: A unified framework for toxic speech detection and explanation. *Preprint*, arXiv:2403.16685.

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)*.

Yuanzhe Jia, Weixuan Wu, Feiqi Cao, and Soyeon Caren Han. 2024. In-game toxic language detection: Shared task and attention residuals (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13):16238–16239.

Ian Kivlichan, Jeffrey Sorensen, Julia Elliott, Lucy Vasserman, Martin Görner, and Phil Culliton. 2020. Jigsaw multilingual toxic comment classification. `https://kaggle.com/competitions/jigsaw-multilingual-toxic-comment-classification`. Kaggle.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal loss for dense object detection. *Preprint*, arXiv:1708.02002.

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.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Amane Sugiyama and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44, Hong Kong, China. Association for Computational Linguistics.

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)*.

Henry Weld, Guanghao Huang, Jean Lee, Tongshu Zhang, Kunze Wang, Xinghong Guo, Siqu Long, Josiah Poon, and Soyeon Caren Han. 2021. Conda: a contextual dual-annotated dataset for in-game toxicity understanding and detection. *Preprint*, arXiv:2106.06213.

Dongxian Wu, Shu tao Xia, and Yisen Wang. 2020. Adversarial weight perturbation helps robust generalization. *Preprint*, arXiv:2004.05884.

Zachary Yang, Nicolas Grenon-Godbout, and Reihaneh Rabbany. 2023. Towards detecting contextual real-time toxicity for in-game chat. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

A Synthetic data prompt

Prompt used for synthetic data generation (using Gemini 3 Pro):

"I am training an NLP model to detect 'Extremism' and 'Threats' in online gaming chat (World of Tanks).

Please generate:

50 distinct examples of 'Extremism' (Class 5): These should focus on radical ideology, promotion of terrorist groups, or extreme political hate, but written in 'gaming chat style' (short, lowercase, maybe some typos).

50 distinct examples of 'Threats' (Class 4): These should be physical threats of violence (e.g., 'I will find where you live', 'kill yourself'), distinct from simple insults.

Format the output as a CSV with columns: message,label (where label is 5 for Extremism and 4 for Threats)."

B Word clouds of each label class

Figure 4: 0 - Non-toxic

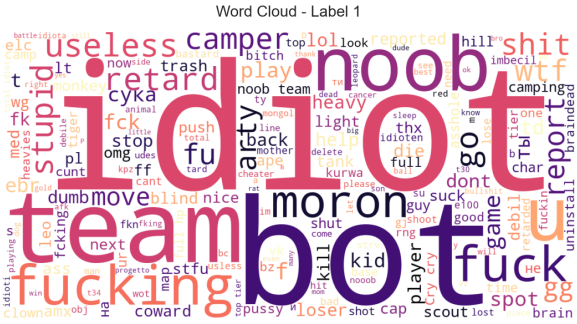


Figure 5: 1 - Insults and flaming

Figure 6: 2 - Other offensive texts

Figure 7: 3 - Hate and harassment

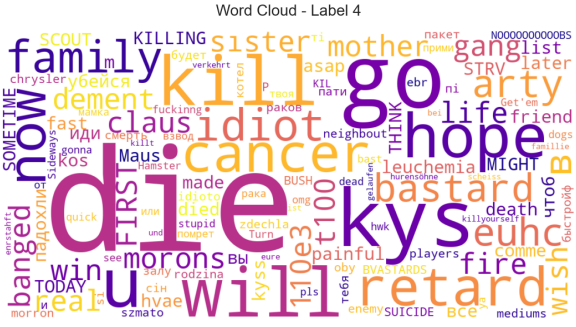


Figure 8: 4 - Threats

Figure 9: 5 - Extremism

wangkongqiang@EEUCA 2026: Understanding Toxic Behavioral Intent in Gaming Chat Logs

Kongqiang Wang
School of Information Science and Engineering, Yunnan University, Yunnan Baiyao Street, 650500, Kunming, Yunnan, China.
wangkongqiang60@gmail.com

Peng Zhang
School of Information Science and Engineering, Yunnan University, Yunnan Baiyao Street, 650500, Kunming, Yunnan, China.
zpp1219@gmail.com

Qingli Tan
College of Ecology and Environment, Yunnan University, Yunnan Baiyao Street, 650500, Kunming, Yunnan, China.
tanqingli@stu.ynu.edu.cn

Abstract

Our team was interested in content classification and labeling from toxicity detection of gaming chat logs in online gaming communities. We joined the shared task on Understanding Toxic Behavioral Intent in Gaming Chat Logs@EEUCA with ACL 2026. In this task, our goal is to assign a content classification label to player's utterance (e.g., Hate and Harassment, Threats, Non-toxic). The objective is to develop systems that can classify the intent of a player's utterance. The dataset for this task will have five labels: Non-toxic (0), Insults and Flaming (1), Other Offensive Texts (2), Hate and Harassment (3), Threats (4) and Extremism (5). The performance will be ranked by F1-score (Macro). The task utilizes 53,000 game chat utterances from *World of Tanks*. Our group used a supervised learning method on multiple pre-trained models and finetuning Qwen2 LLMs. The best result on the test set for shared task were Macro F1 score of 0.5776, Accuracy 0.9075, Precision (Macro) 0.6847, and Recall (Macro) 0.5343 from finetuning qwen2_7B LLM method, ranking 8th among all teams. The complete code of this entire project can be found at our GitHub address[1].

1 Introduction

First of all, let's introduce the overview of shared task on Understanding Toxic Behavioral Intent in Gaming Chat Logs@EEUCA with ACL 2026 (Thapa et al., 2026). The prevalence of toxic behavior in online gaming communities necessitates robust detection methods to ensure user safety. This shared task focuses on detecting toxicity in game chat logs, specifically using the *GameTox* dataset, which captures the complex relationship between user intent and specific linguistic features. In this context, the distinction between Toxic and Non-toxic becomes blurred, as gaming chat logs straddle the line between satire and offense, challenging researchers and platforms alike to navigate the complexities of online content moderation. As one label generally fails to encompass multiple aspects of linguistics, this shared task classifies gaming chat logs on five aspects: Non-toxic (0), Insults and Flaming (1), Other Offensive Texts (2), Hate and Harassment (3), Threats (4) and Extremism (5).

The objective is to develop systems that can classify the intent of a player's utterance (e.g., Hate and Harassment, Threats, Non-toxic). The task utilizes 53,000 game chat utterances from *World of Tanks*. This dataset has been published in NAACL 2025 (Naseem et al., 2025).

2 Background

2.1 Toxicity detection in online games

The internet has become a common platform for everyone to share their ideas and opinions. The user has freedom to post whatever he/she likes in social networking and blogging sites. However, sometimes the content when directed towards certain group of individuals with an intention to incite hate or discrimination, causes a turmoil in the society. Such content is known as hate speech. Hate speech (Bhandari et al., 2023) can be a serious problem to peace and harmony in the society. There are instances where hate speech have led to social unrest and extremism. Thus, hate speech in the internet needs to be monitored (Parihar et al., 2021). In this context, researchers have proposed various frameworks and datasets for automated toxicity detection in online games. (Blackburn and Kwak, 2014) utilized crowdsourced in-game user reports from League of Legends (LoL) for toxic

[1] https://github.com/WangKongQiang/EEUCA2026_Understanding_Toxic_Behavioral_Intent_in_Gaming_Chat_Logs

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 104–111

behavior detection by extracting 534 features from in-game performance, user reports, and chat logs and employed the Random Forest Classifier for toxicity detection. (Stoop et al., 2019) used a similar approach for data collection and introduced the RNN-based HaRe framework that tracked toxicity estimates for each user individually, updated the estimate with every new utterance, concatenated all of the utterances of each user, and classified the combined text. (Märtens et al., 2015) proposed a novel lexicon-based annotation strategy for game chat toxicity detection to devise the DotAlicious dataset consisting of chat replays from 12,923 Defense of the Ancients (DOTA) matches.

2.2 Toxicity and Hate speech datasets

Detection of hate speech and toxicity in online environments has seen significant progress in recent years. (Oz et al., 2023) aimed to explore the perceptions, concerns, and strategies of LGBTQ social media activists in Turkey. Through semi-structured interviews with 20 LGBTQ social media activists, This study investigated how they navigate cultural and political challenges and utilize social media for activism purposes. (Thapa et al., 2024) addressed the need for effective hate speech moderation in contemporary digital discourse, the multimodal hate speech event detection shared task (Thapa et al., 2023) made its debut at russia-ukraine crisis period. (Qian et al., 2019) introduced two labeled hate speech datasets collected from Reddit (22k comments) and Gab (33k comments) containing manually-written intervention responses. (Wijesiriwardene et al., 2020) focused on toxic behaviors among youngsters and introduced ALONE, a dataset for toxic behavior detection among adolescents on Twitter, consisting of 16,901 tweets in 688 interactions and labeled for toxic vs non-toxic classes. (Founta et al., 2018) analyzed abusive behavior on Twitter by releasing a dataset of 80,000 tweets annotated for seven labels: offensive, abusive, hate speech labels, aggressive, cyberbullying, spam, and normal. (Mathew et al., 2020) introduced HateXplain, a dataset for explainable hate speech detection, consisting of 20,148 posts collected from Twitter and Gab annotated for three classes: hate, offensive, and normal, alongside target communities within hate. They further annotated the sections of the post that guide the labeling rationale. (Zampieri et al., 2019) released an offensive language detection dataset comprising 14,100 tweets categorizing offensive language and its targets, consisting of offensiveness detection with three target classes: Individual, Group, and Other. To discern multiple aspects within cyberbullying, (Salawu et al., 2021) curated an extensive dataset for cyberbullying detection comprising 62,587 tweets annotated for multiple aspects including Bullying, Profanity, Sarcasm, Threat, and Spam.

3 Dataset

In this section, we describe various aspects of task dataset including data collection, utterance annotation, and dataset statistics. Task dataset comprises 42,963 text utterance that encompass different intent content relevant to the chat recordings from the game *World Of Tanks*. Organizers collected 53,000 utterances from the WoT Record database, which stores chat recordings from the game *World Of Tanks*. Among these utterances, 42,963 samples contained only English text, and the rest were in other languages or a code-mixed format. The 42,963 English utterances were annotated for intent, and all samples were annotated for slot filling by converting the code-mixed samples to English by using Google Translate[2]. Organizers converted all text to lowercase to ensure uniformity. They removed all duplicated text from the corpus, which may otherwise create biases. Further, they removed all user identifiers such as usernames and gamer tags to preserve the privacy of players.

3.1 Utterance Annotation

Each utterance was labeled to one of 6 labels:Non-toxic (0), Insults and Flaming (1), Other Offensive Texts (2), Hate and Harassment (3), Threats (4), and Extremism (5). Non-toxic if toxicity was not present and one of the five toxicity labels if toxicity was present. Utterance annotation for each label are mentioned below.

Hate and Harassment: Utterances with the presence of identity-based hate or harassment (e.g., racism, sexism, homophobia).

Threats: Utterances with threats of violence, physical harm to another player, employee, or property, terrorism, or releasing a player's real-world personal information (e.g., doxing).

Extremism: Utterances with extremist views (e.g., white supremacy), attempts to groom or recruit for an extremist group, or repeated sharing of

[2]https://translate.google.com

Figure 1: Wordcloud of words in each intent label.

political or religious beliefs.

Insults and Flaming: Insults or attacks on another player or team (not based on player or team's real or perceived identity)

Other Offensive Texts: Any message not covered in the aforementioned categories that is offensive or harms a player's reasonable enjoyment of the game.

Non-Toxic: Utterances without any toxicity.

3.2 Dataset Statistics

Table 1: Dataset statistics for GameTox. The data consists of 42,963 samples for player's utterance toxicity detection task in online gaming communities.

Task	Label	#Samples	%
Intent Classification	Non-Toxic	34679	80.71
	Insults and Flaming	6049	14.07
	Other Offensive Texts	1885	4.38
	Hate and Harassment	274	0.63
	Threats	53	0.12
	Extremism	23	0.053
Task	**Token**	**%**	
Slot Classification	Other	67.17	
	Verb	15.51	
	Game Slang	7.72	
	Toxic	9.59	

Table 1 (*Upper*) provides the class distribution of intent across the 42,963 English utterances, and Table 1 (*Down*) provides the slot filling distribution across all utterances. Most utterances are non-toxic in nature and a notable data imbalance is present. However, this is in line with real world data distributions, where extremely toxic labels such as Hate and Harassment, Threats, and Extremism are often moderated or automatically suppressed. Figure 1 illustrates the word cloud for all intent labels.

4 System Overview

4.1 Fine-tuning Pre-trained Models

Introduction. In recent years, with the rapid development of deep learning technology, large-scale pre-trained models have achieved remarkable results in fields such as natural language processing, computer vision, and multimodal learning. Compared with traditional models trained from scratch, pre-trained models can learn rich semantic representations and general knowledge by pre-training on large-scale general corpora, thereby significantly improving the performance and training efficiency of downstream tasks. However, the knowledge learned by pre-trained models on general corpora often has strong generalization, while specific tasks usually have obvious domain characteristics. Therefore, directly applying pre-trained models to downstream tasks often fails to achieve the best results. To solve this problem, researchers usually adopt the fine-tuning strategy, that is, on the basis of the pre-trained model, further optimize the model parameters using the data of specific tasks, so that it can better adapt to the target task. In this study, to enhance the model's performance in the task of toxic behavioral intent analysis, a pre-trained language model was adopted as the base model and fine-tuned in combination with specific task data (*GameTox* dataset), enabling the model to effectively learn the semantic relationship between gaming chat utterance expressions and their underlying intents.

The classifier in the pre-trained model uses a transformer based classifier. The specific pre-trained models of the classifier are shown in the Table 2.

Table 2: pre-trained model classifier structure for gaming chat content classification.

Model	Batch Size	Num Epochs	Learning Rate
albert/albert-base-v2	32	5	2e-5
google-bert/bert-base-uncased	32	5	2e-5
nghuyong/ernie-2.0-large-en	32	5	1e-5
FacebookAI/roberta-large-mnli	32	5	2e-5
cambridgeltl/trans-encoder-bi-simcse-roberta-large	32	5	1e-5

The Principle of Fine-tuning Pre-trained Models. Pre-trained models typically use large-scale corpora for self-supervised learning, such as tasks like language model prediction, masked language modeling, or autoregressive modeling, thereby learning common language representations. After pre-training, the model parameters already contain a large amount of language knowledge and semantic information. Therefore, in downstream

tasks, only a small amount of labeled data is needed to achieve good performance.

The basic idea of fine-tuning is to introduce supervisory signals from downstream tasks on the basis of the parameters of the pre-trained model and further optimize the model parameters through the gradient descent algorithm. Let the parameters of the pre-trained model be θ, Given the downstream task training dataset $\mathcal{D} = (x_i, y_i)_{i=1}^{N}$, Where x_i represents the input text and y_i represents the corresponding label, then the model training objective can be expressed as:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell\left(f(x_i; \theta), y_i\right) \tag{1}$$

Here, $f(\cdot)$ represents the model prediction function, and $\ell(\cdot)$ is the loss function (such as cross-entropy loss). By minimizing this loss function, the model can gradually adapt to the data distribution of a specific task, thereby enhancing the prediction performance.

In practical applications, fine-tuning typically includes the following two methods:

- Full Fine-tuning: Update all the parameters of the pre-trained model to enable it to fully adapt to the target task.
- Parameter-efficient Fine-tuning: Only update some parameters or introduce additional lightweight modules to reduce training costs, such as Adapter, LoRA and other methods.

In this study, based on the characteristics of the task and the availability of computing resources, the pre-trained model was trained using a parameter-efficient fine-tuning strategy.

Input Data Construction. First of all, the original data needs to be converted into an input format that the model can handle. For text tasks, the following steps are usually required:

- Text Cleaning and Preprocessing: Remove irrelevant symbols or abnormal characters;
- Word Segmentation and Encoding: Use the tokenizer corresponding to the pre-trained model to convert the text into a token sequence;
- Input Sequence Construction: Ultimately, the input text will be represented as a sequence of token ids and input into the pre-trained model for feature encoding.

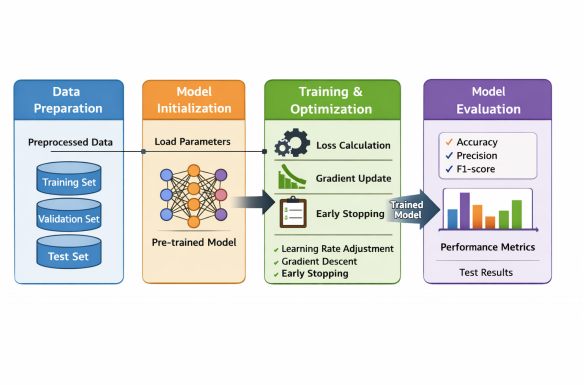


Figure 2: The framework diagram of the fine-tuning pre-trained classification model.

Task Structure Design. During the fine-tuning process, it is necessary to design the corresponding prediction structure based on the specific task. For instance, in the task of toxic behavioral intent analysis in gaming chat logs, the model needs to simultaneously identify toxic behavioral intent categories as well as the corresponding player's utterance information. Therefore, a model is usually composed of the following parts:

- Pre-trained Encoding Layer: Used for extracting semantic representations of text;
- Task-specific Layer: For example, the classification layer or the sequence labeling layer;
- Output Layer: Generate the final prediction result.

By adding task-related structures at the top of the pre-trained model, the model's adaptability to specific tasks can be effectively enhanced.

The overall architecture diagram of the fine-tuing pre-trained model is shown in the Figure 2.

Design of Loss Function. During the training process, it is necessary to select an appropriate loss function based on the type of task. For classification tasks, the cross-entropy loss function is usually adopted.

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(p_i) \tag{2}$$

Here, C represents the number of categories, y_i is the true label, and p_i is the model prediction probability. By minimizing the loss function, the consistency between the model's prediction results and the true labels can be gradually improved.

4.2 Hard Voting Mechanism

The hard voting mechanism is a common model fusion strategy in ensemble learning, mainly used for classification tasks. Its basic idea is: multiple base learners make predictions on the same sample respectively, and then determine the final prediction category through majority voting.

The Principle of Hard Voting Mechanism. Assume that the ensemble model consists of M base classifiers: $h_1(x), h_2(x), \ldots, h_M(x)$, where $h_i(x)$ denotes the prediction of the i-th classifier for input sample x. The final prediction of the hard voting ensemble is determined by majority voting:

$$\hat{y} = \arg\max_{c \in C} \sum_{i=1}^{M} I(h_i(x) = c) \tag{3}$$

where C represents the set of all possible classes and $I(\cdot)$ is an indicator function defined as:

$$I(h_i(x) = c) = \begin{cases} 1, & \text{if } h_i(x) = c \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

In a weighted hard voting scheme, each classifier is assigned a weight w_i, and the final prediction can be written as:

$$\hat{y} = \arg\max_{c \in C} \sum_{i=1}^{M} w_i \cdot I(h_i(x) = c) \tag{5}$$

In our experiment, the weights of each classifier were the same.

4.3 Fine-tuning of the Qwen2 Large Language Model (LLM)

Qwen2 is an open-source large language model (LLM) developed by the Tongyi Qianwen team and created by Alibaba Cloud's Tongyi Lab. Using Qwen2 as the base large language model (LLM) and achieving high-accuracy text classification through instruction fine-tuning is an introductory task for learning the fine-tuning of large language models (LLMs).

Instruction fine-tuning is a process of further training an LLMs on a dataset composed of (instruction, input, output) combine pairs. Among them, the instructions represent the human instructions of the model, the input represent the raw data content from specific dataset, and the output represents the expected output that follows the instructions. This process helps bridge the gap between the next word prediction target of LLMs and the goal of users to have LLMs follow human instructions.

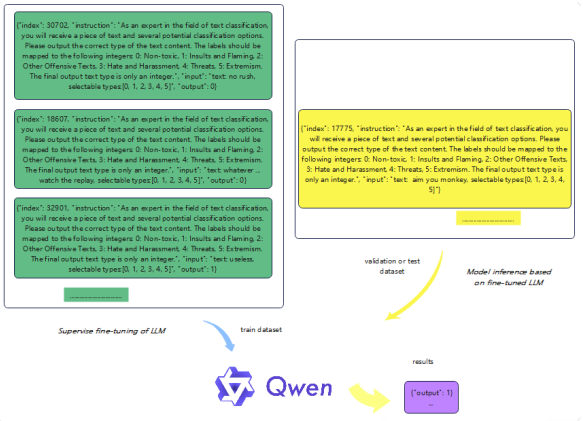


Figure 3: The framework diagram of the fine-tuning qwen2 classification LLM and model inference.

In this toxic behavioral intent classification task in gaming chat logs, we will use the Qwen2-1.5B-Instruct and Qwen2-7B-Instruct model to perform instruction fine-tuning tasks on the dataset, while using SwanLab [3] for monitoring and visualization. The following presents three demonstration formatted data samples for fine-tuning LLM data in the train dataset. Our training task is to ensure that the fine-tuned large language model (LLM) can predict the correct output based on the prompt words composed of text and selectable types.

The complete process of fine-tuning the Qwen2 large language model (LLM) using the train dataset and conducting model inference on validation or test dataset with the fine-tuned large language model (LLM), as shown in Figure 3.

5 Results and Analysis

For results obtained by our fine-tuning pre-trained models and fine-tuning Qwen2 LLM methods on the validation dataset and test dataset are shown in Table 3 and Table 4 respectively. roberta-RNN indicates the addition of a layer of recurrent neural network (RNN) after *FacebookAI/roberta-large-mnli* model, which is the LSTM layer. roberta-cnn indicates adding a convolutional neural network (CNN) layer after *FacebookAI/roberta-large-mnli*, which is the Conv2d layer. bagging refers to the model ensemble decision-making method that employs hard voting mechanism in the four models: roberta, roberta-RNN, simcse-roberta-lstm, roberta-lstm-gru.

[3] https://swanlab.cn

Table 3: The results obtained by our fine-tuning pre-trained models and fine-tuning Qwen2 LLM methods for toxic behavioral intent classification task on the validation dataset.

Pre-trained Model	**Recall (Macro)**	**Precision (Macro)**	**F1 (Macro)**	**Accuracy**
albert/albert-base-v2	0.3416	0.3578	0.348	0.8897
google-bert/bert-base-uncased	0.4118	0.4386	0.4212	0.8975
nghuyong/ernie-2.0-large-en	0.357	0.3737	0.3638	0.9005
FacebookAI/roberta-large-mnli	0.3997	0.4355	0.4131	0.9001
cambridgeltl/trans-encoder-bi-simcse-roberta-large	0.3593	0.4522	0.3702	0.8985
roberta-RNN	0.3624	0.3704	0.366	0.8999
roberta-cnn	0.3502	0.3684	0.3584	0.8975
roberta-lstm-gru	0.3605	0.3638	0.3619	0.8985
simcse-roberta-lstm	0.4047	0.4771	0.4236	0.8985
bagging	0.3562	0.3745	0.3639	0.9018
Large Language Model	**Recall (Macro)**	**Precision (Macro)**	**F1 (Macro)**	**Accuracy**
qwen2_1.5B	-	-	-	-
qwen2_7B	-	-	-	-

Table 4: The results obtained by our fine-tuning pre-trained models and fine-tuning Qwen2 LLM methods for toxic behavioral intent classification task on the test set.

Pre-trained Model	**Recall (Macro)**	**Precision (Macro)**	**F1 (Macro)**	**Accuracy**
albert/albert-base-v2	0.3441	0.3722	0.3553	0.8973
google-bert/bert-base-uncased	0.4109	0.4365	0.4205	0.8928
nghuyong/ernie-2.0-large-en	0.3454	0.3634	0.3533	0.899
FacebookAI/roberta-large-mnli	0.3853	0.4233	0.3992	0.9014
cambridgeltl/trans-encoder-bi-simcse-roberta-large	0.3596	0.5416	0.3785	0.9025
roberta-RNN	0.3547	0.3685	0.3609	0.9003
roberta-cnn	0.3496	0.3699	0.3587	0.9005
roberta-lstm-gru	0.3548	0.3624	0.3584	0.8997
simcse-roberta-lstm	0.4	0.4761	0.4234	0.9031
bagging	0.358	0.5424	0.372	0.9038
Large Language Model	**Recall (Macro)**	**Precision (Macro)**	**F1 (Macro)**	**Accuracy**
qwen2_1.5B	0.4429	0.4646	0.4525	0.9057
qwen2_7B	0.5343	0.6847	0.5776	0.9075

6 Discussion

For shared task on Understanding Toxic Behavioral Intent in Gaming Chat Logs@EEUCA with ACL 2026, we referred to the relevant tasks of CASE 2025 (Hurriyetoglu et al., 2025), CASE 2024 (Thapa et al., 2024) and CASE 2023 (Thapa et al., 2023) shared tasks on multimodal hate speech detection and derived our own method. Although the effect of the experiment needs to be strengthened. However, these contents and ideas have given us a lot of inspiration. Toxic behavioral intent content analysis is a longstanding tradition of the EEUCA workshop series. We believe that with our further research and more detailed optimization on training of the model, we will achieve even greater success in future competitions.

7 Conclusion

We employed multiple methods in detection of toxic behavioral intent in gaming chat logs, which respectively involved the transformer pre-trained models and Qwen2 LLM. Our final leaderboard is shown in the Table 5. The best result of this task was achieved by fine-tuning the Qwen2_7B large language model (LLM) and conducting inference on the test set.

Table 5: The final leaderboard of shared task on Understanding Toxic Behavioral Intent in Gaming Chat Logs@EEUCA with ACL 2026.

#	Username	Recall (Macro)	Precision (Macro)	F1 (Macro)	Accuracy
1	syuhhh-637901	0.7986	0.64	0.7041	0.8982
2	ramihai-572801	0.6846	0.6636	0.6725	0.8992
3	anmolguragain-637916	0.6601	0.6334	0.6441	0.9062
4	srikarkashyap-635409	0.6814	0.5864	0.6234	0.88
5	akshyatshah-636282	0.6497	0.6047	0.6186	0.8902
6	yinloonkhor-636292	0.5946	0.6098	0.5932	0.8925
7	shrinep-637207	0.659	0.554	0.5883	0.9031
8	**wangkongqiang-504685**	**0.5343**	**0.6847**	**0.5776**	**0.9075**
9	dkhonker-536426	0.5815	0.6214	0.5749	0.8865
10	_alexcristea-610819	0.5754	0.5652	0.5632	0.8733
11	akking-609884	0.6002	0.5239	0.5563	0.8876
12	rukesh-shrestha-503743	0.5557	0.5599	0.5539	0.8932
13	nepalshr-637149	0.6476	0.5201	0.5512	0.893
14	merrli-510969	0.6137	0.4798	0.5302	0.8603
15	xiaotian-518453	0.5291	0.5402	0.5301	0.8969
16	runick_allure-508659	0.5328	0.5441	0.5281	0.8772
17	rohanmainali-491803	0.5221	0.5192	0.5192	0.8893
18	linus-636500	0.5134	0.5191	0.5104	0.8716
19	xiaoyu666-603164	0.4884	0.5156	0.4984	0.8951
20	havnis-610798	0.5083	0.4766	0.4895	0.8794
21	giris-585517	0.4895	0.5081	0.4878	0.8964
22	shashi_sah-637803	0.4774	0.5001	0.4869	0.8999
23	wjyyyy-609715	0.4732	0.4962	0.4774	0.8953
24	justdoi-613394	0.5071	0.4487	0.4737	0.8973
25	barkion-610469	0.5002	0.4538	0.4726	0.8781
26	mestecha-623302	0.495	0.4763	0.4686	0.8927
27	binayakkarki-589485	0.4688	0.4647	0.4645	0.8921
28	syhhh-610772	0.5659	0.4198	0.4641	0.7792
29	exterio-610602	0.5084	0.4205	0.4491	0.8443
30	zmin123-554678	0.4568	0.4646	0.4487	0.8506
31	aryankafle-524077	0.4373	0.449	0.4421	0.8962
32	liutianyong-605718	0.4219	0.4701	0.4413	0.9036
33	quasar-501127	0.5357	0.3943	0.4169	0.6471
34	alexandru412-511289	0.6432	0.3315	0.3783	0.7068
35	wenbin-520996	0.1653	0.1629	0.1558	0.7784

8 Limitations of the Work

we are interested in learning about LLMs in computational social science (Thapa et al., 2025), our paper mainly focuses on making discussions on player's utterance for this toxic behavioral intent classification task. This is because we are quite interested in and good at identifying hate and offense categories in the text (Parihar et al., 2021). Due to our lack of utilization of context features, we are unable to make good use of the utterance content in the train dataset of this sharing task. we have chosen the 7B version of Qwen2 due to the limited computing resources. If we could use a larger language model with more parameters, we would achieve better prediction results. These are all our future tasks.

Acknowledgments

We are very grateful to the organizers of the Shared Task on Understanding Toxic Behavioral Intent in Gaming Chat Logs@EEUCA with ACL 2026 (Hürriyetoğlu et al., 2026) and the School of Information Science and Engineering of Yunnan University for providing the experimental environment and equipment.

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*, page 877–888.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Ali Hurriyetoglu, Surendrabikram Thapa, Hristo Tanev, and Surabhi Adhikari. 2025. Findings and insights from the 8th workshop on challenges and applications of automated extraction of socio-political events from text. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts*, pages 1–5, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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)*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. In *AAAI Conference on Artificial Intelligence*.

Marcus Märtens, Siqi Shen, Alexandru Iosup, and Fernando Kuipers. 2015. Toxicity detection in multiplayer online games. In *2015 International Workshop on Network and Systems Support for Games (NetGames)*, page 1–6. IEEE.

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.

Mustafa Oz, Akan Yanik, and Mikail Batu. 2023. Under the shadow of culture and politics: Understanding lgbtq social media activists' perceptions, concerns, and strategies. *Social Media + Society*, 9(3).

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.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech.

Semiu Salawu, Jo Lumsden, and Yulan He. 2021. A large-scale English multi-label Twitter dataset for cyberbullying and online abuse detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 146–156, Online. Association for Computational Linguistics.

Wessel Stoop, Florian Kunneman, Antal van den Bosch, and Ben Miller. 2019. Detecting harassment in real-time as conversations develop. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 19–24.

Surendrabikram Thapa, Farhan Ahmad Jafri, Ali Hürriyetoğlu, Francielle Vargas, Roy Ka Wei Lee, and Usman Naseem. 2023. Multimodal hate speech event detection-shared task 4. In *CASE 2023-Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, associated with 14th International Conference on Recent Advances in Natural Language Processing, RANLP 2023*, pages 151–159. Association for Computational Linguistics.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Hariram Veeramani, Raghav Jain, Sandesh Jain, Francielle Vargas, Ali Hürriyetoğlu, and Usman Naseem. 2024. Extended multimodal hate speech event detection during russia-ukraine crisis-shared task at case 2024. In *7th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, CASE 2024*, pages 221–228. Association for Computational Linguistics.

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)*.

Thilini Wijesiriwardene, Hale Inan, Ugur Kursuncu, Manas Gaur, Valerie L. Shalin, Krishnaprasad Thirunarayan, Amit Sheth, and I. Budak Arpinar. 2020. Alone: A dataset for toxic behavior among adolescents on twitter. In *Social Informatics*, pages 427–439, Cham. Springer International Publishing.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar.

2019. Predicting the type and target of offensive posts in social media.

wangkongqiang@EEUCA 2026: Multimodal Identification of Vaccine Critical Content on Social Media

Kongqiang Wang
School of Information Science and Engineering, Yunnan University, Yunnan Baiyao Street, 650500, Kunming, Yunnan, China.
wangkongqiang60@gmail.com

Peng Zhang
School of Information Science and Engineering, Yunnan University, Yunnan Baiyao Street, 650500, Kunming, Yunnan, China.
zpp1219@gmail.com

Qingli Tan
College of Ecology and Environment, Yunnan University, Yunnan Baiyao Street, 650500, Kunming, Yunnan, China.
tanqingli@stu.ynu.edu.cn

Abstract

Our team was interested in content classification and labeling from multimodal meme detection of vaccine critical content on social media. We joined the shared task on Multimodal Identification of Vaccine Critical Content on Social Media@EEUCA with ACL 2026. In this task, our goal is to assign a content classification label to vaccine-related discourse (e.g., Vaccine critical, Neutral, Pro-vaccine). The objective is to develop systems that can classify the intent of a vaccine-related meme. The dataset for this task will have three labels: Vaccine critical (0), Neutral (1), and Pro-vaccine (2). The performance will be ranked by F1-score (Macro). This shared task is based on the *VaxMeme* dataset, a collection of over 10,000 manually annotated vaccination-related memes, designed to support multimodal vaccine-critical meme detection. Our group used a supervised learning method on finetuning pre-trained models and Large Language Model (LLM), including Qwen2 LLMs and Llama series LLMs based on Llama-Factory. The best result on the test set for shared task were Macro F1 score of 0.8153, Accuracy 0.8185, Precision (Macro) 0.8151, and Recall (Macro) 0.8159 from finetuning qwen2_1.5B LLM method, ranking 12th among all teams. The complete code of this entire project can be found at our GitHub address[1].

1 Introduction

First of all, let's introduce the overview of shared task on Multimodal Identification of Vaccine Critical Content on Social Media@EEUCA with ACL 2026 (Thapa et al., 2026b). Memes have become a powerful and fast-spreading medium for sharing information online, especially around high-impact public health issues such as COVID-19 vaccination. While memes can be used to promote awareness and positive behavior, they are also frequently used to spread misinformation (Thapa et al., 2026a), skepticism (Thapa et al., 2024b), and vaccine-critical narratives, often through sarcasm and implicit context that make automated analysis challenging. In this context, the distinction between vaccine critical and pro-vaccine becomes blurred, as vaccination-related images straddle the line between satire and offense, challenging researchers and platforms alike to navigate the complexities of memes content moderation. As one label generally fails to encompass multiple aspects of linguistics, this shared task classifies memes on three aspects: Vaccine critical (0), Neutral (1), and Pro-vaccine (2).

This shared task is based on the *VaxMeme* dataset (Naseem et al., 2023), a collection of over 10,000 manually annotated vaccination-related memes, designed to support multimodal vaccine-critical meme detection. The task invites participants to develop models that jointly leverage both visual and textual representations to capture the global and local contextual cues embedded in memes. By focusing on fine-grained multimodal understanding, this challenge aims to advance more reliable systems for monitoring vaccine-related discourse, supporting myth debunking efforts, and informing the design of effective public health communication strategies on social media platforms.

2 Background

2.1 Content Detection for Vaccine Critical Posts

Majority of research (Wang et al., 2020) on identifying vaccine critical posts on social media has mainly focused on textual content. (Zhang et al., 2020) presented three models for analysing public sentiment on HPV vaccines on Twitter using

[1] https://github.com/WangKongQiang/EEUCA2026_Multimodal_Identification_of_Vaccine_Critical_Content_on_Social_Media

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 112–121

transfer learning. They fine-tuned bidirectional encoder representations from Transformers (BERT) (Devlin et al., 2019), and their results demonstrated the effectiveness of the proposed framework, which also aided in the discovery of vaccine uptake strategies. Recently, (Naseem et al., 2021) categorised vaccine-related tweeter posts using word representation from the domain-specific context with common knowledge and sentiment data. Their proposed method outperformed several traditional and recent transformer-based pre-trained language models. Previously published architectures, however, only focus on local semantic word representations using a sliding window for textual content. However, long-range and non-consecutive semantic links among feature representation words are required to capture global characteristics. We address this limitation by using a graph-based method to capture both local and global features of textual content.

Previously research has examined the use of multimodal content for detecting hateful memes (Lee et al., 2021), misleading information (Volkova et al., 2019), antisemitism (Chandra et al., 2021), and fake news detection (Wang et al., 2018). Experiments conducted using unimodal and multimodal in previous studies showed that understanding both modalities is essential for detection. Limited research has explored multimodal data to identify vaccine critical memes on social media. Recently, (Wang et al., 2020) created a multimodal dataset from Instagram posts and presented a multimodal framework with semantic and task-level attention to identifying vaccine critical information on social media. In contrast, our work jointly learns global and local representations of the textual and visual content of memes, which provide complementary information to improve the identification of vaccine critical memes on Twitter. We suggest that releasing a robustly annotated dataset to the community will support further advances and benchmarking of methods in this space.

2.2 Vaccine and Multimodal Datasets

Social media is a valuable source of information and has been widely used for various tasks like health mention classification (Naseem et al., 2022c), identifying suicide (Naseem et al., 2022b) and depression (Naseem et al., 2022a) and others. Systematic reviews show the wide range of applications for classifying user-generated content for vaccine hesitancy on social media, such as infectious diseases and outbreaks such as human papillomavirus, measles Influenza, mining misinformation mining.

Only two multimodal datasets are used in the previous studies to identify vaccine critical information on social media. The first of them was presented by (Wang et al., 2020), where authors used Instagram posts with text and visual content collected from January 2016 to October 2019 to identify vaccine critical information on Instagram posts. MMCoVaR (Chen et al., 2021), a multimodal COVID-19 vaccine focused data repository is the second dataset. MMCoVaR comprises 2,593 articles and 24,184 tweets from February 2020 to March 2021 and is limited to only COVID vaccine related posts. Both mentioned datasets are not publicly available, whereas we make our dataset publicly available for further research.

3 Dataset

In this section, we describe various aspects of task dataset including data collection, meme annotation, and dataset statistics. Task dataset comprises 10,244 memes that encompass different intent content relevant to the vaccination.

3.1 Utterance Annotation

Each meme was labeled to one of 3 labels: Vaccine critical (0), Neutral (1), and Pro-vaccine (3). Neutral if meme was not concept of vaccination. meme annotation for each label are mentioned below.

Vaccine critical: A meme (text or image or both) criticises vaccines, contains vaccine misinformation about vaccine side effects, vaccine conspiracy theories, and cases or statistical conclusions against vaccines.

Neutral: A meme (text or image or both) reports the events or others' opinions objectively related to vaccines, such as talking about rights of people related to vaccines, or news or statistical charts about vaccines showing no content in favor or against vaccines.

Pro-vaccine: A meme (text or image or both) contains a content in favor of vaccines, advising people to get vaccinated, a content about any event or place that is open only for vaccinated people or promoting and selling products with slogans in favor of vaccines.

Table 1: Dataset statistics for *VaxMeme*. The data consists of 10,244 samples for the multimodal identification task of vaccine critical memes on Twitter.

Data	Number of Pro-Vaccine	Number of Vaccine critical	Number of Neutral	Total
Full	3983	3441	2820	10244
Timeline	**Number of Pro-Vaccine**	**Number of Vaccine critical**	**Number of Neutral**	**Total**
T1	452	1679	1027	3158
T2	1040	747	1062	2849
T3	2491	1015	731	4237

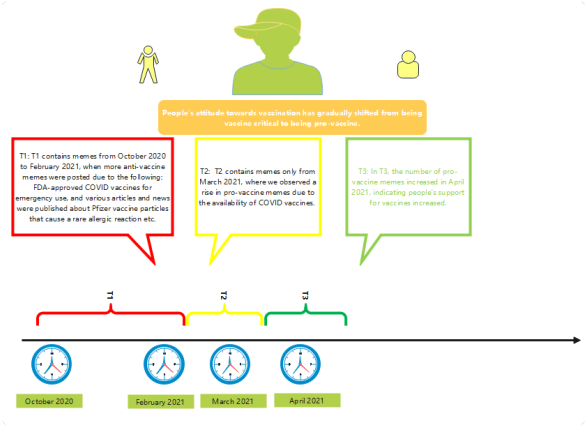


Figure 1: Timelines of the general vaccine critical memes gradually shifted in intention label.

3.2 Dataset Statistics

Table 1 (*Upper*) provides the class distribution of intent across the 10,244 English memes, and Table 1 (*Down*) provides the different timelines dividing distribution across all memes. Most memes are Pro-Vaccine or Vaccine critical in nature and a approximately data balance is present. However, this is in line with real world data distributions, where different people have different views on getting vaccinated. Figure 1 illustrates the timelines for vaccine critical memes.

4 System Overview

4.1 Fine-tuning Pre-trained Models

Introduction. In recent years, with the rapid development of deep learning technology, large-scale pre-trained models have achieved remarkable results in fields such as natural language processing, computer vision, and multimodal learning. Compared with traditional models trained from scratch, pre-trained models can learn rich semantic representations and general knowledge by pre-training on large-scale general corpora, thereby significantly improving the performance and training efficiency of downstream tasks. However, the knowledge learned by pre-trained models on general corpora often has strong generalization, while specific tasks usually have obvious domain characteristics. Therefore, directly applying pre-trained models to downstream tasks often fails to achieve the best results. To solve this problem, researchers usually adopt the fine-tuning strategy, that is, further optimize the model parameters using the data of specific tasks on the basis of the pre-trained model, so that it can better adapt to the target task. In this study, to enhance the model's performance in the task of vaccine critical content analysis on social media, a pre-trained language model was adopted as the base model and fine-tuned in combination with specific task data (*VaxMeme* dataset), enabling the model to effectively learn the semantic relationship between vaccination-related meme expressions and their underlying intents.

The classifier in the pre-trained model uses a transformer based classifier. The specific pre-trained models of the classifier are shown in the Table 2.

Table 2: pre-trained model classifier structure for vaccine critical content classification.

Model	Batch Size	Num Epochs	Learning Rate
alber/albert-base-v2	4	5	2e-5
google-bert/bert-base-uncased	4	5	2e-5
nghuyong/ernie-2.0-large-en	4	5	1e-5
nghuyong/ernie-1.0-base-zh	4	5	1e-5
FacebookAI/roberta-large-mnli	4	5	2e-5
cambridgeltl/trans-encoder-bi-simcse-roberta-large	4	5	1e-5

The Principle of Fine-tuning Pre-trained Models. Pre-trained models typically use large-scale corpora for self-supervised learning, such as tasks like language model prediction, masked language modeling, or autoregressive modeling, thereby learning common language representations. After pre-training, the model parameters already contain a large amount of language knowledge and semantic information. Therefore, in downstream tasks, only a small amount of labeled data is needed to achieve good performance.

The basic idea of fine-tuning is to introduce supervisory signals from downstream tasks on the basis of the parameters of the pre-trained model and further optimize the model parameters through the gradient descent algorithm. Let the parameters of the pre-trained model be θ, Given the downstream task training dataset $\mathcal{D} = (x_i, y_i)_{i=1}^{N}$, Where x_i represents the input text and y_i represents the corresponding label, then the model training objective can be expressed as:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell \left(f(x_i; \theta), y_i \right) \quad (1)$$

Here, $f(\cdot)$ represents the model prediction func-

tion, and $\ell(\cdot)$ is the loss function (such as cross-entropy loss). By minimizing this loss function, the model can gradually adapt to the data distribution of a specific task, thereby enhancing the prediction performance.

In practical applications, fine-tuning typically includes the following two methods:

- Full Fine-tuning: Update all the parameters of the pre-trained model to enable it to fully adapt to the target task.
- Parameter-efficient Fine-tuning: Only update some parameters or introduce additional lightweight modules to reduce training costs, such as Adapter, LoRA and other methods.

In this study, based on the characteristics of the task and the availability of computing resources, the pre-trained model was trained using a parameter-efficient fine-tuning strategy.

Input Data Construction. First of all, the original data needs to be converted into an input format that the model can handle. For text tasks, the following steps are usually required:

- Text Cleaning and Preprocessing: Remove irrelevant symbols or abnormal characters;
- Word Segmentation and Encoding: Use the tokenizer corresponding to the pre-trained model to convert the text into a token sequence;
- Input Sequence Construction: Ultimately, the input text will be represented as a sequence of token ids and input into the pre-trained model for feature encoding.

Task Structure Design. During the fine-tuning process, it is necessary to design the corresponding prediction structure based on the specific task. For instance, in the task of vaccine critical memes intent analysis on social media, the model needs to simultaneously identify vaccination-related discourse intent categories as well as the corresponding memes content information. Therefore, a model is usually composed of the following parts:

- Pre-trained Encoding Layer: Used for extracting semantic representations of text;
- Task-specific Layer: For example, the classification layer or the sequence labeling layer;

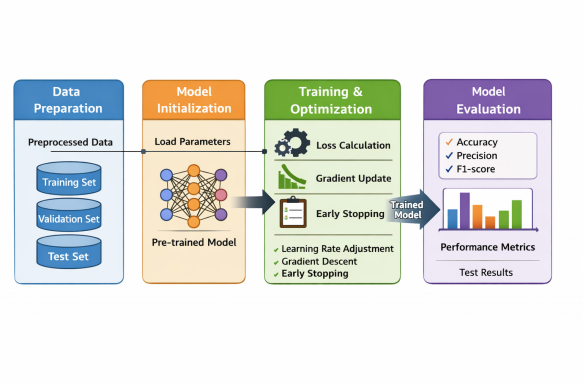


Figure 2: The framework diagram of the fine-tuning pre-trained classification model.

- Output Layer: Generate the final prediction result.

By adding task-related structures at the top of the pre-trained model, the model's adaptability to specific tasks can be effectively enhanced.

The overall architecture diagram of the fine-tuing pre-trained model is shown in the Figure 2.

Design of Loss Function. During the training process, it is necessary to select an appropriate loss function based on the type of task. For classification tasks, the cross-entropy loss function is usually adopted.

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(p_i) \tag{2}$$

Here, C represents the number of categories, y_i is the true label, and p_i is the model prediction probability. By minimizing the loss function, the consistency between the model's prediction results and the true labels can be gradually improved.

4.2 Hard Voting Mechanism

The hard voting mechanism is a common model fusion strategy in ensemble learning, mainly used for classification tasks. Its basic idea is: multiple base learners make predictions on the same sample respectively, and then determine the final prediction category through majority voting.

The Principle of Hard Voting Mechanism. Assume that the ensemble model consists of M base classifiers: $h_1(x), h_2(x), \ldots, h_M(x)$, where $h_i(x)$ denotes the prediction of the i-th classifier for input sample x. The final prediction of the hard voting ensemble is determined by majority voting:

$$\hat{y} = \arg\max_{c \in C} \sum_{i=1}^{M} I(h_i(x) = c) \tag{3}$$

where C represents the set of all possible classes and $I(\cdot)$ is an indicator function defined as:

$$I(h_i(x) = c) = \begin{cases} 1, & \text{if } h_i(x) = c \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

In a weighted hard voting scheme, each classifier is assigned a weight w_i, and the final prediction can be written as:

$$\hat{y} = \arg\max_{c \in C} \sum_{i=1}^{M} w_i \cdot I(h_i(x) = c) \quad (5)$$

In our experiment, the weights of each classifier were the same.

4.3 Fine-tuning of the Qwen2 and Llama series Large Language Model (LLM)

Qwen2 is an open-source large language model (LLM) developed by the Tongyi Qianwen team and created by Alibaba Cloud's Tongyi Lab. Using Qwen2 as the base large language model (LLM) and achieving high-accuracy text classification through instruction fine-tuning is an introductory task for learning the fine-tuning of large language models (LLMs). The Llama-2 7B (Touvron et al., 2023b) model approximately 7 billion parameters in an open-source large language model (LLM) released by Meta in 2023, belongs to a typical Transformer decoder architecture. Parameter scale is approximately 7B; Context length is approximately 4K tokens; Architecture is Standard Transformer + Multi-Head Attention (MHA). Features: Mature structure and stable reasoning; Suitable for deployment in resource-constrained environments; It is mostly used for basic NLP tasks and lightweight applications. The Llama-3 8B (Grattafiori et al., 2024) model of is a new generation version released in 2024, featuring significant upgrades in both architecture and training data. Parameter scale is approximately 8B; Context length is approximately 8K tokens (longer context); Architecture improvement based on enhance efficiency by using Grouped Query Attention (GQA); Tokenizer based on word list from 32K to 128K (stronger expressive power); The training data scale has significantly expanded (about 7 times).

Instruction fine-tuning is a process of further training an LLMs on a dataset composed of (instruction, input, output) combine pairs. Among them, the instructions represent the human instructions of the model, the input represent the raw data content from specific dataset, and the output represents the expected output that follows the instructions. This process helps bridge the gap between the next word prediction target of LLMs and the goal of users to have LLMs follow human instructions.

In this vaccine-related behavioral intent classification task on social media, we will use the Qwen2-1.5B-Instruct [2] and Qwen2-7B-Instruct [3] model to perform instruction fine-tuning tasks on the dataset, while using SwanLab [4] for monitoring and visualization. Llama3-8B, while maintaining a lightweight scale, has achieved a significant performance improvement over Llama2-7B through a larger vocabulary, longer context window, and GQA attention mechanism, making it a strong competitor among current open-source small-parameter models. The fine-tuning of the Llama series models (Touvron et al., 2023a) is mainly carried out using the Llama-Factory tool [5]. The following presents three demonstration formatted data samples for fine-tuning LLM data in the train dataset. Our training task is to ensure that the fine-tuned large language model (LLM) can predict the correct output based on the prompt words composed of post_text, image_text and selectable types.

The complete process of fine-tuning the Qwen2 and Llama series large language model (LLM) using the train dataset and conducting model inference on validation or test dataset with the fine-tuned large language model (LLM), as shown in Figure 3.

5 Results and Analysis

For results obtained by our fine-tuning pre-trained models and fine-tuning Qwen2 or Llama series LLM methods on the train dataset, and implement model inference on validation dataset and test dataset are shown in Table 3 and Table 4 respectively. roberta-RNN indicates the addition of a layer of recurrent neural network (RNN) after *FacebookAI/roberta-large-mnli* model, which is the LSTM layer. roberta-cnn indicates adding a convolutional neural network (CNN) layer after *FacebookAI/roberta-large-mnli*, which is the Conv2d layer. bagging refers to the model ensemble decision-making method that employs hard vot-

[2]https://huggingface.co/Qwen/Qwen2-1.5B-Instruct
[3]https://huggingface.co/Qwen/Qwen2-7B-Instruct
[4]https://swanlab.cn
[5]GitHub: https://github.com/hiyouga/LlamaFactory

Figure 3: The framework diagram of the fine-tuning qwen2 and Llama series classification LLM and model inference.

ing mechanism in the four models: roberta, roberta-RNN, simcse-roberta-lstm, roberta-lstm-gru.

Table 3: The results obtained by our fine-tuning pre-trained models and fine-tuning Qwen2 or Llama series LLM methods for vaccine-related discourse intent classification task on the validation dataset.

Pre-trained Model	**Recall (Macro)**	**Precision (Macro)**	**F1 (Macro)**	**Accuracy**
albert/albert-base-v2	-	-	-	-
google-bert/bert-base-uncased	-	-	-	-
nghuyong/ernie-2.0-large-en	-	-	-	-
nghuyong/ernie-1.0-base-zh	-	-	-	-
FacebookAI/roberta-large-mnli	-	-	-	-
cambridgeltl/trans-encoder-bi-simcse-roberta-large	-	-	-	-
roberta-RNN	-	-	-	-
roberta-cnn	-	-	-	-
roberta-lstm-gru	-	-	-	-
simcse-roberta-lstm	-	-	-	-
bagging	-	-	-	-
Large Language Model	**Recall (Macro)**	**Precision (Macro)**	**F1 (Macro)**	**Accuracy**
qwen2_1.5B	0.7945	0.7948	0.7943	0.7969
qwen2_7B	-	-	-	-
Llama2_7B	-	-	-	-
Llama3_8B	-	-	-	-

Table 4: The results obtained by our fine-tuning pre-trained models and fine-tuning Qwen2 or Llama series LLM methods for vaccine-related discourse intent classification task on the test set.

Model	**Recall (Macro)**	**Precision (Macro)**	**F1 (Macro)**	**Accuracy**
albert/albert-base-v2	0.7618	0.7916	0.7611	0.76
google-bert/bert-base-uncased	0.763	0.7734	0.7649	0.7727
nghuyong/ernie-2.0-large-en	0.7918	0.8016	0.7915	0.7941
nghuyong/ernie-1.0-base-zh	0.756	0.7652	0.7552	0.76
FacebookAI/roberta-large-mnli	0.8029	0.808	0.8025	0.8049
cambridgeltl/trans-encoder-bi-simcse-roberta-large	0.7933	0.7986	0.7935	0.7971
roberta-RNN	0.7985	0.8019	0.7971	0.799
roberta-cnn	0.7982	0.8	0.7988	0.8049
roberta-lstm-gru	0.7945	0.8076	0.7928	0.7922
simcse-roberta-lstm	0.804	0.8047	0.8036	0.8078
bagging	0.804	0.809	0.8025	0.8039
Large Language Model	**Recall (Macro)**	**Precision (Macro)**	**F1 (Macro)**	**Accuracy**
qwen2_1.5B	0.8159	0.8151	0.8153	0.8185
qwen2_7B	0.8135	0.815	0.8134	0.8166
Llama2_7B	0.8006	0.8026	0.7998	0.8029
Llama3_8B	0.8051	0.8044	0.8046	0.8098

6 Discussion

It is highly unusual that Qwen2-1.5B outperformed Llama-3-8B. A plausible explanation for this anomaly is that the 8B model may have been more sensitive to hyperparameter settings, such as learning rate or regularization, leading to suboptimal fine-tuning or mild overfitting on the training data. In contrast, the smaller 1.5B model could have benefited from better generalization under the same setup. Additionally, differences in pretraining data quality and alignment strategies between Qwen and Llama models may also have contributed to the performance gap. For shared task on Multimodal Identification of Vaccine Critical Content on Social Media@EEUCA with ACL 2026, we referred to the relevant tasks of CASE 2025 (Hurriyetoglu et al., 2025), CASE 2024 (Thapa et al., 2024a) and CASE 2023 (Thapa et al., 2023) shared tasks on multimodal hate speech detection and derived our own method. Although the effect of the experiment needs to be strengthened. However, these contents and ideas have given us a lot of inspiration. Vaccine-related discourse intent content analysis is a longstanding tradition of the EEUCA workshop series. We believe that with our further research and more detailed optimization on training of the model, we will achieve even greater success in future competitions.

7 Conclusion

We employed multiple methods in detection of vaccine-related discourse behavioral intent on social media, which respectively involved the transformer pre-trained models and Qwen2 or Llama series LLM. Our final leaderboard is shown in the Table 5. The best result of this task was achieved by fine-tuning the Qwen2_1.5B large language model (LLM) and conducting inference on the test set.

Table 5: The final leaderboard of shared task on Multimodal Identification of Vaccine Critical Content on Social Media@EEUCA with ACL 2026.

#	Username	Recall (Macro)	Precision (Macro)	F1 (Macro)	Accuracy
1	lili12-637947	0.8517	0.8494	0.8494	0.8517
2	wangxiuxian-637268	0.8409	0.8386	0.8389	0.842
3	rishta_19-611897	0.8359	0.8383	0.8357	0.839
4	_alexcristea-636983	0.8351	0.8338	0.834	0.838
5	sumaiya_110-594217	0.834	0.8345	0.8332	0.8361
6	anchy-637928	0.8309	0.8309	0.8308	0.8341
7	myname-637930	0.8309	0.8309	0.8308	0.8341
8	quasar-637336	0.8324	0.8331	0.8306	0.8322
9	wenbin-634065	0.8218	0.8205	0.8205	0.8244
10	naturia_beast-636958	0.8209	0.8212	0.8201	0.8244
11	vinaybabu-637935	0.819	0.8216	0.8184	0.8215
12	**wangkongqiang-495416**	**0.8159**	**0.8151**	**0.8153**	**0.8185**
13	ratpier-637076	0.8161	0.817	0.815	0.8176
14	yjwong1999-494691	0.8141	0.8189	0.8122	0.8137
15	linus-637363	0.8123	0.8106	0.8105	0.8137
16	havnis-636808	0.8083	0.808	0.8067	0.8117
17	alishba-wazir-604227	0.8071	0.8132	0.8067	0.8088
18	zmin123-553584	0.8013	0.8005	0.7997	0.8039
19	lin123-637530	0.8007	0.7992	0.7994	0.8039
20	barkion-636765	0.7986	0.808	0.7976	0.799
21	merli-636903	0.7982	0.8058 (19)	0.7972	0.799
22	exterio-636705	0.7846	0.7964	0.7861	0.7912
23	abs123-504332	0.7864	0.7868	0.7846	0.7912
24	thatgrass-519137	0.7802	0.7858	0.7754	0.7844
25	kannanrk-615633	0.7437	0.7435	0.7436	0.7502

8 Limitations of the Work

we are interested in learning about LLMs in computational social science (Thapa et al., 2025), our paper mainly focuses on making discussions on

vaccine-related discourse for this meme behavioral intent classification task. This is because we are quite interested in and good at identifying hate (Bhandari et al., 2023) and offense categories in the text (Parihar et al., 2021). Due to our lack of utilization of context features, we are unable to make good use of the image content in the train dataset of this sharing task. we have chosen the 7B version of Qwen2 due to the limited computing resources. If we could use a larger language model with more parameters, we would achieve better prediction results. These are all our future tasks.

Acknowledgments

We are very grateful to the organizers of the Shared Task on Multimodal Identification of Vaccine Critical Content on Social Media@EEUCA with ACL 2026 (Hürriyetoğlu et al., 2026) and the School of Information Science and Engineering of Yunnan University for providing the experimental environment and equipment.

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.

Mohit Chandra, Dheeraj Pailla, Himanshu Bhatia, Aadilmehdi Sanchawala, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. "subverting the jewtocracy": Online antisemitism detection using multimodal deep learning. In *13th ACM Web Science Conference*, page 148–157.

Mingxuan Chen, Xinqiao Chu, and KP Subbalakshmi. 2021. Mmcovar: multi-modal covid-19 vaccine focused data repository for fake news detection anda baseline architecture for classification. In *Proceedings of the 2021 IEEE/ACMInternational Conference on Advances in Social Networks Analysis and Mining*, pages 31–38.

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 Associationfor Computational Linguistics: Human Language Technologies*, volume 1 (Long andShort Papers), pages 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias

Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Ali Hurriyetoglu, Surendrabikram Thapa, Hristo Tanev, and Surabhi Adhikari. 2025. Findings and insights from the 8th workshop on challenges and applications of automated extraction of socio-political events from text. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts*, pages 1–5, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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)*.

Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. 2021. Disentangling hate in online memes. In *Proceedings of the 29th ACM International Conference on Multimedia*, page 5138–5147.

Usman Naseem, Adam G Dunn, Jinman Kim, and Matloob Khushi. 2022a. Early identification of depression severity levels on reddit using ordinal classification. In *Proceedings of the ACM Web Conference*, pages 2563–2572.

Usman Naseem, Matloob Khushi, Jinman Kim, and Adam G. Dunn. 2021. Classifying vaccine sentiment tweets by modelling domain-specific representation and commonsense knowledge into context-aware attentive gru.

Usman Naseem, Matloob Khushi, Jinman Kim, and Adam G Dunn. 2022b. Hybrid text representation for explainable suicide risk identification on social media. *IEEE transactions on computational social systems*.

Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G Dunn. 2022c. Identification of disease or symptom terms in reddit to improve health mention classification. In *Proceedings of the ACM Web Conference*, pages 2573–2581.

Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G Dunn. 2023. A multimodal framework for the identification of vaccine critical memes on twitter. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 706–714.

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.

Laxmi Thapa, Aryaman Jain, Lakshmojee Koduru, Surabhi Adhikari, Junaid Rashid, Jungeun Kim, Surendrabikram Thapa, and Usman Naseem. 2026a. Concept-grounded detection of vaccine misinformation in multimodal content using interpretable vision-language models. In *Companion Proceedings of the ACM on Web Conference 2026*.

Surendrabikram Thapa, Farhan Ahmad Jafri, Ali Hürriyetoğlu, Francielle Vargas, Roy Ka Wei Lee, and Usman Naseem. 2023. Multimodal hate speech event detection-shared task 4. In *CASE 2023-Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, associated with 14th International Conference on Recent Advances in Natural Language Processing, RANLP 2023*, pages 151–159. Association for Computational Linguistics.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Hariram Veeramani, Raghav Jain, Sandesh Jain, Francielle Vargas, Ali Hürriyetoğlu, and Usman Naseem. 2024a. Extended multimodal hate speech event detection during russia-ukraine crisis-shared task at case 2024. In *7th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, CASE 2024*, pages 221–228. Association for Computational Linguistics.

Surendrabikram Thapa, Kritesh Rauniyar, Hariram Veeramani, Aditya Shah, Imran Razzak, and Usman Naseem. 2024b. Did you tell a deadly lie? evaluating large language models for health misinformation identification. In *International Conference on Web Information Systems Engineering*, pages 391–405. Springer.

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. 2026b. Multimodal identification of vaccine content stance on social media. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Svitlana Volkova, Ellyn Ayton, Dustin L Arendt, Zhuanyi Huang, and Brian Hutchinson. 2019. Explaining multimodal deceptive news prediction

model. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 659–662.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conferenceon knowledge discovery data mining*, pages 849–857.

Zuhui Wang, Zhaozheng Yin, and Young Anna Argyris. 2020. Detecting medical misinformation on social media using multimodal deep learning. *IEEE Journalof Biomedical and Health Informatics*, 25(6):2193–2203.

Li Zhang, Haimeng Fan, Chengxia Peng, Guozheng Rao, and Qing Cong. 2020. Sentiment analysis methods for hpv vaccines tweets based on transfer learning. *Healthcare*, 8(307).

Quasar@EEUCA 2026: Multimodal Deep Learning for Vaccine Stance Detection in Memes

Adiba Fairooz Chowdhury, MD. Sagor Chowdhury
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2004014, u2004010}@student.cuet.ac.bd

Abstract

Vaccine stance detection in multimodal memes has emerged as an important yet challenging task, requiring models to interpret both textual and visual cues that jointly convey opinions. The difficulty lies in capturing subtle semantic interactions and handling class imbalance across stance categories. In this paper, we present our system developed for the VaxMeme 2026 Shared Task at EEUCA 2026. Our approach leverages a soft-voting ensemble of complementary models, combining DeBERTa-v3-large and RoBERTa-large for rich textual representation with CLIP (ViT-B/32) for joint vision-language understanding. We incorporate domain-specific preprocessing, techniques such as random token deletion, image enhancement, and balanced class oversampling to address dataset limitations. Through extensive ablation studies, we identify balanced class oversampling as the most effective component, significantly improving performance across models. Our final system achieves a macro F1-score of 0.8306, securing 8th place among 25 teams, demonstrating the effectiveness of ensemble-based multimodal learning for stance detection.

1 Introduction

Vaccine hesitancy has emerged as a critical public health challenge, with social media playing a significant role in the spread of both pro-vaccine advocacy and vaccine misinformation (Sallam, 2021). Among the many forms of health-related content on social media, memes have proven particularly influential: they combine image and text into a single communicative act, exploiting sarcasm, irony, and cultural reference to encode stances that neither modality alone reveals (Kiela et al., 2020). The multimodal nature of memes makes automatic stance detection substantially harder than text-only misinformation detection, and has direct relevance to public health (Thapa et al., 2024).

The VaxMeme 2026 Shared Task (Thapa et al., 2026b), organised as part of the EEUCA 2026 workshop (Hürriyetoğlu et al., 2026), which focuses on event extraction and understanding challenges has introduced a benchmark for this problem. The task requires systems to classify English-language vaccine memes from the VaxMeme dataset (Naseem et al., 2023; Thapa et al., 2026a; Bhandari et al., 2023) into three stance categories: *vaccine-critical*, *neutral*, and *pro-vaccine*. With 8,195 memes and moderate class imbalance, the dataset presents challenges for both model selection and training strategy.

To address this task, we have developed a multimodal ensemble system. We have systematically evaluated text-only models (TF-IDF, BERT, RoBERTa variants, DeBERTa variants), image-only models (ResNet-50, ViT, Swin, ConvNeXt, EfficientNet, CLIP Vision), and multimodal models (CLIP, BLIP, LLaVA), applying domain-specific text preprocessing, random token deletion augmentation, image enhancement, and class balancing via data augmentation. Our experiments show that balancing the dataset through augmentation is the single most impactful intervention, boosting even simple TF-IDF models by approximately 4.3 macro F1 points, while the specific choice of augmentation strategy has minimal effect. Furthermore, CLIP multimodal with image enhancement and balanced training data outperforms all text-only models. Our final soft-voting ensemble of DeBERTa-v3-large, RoBERTa-large, and CLIP multimodal achieves a macro F1 of 0.8306, placing us 8th out of 25 participating teams.

The main contributions of this work are:

- We have conducted a comprehensive comparison of text-only, image-only, and multimodal architectures across multiple preprocessing and augmentation configurations, providing a systematic ablation of what helps for vaccine

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 122–132

stance detection in memes.

- We have shown that addressing class imbalance is the most impactful single intervention, improving all model families substantially and implicating class imbalance as a primary bottleneck.
- We have shown that random token deletion generally matches or slightly outperforms synonym replacement as a text augmentation strategy for the short, domain-specific vocabulary of vaccine memes.
- We have developed a domain-adapted preprocessing pipeline for meme text that deliberately preserves stance-relevant informal signals such as emoji, hashtags, and repeated characters.

Further implementation details will be available via our code repository.[1]

2 Background

The VaxMeme 2026 Shared Task (Thapa et al., 2026b) presents a three-class stance classification problem: given a meme consisting of an image and its OCR-extracted text overlay, a system must assign one of three labels—vaccine-critical, neutral, or pro-vaccine. For example, a meme showing a syringe with the caption "they want to inject you with poison" would be labelled vaccine-critical, while one showing a vaccination queue with "protecting our community" would be pro-vaccine; a neutral meme might present statistics without implicit endorsement or criticism. Table 1 provides representative examples of multimodal inputs and their corresponding stance labels. The dataset (Naseem et al., 2023; Thapa et al., 2026a; Bhandari et al., 2023) contains 8,195 English-language social media memes with moderate class imbalance: pro-vaccine (39.0%), vaccine-critical (30.9%), and neutral (30.0%). For the official evaluation, 8,195 training samples, 1,024 validation samples (labels released at the test phase), and 1,025 unlabeled test samples are provided. In the development phase we use a stratified split without access to official validation labels: 5,736 train / 820 val / 1,639 test. The primary evaluation metric is macro F1, which equally weights performance across all three classes regardless of their frequency. We participated in the single track of this shared task.

[1] GitHub Repository

The task is organised as part of the EEUCA 2026 workshop (Hürriyetoğlu et al., 2026), which focuses on event extraction and understanding challenges. Meme classification has been studied extensively in the context of hate speech (Kiela et al., 2020), where multimodal models such as VisualBERT and UNITER demonstrated that jointly encoding image and text substantially outperforms unimodal approaches. The CrisisHateMM dataset (Bhandari et al., 2023) provides an annotation schema for multimodal hate content in social media images, which informs the VaxMeme annotation design. Vaccine misinformation detection on text has received considerable attention (Jennings et al., 2021; Hayawi et al., 2022; Thapa et al., 2024), and the VaxMeme dataset (Naseem et al., 2023; Thapa et al., 2026a) extends this line of work to multimodal meme-level stance classification. Thapa et al. (2025) survey the use of large language models in computational social science, providing broader context for NLP-based misinformation detection.

CLIP (Radford et al., 2021) and BLIP (Li et al., 2022) provide state-of-the-art joint vision-language pretraining suitable for multimodal meme understanding. DeBERTa (He et al., 2021) introduces disentangled attention that separately encodes content and positional representations, yielding strong results on social media text. Prior ensemble work on social media classification (Cai et al., 2019) confirms that combining diverse model families reliably outperforms any single model.

3 System Overview

Figure 1 illustrates the overall system pipeline. Each meme is processed through parallel text and image pipelines. These pipelines perform preprocessing and optional data augmentation before extracting features using their respective models. The extracted features are then fed into three different model families, and their output probability distributions are combined via soft voting to produce the final predictions. Figure 2 shows the detailed architecture of the final ensemble, highlighting how features from different models are integrated to improve classification performance.

3.1 Preprocessing

Meme text is often noisy and informal. We apply a lightweight normalization pipeline consisting of: (1) whitespace trimming and normaliza-

Input			Output
Image	**Post Text**	**Image Text**	**Label**
	Unvaccinated is Sexy AF. https://t.co/SJQ6DpBh17	Unvaccinated MATTER	0
	@pauraenisciun Unvaccinated https://t.co/KMDvQ...	Bullying in 2022	1
	IM VAXXED YALL https://t.co/xfixZZV9oI	D	2

Table 1: Example memes from the VaxMeme dataset. The first three columns represent the multimodal inputs (image, post text, and OCR-extracted image text), while the final column is the stance label. Labels: 0 = Vaccine-critical, 1 = Neutral, 2 = Pro-vaccine.

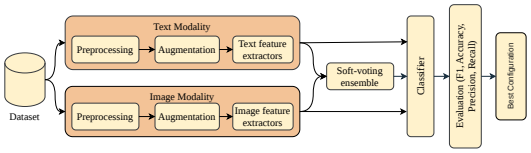


Figure 1: Overall system pipeline for processing memes through text and image models, with outputs combined via soft voting.

tion; (2) URL replacement with `[URL]`; (3) emoji demojization (e.g. `:syringe:`); (4) reduction of repeated characters; (5) hashtag normalization; (6) mention normalization; and (7) removal of non-alphanumeric characters except common punctuation. Although some steps (e.g., whitespace cleanup and emoji handling) have minimal effect on this dataset, they are retained for consistency and robustness. Importantly, normalization is conservative to preserve stance-indicative tokens such as hashtags and informal expressions. Table 2 shows representative examples of each step.

All images are uniformly resized to 224×224 pixels and enhanced using a three-step procedure: contrast adjustment ($\times 1.2$), brightness scaling ($\times 1.1$), and sharpness enhancement ($\times 1.1$). Figure 3 illustrates the enhancement effect on example memes.

3.2 Class Balancing and Augmentation

The original training set is moderately imbalanced: pro-vaccine 39.0%, vaccine-critical 30.9%, and neutral 30.0%. To address this, we perform class-balanced oversampling, increasing each class to a fixed target size (e.g., 3,200 or 4,000 samples per class depending on the experiment). Oversampling is implemented via sampling with replacement. For the duplicated samples only, we apply random token deletion as a text augmentation strategy. Specifically, each token is removed with probability $p = 0.1$, and augmentation is applied to a sample with probability 0.3. We intentionally avoid geometric augmentations (e.g., flipping, rotation, cropping), as preliminary experiments in our ablation study (Appendix A.4) showed a drop in validation F1 from 0.8140 to 0.8069 when these transformations were applied. This suggests that meme semantics depend heavily on layout and text placement, which such transformations may distort.

This approach preserves domain-specific vocabulary (e.g., vaccine names, slang) while introducing slight variability, making it more suitable than synonym replacement, which risks altering semantic

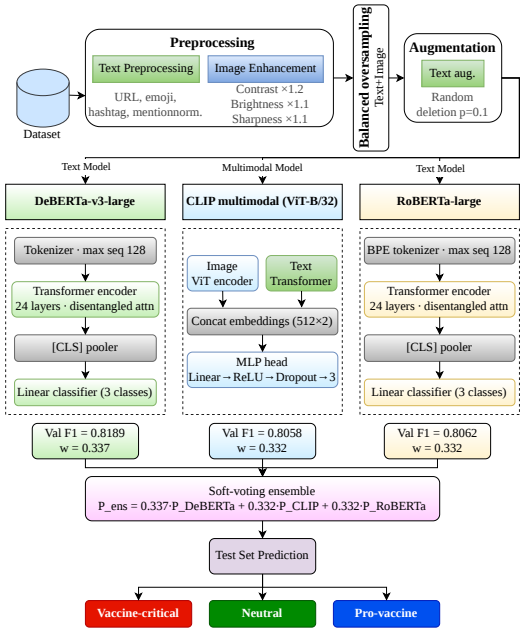


Figure 2: Final ensemble architecture: DeBERTa-v3-large, CLIP multimodal (ViT-B/32), and RoBERTa-large with soft voting. Weights are proportional to individual validation F1.

Before	Preprocessing	After
`Vaxxed!  Yass! https://t.co/xxx`	Normalise space	`Vaxxed! Yass! https://t.co/xxx`
`Vaxxed https://t.co/xxx`	Remove URL	`Vaxxed [URL]`
`vaxxed???????? [URL]`	Repeat char	`vaxxed?? [URL]`
`I got #vaxxed [URL]`	Hashtag	`I got hashtag_vaxxed [URL]`
`@UKvaxxed Vaxxed [URL]`	Mention	`mention_UKvaxxed Vaxxed [URL]`
`Vaxxed [URL]`	Clean	`Vaxxed URL`

Table 2: Preprocessing pipeline with before-and-after examples for each step.

Figure 3: Image enhancement examples: original (left) and enhanced (right).

meaning. Notably, augmentation is applied only to oversampled instances, while original samples remain unchanged.

Class balancing is the most impactful intervention in our pipeline. As shown in Table 10, even simple models such as TF-IDF classifiers benefit significantly (+4–5 macro F1 points), indicating that performance gains primarily stem from improved class distribution rather than augmentation alone. Figure 4 shows the class distribution before and after balancing.

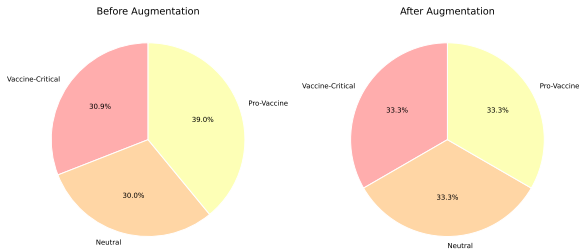


Figure 4: Class distribution before (left) and after (right) balancing to 4,000 samples per class.

3.3 Model Selection and Justification

We select three models for the final ensemble, each chosen based on ablation evidence:

DeBERTa-v3-large achieves the strongest text-only result in ablation (on the development split: 0.8107 without augmentation, 0.8110 with P+D). Its disentangled attention mechanism, which separately encodes content and positional information (He et al., 2021), is particularly suited to the short, stylised text of memes.

RoBERTa-large reaches 0.8130 with P+D augmentation on the development split, complementing DeBERTa through a different pretraining objective and corpus. Despite similar macro F1, the two models make different errors, making them effective ensemble partners.

CLIP multimodal (ViT-B/32) is the only model in our system that jointly encodes image and text. With image enhancement, balanced oversampling, text preprocessing and text deletion augmentation, it reaches macro F1=0.8502 on the development set—outperforming all text-only models—confirming that visual features carry stance signals not captured by text alone. BLIP was evaluated but plateaued at 0.7892 regardless of pipeline configuration, likely due to insufficient fine-tuning, and was excluded. We evaluated ViT-B/32 as it fits within the 16 GB memory budget of a single Kaggle P100; ViT-L/14 requires approximately 2× the GPU memory for the same batch size and was not feasible under our computational constraints.

3.4 Ensemble Strategy

We combine the three models via soft voting, averaging class probability vectors with weights proportional to validation macro F1:

$$P_{\text{ens}} = w_1 P_{\text{DeBERTa}} + w_2 P_{\text{CLIP}} + w_3 P_{\text{RoBERTa}} \quad (1)$$

where $w_1 = 0.337$, $w_2 = 0.332$, $w_3 = 0.332$. We explored majority voting and four-model configurations (adding TF-IDF and BLIP); none exceeded the three-model soft-voting result (Section 5).

3.5 End-to-End Inference Example

Figure 5 shows the inference pipeline for a representative sample (index 30): an image of a man holding a CDC COVID-19 Vaccination Record Card and the post text `Fully vaxxed w/Pfizer. https://t.co/OzvacyQTT9`. Text is preprocessed via MemeTextPreprocessor (URL replaced with

[URL], whitespace normalized, special characters removed), yielding Fully vaxxed wPfizer. URL". The 900×900 image is enhanced (contrast $\times 1.2$, brightness $\times 1.1$, sharpness $\times 1.1$) and resized to 224×224. DeBERTa-v3-large and RoBERTa-large encode the text, CLIP encodes image and text embeddings which are concatenated and passed through a 2-layer classification head. Weighted soft voting ($w = (0.337, 0.332, 0.332)$) combines outputs: DeBERTa predicts Neutral (0.638), CLIP and RoBERTa predict Pro-vaccine (0.968, 0.720), giving a final correct Pro-vaccine prediction (ensemble probability 0.611, computed from the full class probability vectors), showing how strong visual cues can override sparse textual input.

4 Experimental Setup

We use the official shared task splits described in Section 3. All models are trained on the full training set with balanced augmentation. Validation macro F1 guides model selection and weight calibration.

DeBERTa-v3-large and RoBERTa-large are fine-tuned with AdamW (lr=1e-5, weight decay=0.01), linear schedule with 10% warmup, batch size 8 (gradient accumulation steps 4 and 2 respectively), fp16, for 5 epochs with early stopping (patience=3 on val macro F1). CLIP uses AdamW (lr=2e-5), ReduceLROnPlateau scheduler, batch size 16, up to 10 epochs with early stopping (patience=3). Full hyperparameters are in Appendix B.

We use HuggingFace Transformers v4.40.0[2] for all transformer models, OpenAI CLIP[3] for the multimodal encoder, Salesforce BLIP[4] for the BLIP baseline, scikit-learn v1.4.2[5] for TF-IDF and logistic regression, and Pillow v10.3.0[6] for image enhancement.

All experiments use a single NVIDIA P100 (16 GB) via Kaggle.

Macro F1 is the primary metric, consistent with the shared task evaluation. We additionally report accuracy, per-class precision, recall, and F1, along with a confusion matrix for error analysis.

Formally, let $\mathbf{C} \in N^{K \times K}$ denote the confusion matrix, where C_{ij} is the number of samples with true class i predicted as class j. For class k, precision, recall, and F1 are defined as:

$$\text{Precision}_k = \frac{C_{kk}}{\sum_j C_{jk}}, \tag{2}$$

$$\text{Recall}_k = \frac{C_{kk}}{\sum_j C_{kj}}, \tag{3}$$

$$\text{F1}_k = \frac{2 \cdot \text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}. \tag{4}$$

Macro F1 is computed as:

$$\text{Macro-F1} = \frac{1}{K} \sum_{k=1}^{K} \text{F1}_k. \tag{5}$$

Overall accuracy is defined as:

$$\text{Accuracy} = \frac{\sum_k C_{kk}}{\sum_{i,j} C_{ij}}. \tag{6}$$

5 Results

5.1 Progressive System Development

Table 3 summarises our experiments as a progression from simple baselines to the final submission, grouped by model category. All results are on the official test set (1,025 samples) unless otherwise noted. Full per-model breakdowns are in Appendix D.

The results tell a clear story. Image-only models establish a ceiling of 0.7156, confirming text as the dominant modality. Plain text baselines without augmentation reach 0.8046 (TF-IDF+LogReg), which large pretrained transformers with augmentation push to 0.8252 (RoBERTa-large). Adding CLIP multimodal to the ensemble breaks the text-only ceiling, and increasing the augmentation target from 3,200 to 4,000 samples per class yields the final gain to 0.8306. Zero-shot LLaVA-1.5-13B (0.3447) confirms that general-purpose large vision-language models require task-specific fine-tuning for this domain.

5.2 Final Ensemble and Submission

Table 4 shows the individual validation scores of each component in the final ensemble. Weights are proportional to validation F1.

Adding TF-IDF or BLIP to the ensemble does not improve results: TF-IDF's lexical patterns are largely subsumed by the large transformers, and BLIP's contribution is already covered by CLIP.

[2] https://github.com/huggingface/transformers
[3] https://github.com/openai/CLIP
[4] https://github.com/salesforce/BLIP
[5] https://scikit-learn.org
[6] https://python-pillow.org

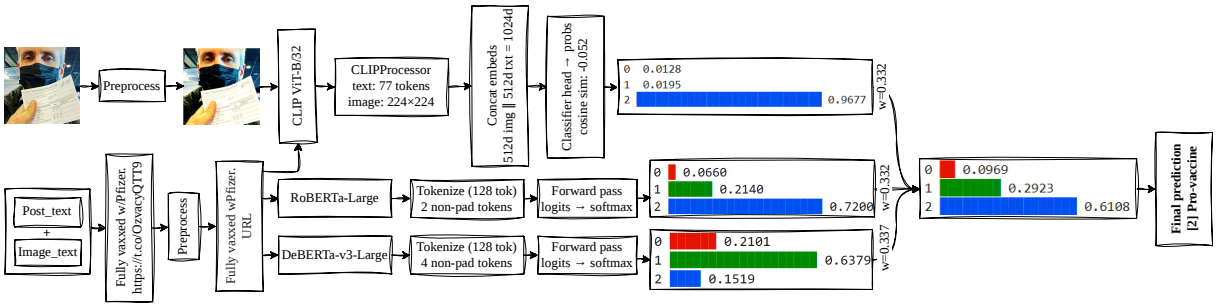


Figure 5: End-to-end inference for sample 30. The image and post text ("`Fully vaxxed w/Pfizer. [URL]`") are preprocessed separately and fed to DeBERTa, CLIP, and RoBERTa. Bars show class probabilities; weighted soft voting produces the final Pro-vaccine prediction, illustrating how visual signals can outweigh sparse textual cues.

Stage	Model / Configuration	F1
Image-only baselines		
	ResNet-50	0.6875
	ConvNeXt-base	0.6603
	CLIP Vision (image only)	0.7156
Text-only baselines (no aug)		
	TF-IDF + LogReg	0.8046
	BERT-base	0.7962
	RoBERTa-base	0.8028
Text models + preprocessing & aug		
	DeBERTa-v3-base	0.7967
	BERT-base	0.8088
	RoBERTa-base	0.8200
Large text models (+ preprocessing & aug)		
	XLM-RoBERTa-large	0.8026
	DeBERTa-v3-large	0.8146
	RoBERTa-large	**0.8252**
Multimodal		
	CLIP Multimodal (no aug)	0.7848
	CLIP Multimodal (aug=3200)	0.7957
	CLIP Multimodal (aug=4000)	0.8002
	LLaVA-1.5-13B (zero-shot)	0.3447
Ensemble configurations		
	RoBERTa + CLIP + TF-IDF	0.8252
	CLIP + RoBERTa + DeBERTa (3200)	0.8286
	RoBERTa + TF-IDF + CLIP	0.8299
	Stacking (+ BLIP)	0.8262
	CLIP + RoBERTa + DeBERTa (4000)	**0.8306**

Table 3: Progression of results from image-only baselines to the final submission. Bold = best in group.

5.3 Per-class Analysis and Error Analysis

Table 5 and Figure 6 show per-class results and the confusion matrix for the final ensemble on validation.

The neutral class is the hardest (F1=0.755), misclassified in both directions: 40 neutral memes are predicted vaccine-critical and 28 are predicted pro-vaccine. Neutral memes tend to present factual information without explicit stance markers, making them inherently ambiguous. Vaccine-critical memes have the highest recall (0.815), suggesting they carry strong distinctive lexical and visual signals. The test prediction distribution—vaccine-critical 30.6%, neutral 35.1%, pro-vaccine 34.2%—closely mirrors the training distribution, indicating no systematic class bias.

Model	Val F1	Acc.	Weight
DeBERTa-v3-large	0.8189	0.8213	0.337
CLIP Multimodal (ViT-B/32)	0.8058	0.8086	0.332
RoBERTa-large	0.8062	0.8076	0.332
Ensemble (submitted)	**0.8140**	**0.8154**	—

Table 4: Final ensemble components, validation scores, and soft-voting weights. Test set macro F1 = **0.8306**.

Class	Prec.	Recall	F1
Vaccine-critical	0.826	0.815	0.820
Neutral	0.721	0.792	0.755
Pro-vaccine	0.900	0.836	0.867
Macro avg	0.816	0.814	**0.814**

Table 5: Per-class validation results for the final ensemble.

6 Discussion

Balanced class oversampling emerges as the most impactful intervention in our study. While text-only models outperform CLIP individually (RoBERTa-large: 0.8062 val, DeBERTa-v3-large: 0.8189 val vs. CLIP: 0.8058 val), CLIP still improves ensemble performance. To understand this, we analysed pairwise prediction agreement between models on the test set: CLIP agrees with DeBERTa on only 86.3% of test samples and with RoBERTa on 85.8%, whereas the two text models agree with each other on 92.9%. Although DeBERTa and RoBERTa individually outperform CLIP on their respective disagreements, CLIP's lower correlation with both text models introduces sufficient diver-

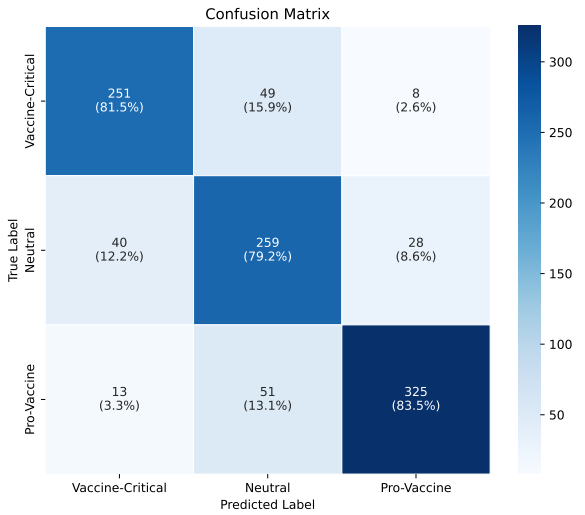


Figure 6: Confusion matrix for the final ensemble on the validation set (1,024 samples).

sity for soft voting to exploit—consistent with ensemble theory, where diversity among members is as important as individual accuracy. In terms of augmentation, random token deletion proves more effective than synonym replacement across all eight transformer models (Table 7), as synonym substitution risks altering domain-specific vocabulary such as vaccine names and slang, whereas deletion preserves the original distribution while introducing useful variation. Finally, large models such as DeBERTa-v3-large and RoBERTa-large exhibit early overfitting, peaking at epochs 2 and 3 respectively, with validation loss increasing thereafter, suggesting that the dataset size (8,195 samples) is limiting and that early stopping is crucial for stable performance. An experiment in which random crop, horizontal flip, colour jitter, and rotation were applied during training reduced ensemble validation F1 from 0.8140 to 0.8069 (Appendix A.4). We attribute this to the nature of meme images: meaning is encoded through composition and text placement, which geometric transforms disrupt. Image enhancement alone (contrast, brightness, sharpness) proved sufficient.

7 Conclusion

We presented a multimodal ensemble system for vaccine stance detection in memes, achieving 0.8306 macro F1 and ranking 8th out of 25 teams in the VaxMeme 2026 Shared Task (Thapa et al., 2026b). Our key finding is that balanced class oversampling is the most impactful intervention: it substantially boosts all model families and should be the first consideration on imbalanced meme datasets. A soft-voting ensemble of DeBERTa-v3-large, RoBERTa-large, and CLIP multimodal—trained with balanced oversampling and domain-adapted preprocessing—achieves competitive performance close to the top systems. Future work will explore cross-modal attention mechanisms and task-specific fine-tuning of larger vision-language models.

Limitations

Our system is trained on English-language COVID-19 vaccine memes; generalisation to other languages or vaccine contexts is unverified. Text deletion augmentation may introduce noise for very short meme texts. Computational constraints limited hyperparameter search to a single run per configuration.

Ethics Statement

All data consists of publicly available social media content. Stance detection tools carry a risk of misuse for content censorship; we advocate for their application in public health research and discourse analysis only.

Acknowledgments

We thank the VaxMeme 2026 shared task organisers for providing the dataset and evaluation infrastructure, and the EEUCA 2026 workshop chairs for hosting the competition. We also thank Kaggle for providing the computational resources used in this work.

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.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515.

Kheir Hayawi, Sakib Shahriar, Mohamed Adel Serhani, Ikbal Taleb, and Sujith Samuel Mathew. 2022. ANTi-Vax: A novel twitter dataset for COVID-19 vaccine misinformation detection. *Public Health*, 203:23–30.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

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

Will Jennings, Gerry Stoker, Hannah Bunting, Viktor Orri Valgarßson, Jennifer Gaskell, Daniel Devine, Lawrence McKay, and Melinda C. Mills. 2021. Lack of trust, conspiracy beliefs, and social media use predict COVID-19 vaccine hesitancy. *Vaccines*, 9(6):593.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900.

Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G Dunn. 2023. A multimodal framework for the identification of vaccine critical memes on Twitter. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 706–714.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Malik Sallam. 2021. COVID-19 vaccine hesitancy worldwide: a concise systematic review of vaccine acceptance rates. *Vaccines*, 9(2):160.

Laxmi Thapa, Aryaman Jain, Lakshmojee Koduru, Surabhi Adhikari, Junaid Rashid, Jungeun Kim, Surendrabikram Thapa, and Usman Naseem. 2026a. Concept-grounded detection of vaccine misinformation in multimodal content using interpretable vision-language models. In *Companion Proceedings of the ACM on Web Conference 2026*.

Surendrabikram Thapa, Kritesh Rauniyar, Hariram Veeramani, Aditya Shah, Imran Razzak, and Usman Naseem. 2024. Did you tell a deadly lie? Evaluating large language models for health misinformation identification. In *International Conference on Web Information Systems Engineering*, pages 391–405. Springer.

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. 2026b. Multimodal identification of vaccine content stance on social media. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

A Ablation Study

All ablation results use a stratified development split: 5,736 train / 820 val / 1,639 test. Macro F1 is reported on the held-out test partition throughout.

A.1 Simple Baselines

Model	Macro F1
ResNet-50 (image only)	0.6518
Late Fusion (BERT+ResNet)	0.7041
TF-IDF + SVM	0.7919
BERT-base	0.7949
TF-IDF + LogReg	0.7995
RoBERTa-base	0.7959
CLIP Multimodal	0.7871
Enhanced TF-IDF	0.8017
Super Ensemble	**0.8075**

Table 6: Initial baselines without any preprocessing or augmentation.

A.2 Text Transformer Ablation

Model	None	Prep	Syn	Del	P+D
Base-sized models					
BERT-base	0.7912	0.8014	0.7938	0.7955	0.7998
RoBERTa-base	0.7990	0.7894	0.7888	0.7894	0.7795
Twitter-RoBERTa	0.7985	0.7988	0.7990	0.8012	0.7997
XLM-RoBERTa-base	0.7945	0.7912	0.7956	0.7889	0.7844
DeBERTa-v3-base	0.8058	0.8062	0.7993	0.8050	0.7966
Large models					
RoBERTa-large	0.7970	0.8026	0.8031	0.8050	0.8130
XLM-RoBERTa-large	0.7997	0.8014	0.7924	0.7975	0.8049
DeBERTa-v3-large	**0.8107**	0.8054	0.8001	0.8021	**0.8110**

Table 7: Text transformer macro F1 across all preprocessing and augmentation configurations (development split). Prep = text preprocessing; Syn = synonym augmentation; Del = random deletion; P+D = Prep+Del. Bold = best per group.

A.3 Image Enhancement and Oversampling

Table 8 reports image-only macro F1 before and after applying image enhancement and balanced oversampling.

Model	Baseline	w/ Enh+Oversample
ConvNeXt-base	0.6213	0.6964
EfficientNet-B3	0.6383	0.6753
ResNet-50	0.6518	0.7118
Swin-base	0.6670	0.7699
ViT-base	0.6672	**0.7834**
CLIP Vision	0.6915	0.7725

Table 8: Image-only macro F1 before and after image enhancement and balanced oversampling (development split).

A.4 Impact of Image Geometric Augmentation

We conducted a small ablation to evaluate the effect of geometric augmentation (flipping, cropping, rotation) on validation performance. Table 9 summarizes the results for our CLIP multimodal ensemble.

Pipeline	Val F1
Img Enh + Oversample only	0.8140
Img Enh + Oversample + Geometric Aug	0.8069

Table 9: Effect of geometric image augmentation on validation macro F1. Augmentation reduces performance, likely due to disruption of text layout and composition.

A.5 Effect of Balanced Class Distribution on TF-IDF

Table 10 reports macro F1 for TF-IDF models before and after balanced class distribution.

Model	Imbalanced	Balanced (Syn)	Balanced (Deletion)
TF-IDF + SVM	0.7919	0.8412	0.8423
TF-IDF + LogReg	0.7995	**0.8427**	0.8401

Table 10: TF-IDF macro F1 before and after balanced class distribution (3,200 per class) and with additional text deletion augmentation. Gains are due to class balancing alone.

A.6 Multimodal Development Results

Table 11 summarizes multimodal model results under different pipeline configurations.

Model	Pipeline	Macro F1
BLIP-base (5ep)	None	0.7892
CLIP Multimodal	None	0.7871
CLIP Multimodal (15ep)	Img Enh + Oversample	0.8415
BLIP-base (5ep)	Img Enh + Oversample	0.7892
CLIP Multimodal (15ep)	Img Enh + Oversample + Text Aug	**0.8502**

Table 11: Multimodal model results under different pipeline configurations (development split).

A.7 All Text Models (Main Evaluation)

Table 12 reports validation macro F1 for all text-only models with and without preprocessing and augmentation.

Model	w/o Pre+Aug	w/ Pre+Aug
Base-sized models		
BERT-base	0.7962	0.8088
DeBERTa-v3-base	0.8031	0.7967
RoBERTa-base	0.8028	0.8200
Twitter-RoBERTa	0.8092	0.8178
XLM-RoBERTa-base	0.8096	0.8136
4-model ensemble	0.8151	0.8165
Large models (w/ Pre+Aug only)		
XLM-RoBERTa-large	—	0.8026
DeBERTa-v3-large	—	0.8146
RoBERTa-large	—	**0.8252**
TF-IDF models		
TF-IDF + SVM	0.8003	0.7902
TF-IDF + LogReg	0.8046	0.7996

Table 12: All text-only model validation macro F1. Pre+Aug = text preprocessing + random deletion augmentation.

A.8 All Image and Multimodal Models (Main Evaluation)

Table 13 reports validation macro F1 for image-only and multimodal models.

Model	Aug Target	Val F1
ConvNeXt-base	—	0.6603
EfficientNet-B3	—	0.6622
CLIP Vision (image only)	—	0.7156
CLIP Multimodal	3200	0.7957
CLIP Multimodal	4000	**0.8002**
CLIP Multimodal (TTA)	4000	0.7856
LLaVA-1.5-13B (zero-shot)	—	0.3447

Table 13: Image-only and multimodal model validation macro F1. Aug Target = balanced samples per class.

B Training Hyperparameters

As shown in Table 14, CLIP multimodal (ViT-B/32) uses a 2-layer MLP classification head over concatenated image and text embeddings

Setting	DeBERTa-v3-large	RoBERTa-large
Learning rate	1e-5	1e-5
Batch size	8	8
Grad. accum. steps	4	2
Epochs	5	5
Best epoch (val F1)	2	3
Optimizer	AdamW, weight decay 0.01	
Scheduler	Linear decay, warmup ratio 0.1	
Max seq. len	128 tokens	
Precision	fp16	
Early stopping	Patience = 3 (val macro F1)	

Table 14: Transformer training hyperparameters for the final submission run.

(dim=512 $\times$ 2), trained with AdamW (lr=2e-5, weight decay=0.01) and a ReduceLROnPlateau scheduler (factor=0.5, patience=1). Batch size=16, up to 10 epochs with early stopping (patience=3 on val macro F1); best checkpoint at epoch 2 (val F1=0.8058). Images resized to 224 $\times$ 224 with enhancement (contrast $\times$1.2, brightness $\times$1.1, sharpness $\times$1.1). No geometric augmentation was applied. Cross-entropy loss with class-balanced weights. All models trained on a single NVIDIA P100 (16 GB) via Kaggle.

C Ensemble Weights

Final ensemble weights are proportional to validation macro F1: DeBERTa-v3-large: 0.337, CLIP Multimodal: 0.332, RoBERTa-large: 0.332 (normalised to sum to 1.0). Soft voting averages class probability vectors; majority voting was also evaluated and produced identical results on several configurations, suggesting the models largely agree.

D Full Test-Phase Results

All results below are on the official test set (1,025 samples) unless marked † (validation set). Models are grouped by family. This appendix covers all 65 experimental configurations run during the test phase.

D.1 Text-Only Models

Table 15 lists all text-only model results on the official test set.

D.2 Image-Only Models

Table 16 lists all image-only model results on the official test set.

D.3 Multimodal Models

Table 17 lists all multimodal model results on the official test set.

Model	Pre+Aug	F1	Acc.	Prec.
TF-IDF models				
TF-IDF + LogReg	No	0.8046	0.8068	0.8061
TF-IDF + SVM	No	0.8003	0.8020	0.8032
TF-IDF + LogReg	Yes	0.7996	0.8020	0.8016
TF-IDF + SVM	Yes	0.7902	0.7922	0.7943
Base-sized transformers (no Pre+Aug)				
BERT-base	No	0.7962	0.8000	0.7971
RoBERTa-base	No	0.8028	0.8059	0.8036
Twitter-RoBERTa	No	0.8092	0.8137	0.8088
XLM-RoBERTa-base	No	0.8096	0.8127	0.8120
DeBERTa-v3-base	No	0.8031	0.8059	0.8066
Base-sized transformers (with Pre+Aug)				
BERT-base	Yes	0.8088	0.8117	0.8101
RoBERTa-base	Yes	0.8200	0.8215	0.8232
Twitter-RoBERTa	Yes	0.8178	0.8205	0.8193
XLM-RoBERTa-base	Yes	0.8136	0.8156	0.8164
DeBERTa-v3-base	Yes	0.7967	0.8010	0.7978
Large models (with Pre+Aug)				
XLM-RoBERTa-large	Yes	0.8026	0.8049	0.8062
DeBERTa-v3-large	Yes	0.8146	0.8185	0.8147
BERT-base (large run)	Yes	0.8147	0.8156	0.8196
RoBERTa-large	Yes	0.8252	0.8293	0.8250

Table 15: All text-only model results on the test set.

Model	F1	Acc.	Prec.
ResNet-50	0.6875	0.6907	0.6904
ConvNeXt-base	0.6603	0.6663	0.6612
EfficientNet-B3	0.6622	0.6663	0.6671
CLIP Vision (image only)	0.7156	0.7220	0.7178
ResNet-50 (with enh+oversample)	0.6784	0.6810	0.6841

Table 16: All image-only model results on the test set.

D.4 Ensemble Configurations

Table 18 lists all ensemble configurations explored; the final submission is bolded.

Model / Configuration	F1	Acc.
CLIP Multimodal (baseline, no aug)	0.7848	0.7873
CLIP Multimodal (aug=3200)	0.7957	0.8000
CLIP Multimodal (aug=4200)	0.7849	0.7873
CLIP Multimodal (TTA)	0.7856	0.7912
CLIP Multimodal (aug=4000)	**0.8002**	**0.8020**
BLIP-base (5ep)†	0.7892	0.7906
LLaVA-1.5-13B (zero-shot)	0.3447	0.3971

Table 17: All multimodal model results. †Validation F1 (test set unlabeled).

Ensemble Configuration	F1	Acc.
RoBERTa + CLIP + TF-IDF (weighted)	0.8252	0.8293
RoBERTa + CLIP + TF-IDF (majority)	0.8252	0.8293
TF-IDF + CLIP + RoBERTa + DeBERTa	0.8221	0.8244
CLIP + RoBERTa + DeBERTa (aug=3200)	0.8286	0.8312
Stacking (CLIP+RoBERTa+DeBERTa+BLIP)	0.8262	0.8283
RoBERTa + TF-IDF + CLIP (majority)	0.8299	0.8312
RoBERTa + TF-IDF + CLIP (weighted)	0.8299	0.8312
CLIP + RoBERTa only	0.8132	0.8156
CLIP + RoBERTa + DeBERTa (aug=4000)	**0.8306**	**0.8322**

Table 18: All ensemble configurations explored. Final submission is bolded.

CUET_SYNTHETICA@EEUCA 2026: Gated Cross-Modal Attention with Domain-Adapted Text Encoding for Vaccine-Critical Meme Detection

Sumaiya Zaman, Miftahul Jannat Rishta, Shiti Chowdhury
Department of Computer Science and Engineering,
Chittagong University of Engineering and Technology, Bangladesh
{u2104110, u2104019, u2004027}@student.cuet.ac.bd

Abstract

Vaccine-critical memes have emerged as a growing challenge for public health communication, combining images and text to spread misinformation in ways that are difficult to detect automatically. In this paper, we have described our system for the EEUCA 2026 Shared Task on Multimodal Vaccine-Critical Meme Detection, classifying memes from the VaxMeme dataset into Vaccine-Critical, Neutral and Pro-Vaccine categories. We have experimented with multiple text encoders and visual backbones, finding that Twitter-RoBERTa fused with CLIP ViT-L/14 through gated cross-modal attention has achieved a test macro F1 of 0.8357. We have further shown that domain-specific pretraining has outperformed larger general-purpose models, highlighting the importance of domain adaptation over raw model scale. Finally, our system has secured the 3rd position on the shared task leaderboard.

1 Introduction

Memes have evolved into powerful tools for spreading vaccine misinformation online, combining text and images in ways that are difficult to counter through traditional fact-checking methods. While researchers have made progress in detecting harmful memes broadly (Suryawanshi et al., 2020), vaccine-specific meme detection has remained largely underexplored (Naseem et al., 2023).
The shared task has provided the VaxMeme dataset, consisting of over 10,000 manually annotated vaccination-related memes labeled as Vaccine-Critical, Neutral and Pro-Vaccine. It has encouraged the development of multimodal systems that jointly leverage textual and visual information for fine-grained meme understanding (Naseem et al., 2023). In the EEUCA 2026 Shared Task on Multimodal Vaccine-Critical Meme Detection, we have experimented with multiple text encoders and visual backbones, finding that Twitter-RoBERTa paired with CLIP ViT-L/14 under gated cross-attention fusion has proven the most effective, achieving a macro F1 of 0.8357 on the official test set. The core contributions of our work are as follows:

- We have implemented a gated cross-attention fusion mechanism that has learned to balance text and image features for each meme individually.
- We have shown that Twitter-RoBERTa has outperformed larger general-purpose encoders, highlighting the value of domain-specific pretraining over raw model size.
- We have developed a multi-stage training pipeline incorporating weighted focal loss, a weighted ensemble and retraining on combined training and evaluation data.

Code is available at: https://github.com/Mif-taa/VaxMemeStance.

2 Related Work

Detecting harmful content in memes has become an increasingly important research problem as social media platforms have grown into primary channels for health-related discourse. Early approaches have tackled this problem using unimodal models, but these have consistently fallen short in capturing how the two modalities interact to construct meaning (Suryawanshi et al., 2020), motivating multimodal fusion strategies that have shown meaningful gains (Koutlis et al., 2023). Pretrained transformers such as BERT and RoBERTa have proven effective for social media text understanding (Devlin et al., 2019; Liu et al., 2019), while vision-language models like CLIP (Radford et al., 2021) and FLAVA (Singh et al., 2022) have pushed multimodal representation further. Twitter-RoBERTa (Barbieri et al., 2020) has demonstrated strong

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 133–140

performance on Twitter-oriented tasks through domain-specific pretraining and DINOv2 (Oquab et al., 2023) has expanded the toolkit for image feature extraction without text-image contrastive objectives. Most directly relevant to our work, Naseem et al. (2023) have developed a multimodal system for vaccine-critical meme detection on Twitter, providing the foundational benchmark for our work. Building on these works, our approach has addressed the challenge of capturing cross-modal interactions in memes by combining Twitter-RoBERTa, CLIP ViT-L/14 and OCR-extracted text, enabling more effective understanding of meme semantics for vaccine-critical detection.

3 Dataset & Task Description

This work has addressed the shared task on Multimodal Identification of Vaccine Content Stance on Social Media (Thapa et al., 2026b), organized as part of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA) (Hürriyetoğlu et al., 2026). The task has required classifying vaccine-related memes into one of three stance categories: Pro-Vaccine, Neutral and Vaccine-Critical, to support automatic detection of health misinformation and stance in multimodal social media content (Thapa et al., 2024, 2025).

The competition has been hosted on CodaBench[1] and participants have been evaluated on a held-out test set. The task has built upon prior work in multimodal vaccine-critical meme identification (Naseem et al., 2023) and concept-grounded detection of vaccine misinformation (Thapa et al., 2026a), with the annotation schema shared with the CrisisHateMM framework for hate speech analysis in crisis contexts (Bhandari et al., 2023).

3.1 The VaxMeme Dataset

The VaxMeme dataset has consisted of multimodal samples, each comprising a meme image, associated `post_text` and OCR-extracted `image_text`. The dataset has been drawn from prior WSDM and WebConf resources (Naseem et al., 2023; Thapa et al., 2026a) and has been partitioned into Train, Evaluation and Test splits. The label distribution across splits has been presented in Table 1.

The training set has contained 8,195 labelled memes, with a mild class imbalance toward the Pro-Vaccine category, followed by Vaccine-Critical and Neutral. The evaluation set has contained 1,024 labelled samples, while the test set has contained 1,025 labelled memes used for final leaderboard evaluation.

Labels	Train	Evaluation	Test
Vaccine-Critical	2,535	308	314
Neutral	2,461	327	316
Pro-Vaccine	3,199	389	395
Total	**8,195**	**1,024**	**1,025**

Table 1: Data distribution across Train, Evaluation, and Test sets.

4 System Overview

4.1 Problem Formulation

Each meme in the VaxMeme dataset has been assigned one of three stances: Vaccine-Critical (0), Neutral (1) or Pro-Vaccine (2). We have trained three classifiers and combined their predictions via weighted soft voting. The final label has been determined by the highest weighted average probability:

$$\hat{y} = \arg\max_{k} \sum_{i=1}^{3} w_i \, p_i(k \mid x), \quad w_i = \frac{f_i}{\sum_j f_j} \tag{1}$$

where f_i is the macro F1 score of model i, ensuring stronger models influence the final decision, especially for borderline Neutral cases.

4.2 Model Architecture

4.2.1 Text-Only Classifier

The text-only classifier has been built on the Twitter-RoBERTa model, a RoBERTa-base model pretrained on 124 million tweets and fine-tuned on sentiment, making it well-suited to the informal language of vaccine-related social media. A classification head has taken the average of the `[CLS]` token and the masked mean-pooled embeddings, has passed through dropout and a linear projection to three output classes:

$$\mathbf{h} = \frac{\mathbf{h}_{\text{CLS}} + \text{MeanPool}(\mathbf{H}, \mathbf{m})}{2}$$
$$\hat{p} = \text{softmax}(\mathbf{W}\,\text{Dropout}(\mathbf{h})) \tag{2}$$

where $\mathbf{H}$ is the full last hidden state and $\mathbf{m}$ is the attention mask.

4.2.2 Gated Multimodal Model

The multimodal model has fused the OCR-augmented text stream with a visual stream from

[1] https://www.codabench.org/competitions/12085/

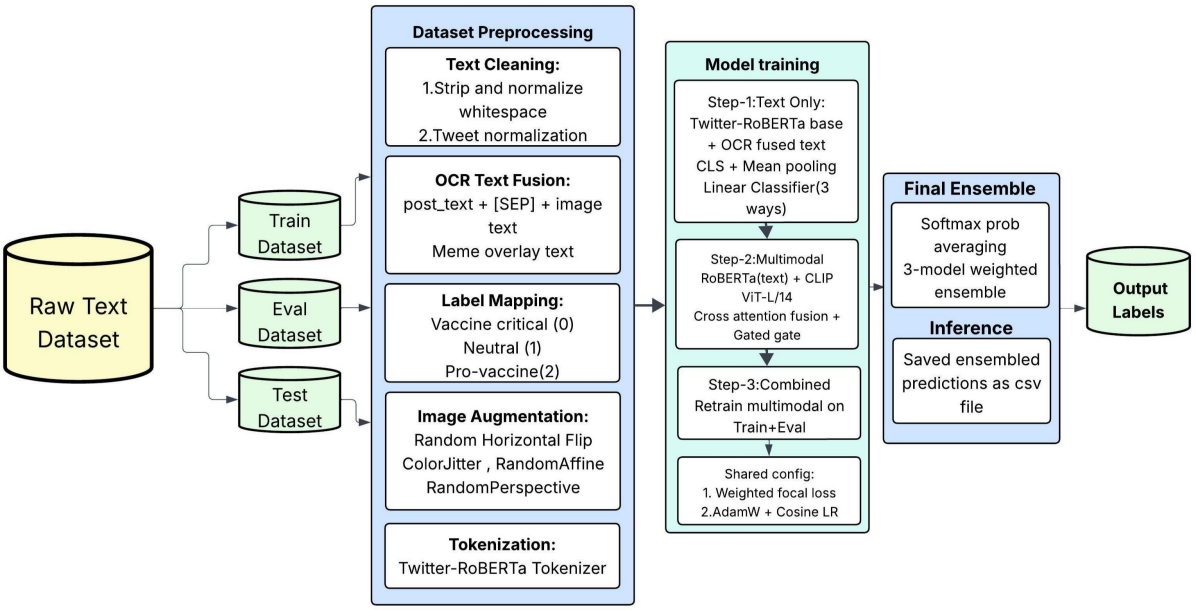


Figure 1: End-to-end pipeline of the 3-model ensemble for vaccine-meme detection.

the meme image. The text branch has reused the same Twitter-RoBERTa encoder and pooling strategy, while the visual branch has employed a CLIP ViT-L/14 encoder producing 768-dimensional image embeddings. For the first four epochs CLIP has been kept frozen to stabilise the fusion layers, after which CLIP parameters have been unfrozen with a conservative learning rate of 1×10^{-7} to allow gradual domain adaptation without catastrophic forgetting.

Both representations have been projected into a shared 512-dimensional fusion space via linear projections with LayerNorm and GELU activation. Cross-modal interaction has been modelled through 8-head cross-attention, where the text embedding has acted as query and the image embedding as key and value:

$$\mathbf{c} = \text{LayerNorm}\big(\text{CrossAttn}(\mathbf{q}_{\text{text}},\, \mathbf{k}_{\text{img}},\, \mathbf{v}_{\text{img}})\big) \tag{3}$$

A learned scalar gate $g \in (0, 1)$ has then blended the cross-attended representation with the raw text projection, allowing the model to down-weight the image when it has provided no additional discriminative signal:

$$\mathbf{z} = g \cdot \mathbf{c} \;+\; (1-g) \cdot \mathbf{h}_{\text{text}}, \quad g = \sigma\big(\mathbf{W}_g\,[\mathbf{h}_{\text{text}};\, \mathbf{c}]\big) \tag{4}$$

The fused representation $\mathbf{z}$ has been passed to a two-layer MLP classifier ($512 \to 256 \to 3$).

4.2.3 Combined-Data Model

The third model has been an additional instance of the gated multimodal architecture retrained on the union of the training and validation sets, exposing the model to every labelled example before test-time inference. Training has been run for a fixed 4 epochs with CLIP kept frozen throughout. Its logits have been included in the final ensemble using the validation F1 of the standalone multimodal model as a proxy weight, avoiding data leakage in the ensemble estimate.This proxy does not inflate ensemble performance because the standalone multimodal model's validation F1 was computed on a held-out evaluation set that was never used during combined-data training. The combined-data model thus contributes a conservatively weighted vote, and any overestimation of its weight would only marginally affect the final ensemble given that all three models produce similar probability distributions.

Figure 1 illustrates the pipeline.

5 Experimental Setup

5.1 Data Splits

The dataset has been divided into three splits: a training set of 8,195 samples, an evaluation set of 1,024 samples and a test set of 1,025 samples. During Stage 1 and Stage 2, the model has been trained on the training set and validated on the evaluation set. In Stage 3, the model has been retrained on the combined training and evaluation

set of 9,219 samples before running inference on the test set. The test set has been kept unseen throughout all training stages.

5.2 Preprocessing

5.2.1 Text

Each sample's textual content has been constructed by concatenating the post text with OCR-extracted meme overlay text separated by a special delimiter:

$$t = \texttt{post\_text} \parallel \texttt{[SEP]} \parallel \texttt{image\_text} \quad (5)$$

When no OCR text is available, only `post_text` has been used. This fusion has ensured that stance signals in meme overlays, like slogans and hashtags are preserved.

5.2.2 Image

Each meme image has been loaded, converted to RGB and preprocessed using the CLIP ViT-L/14 preprocessor, resizing and centre-cropping to 224×224 pixels with ImageNet normalization. During training, augmentations such as random flipping, colour jitter, affine transformation, and perspective distortion has been applied, with no augmentation during evaluation or inference. Missing images has been replaced with a zero tensor of the same shape.

5.3 Feature Extraction and Fusion

Textual features have been extracted using `cardiffnlp/twitter-roberta-base-sentiment-latest` with the final representation computed as the average of the `[CLS]` token and the mean-pooled token embeddings, both of dimensionality 768. Visual features have been extracted using the CLIP ViT-L/14 visual encoder, also producing 768-dimensional embeddings. Both representations have been mapped to a shared fusion space of dimension 512 via linear projection layers with LayerNorm, GELU activation, and dropout of 0.1.

Gated cross-modal attention has been used for fusion, where the text representation has served as the query and the image representation as the key and value in an 8-head multi-head attention layer. A learned gate has controlled the balance between the attended image features and the text features:

$$\mathbf{f} = g \cdot \mathbf{v}_{\text{cross}} + (1-g) \cdot \mathbf{f}_{\text{text}}, \quad g = \sigma\left(W[\mathbf{f}_{\text{text}}; \mathbf{v}_{\text{cross}}]\right) \quad (6)$$

The fused representation has then been passed through a two-layer classification head ($512 \rightarrow 256 \rightarrow 3$) with GELU activation and dropout.

5.4 Training

Training has proceeded in three stages. In Stage 1, a text-only model (Twitter-RoBERTa) has been trained for 5 epochs on the training set. In Stage 2, the full multimodal model has been trained for 8 epochs, with the CLIP encoder kept frozen for the first 4 epochs and gradually unfrozen with a low learning rate of 1×10^{-7} thereafter. In Stage 3, the multimodal model has been retrained for 4 epochs on the combined training and evaluation set to make use of all labelled data before running inference on the test set.All models have been optimised with AdamW using a cosine learning rate schedule with 10% linear warmup and gradient clipping at 1.0. The text encoder has used 2×10^{-5}, fusion and classification layers 5×10^{-5} and the CLIP encoder once unfrozen 1×10^{-7}. To address class imbalance, all models have been trained with a Weighted Focal Loss:

$$\mathcal{L} = -\sum_{k} w_k \, (1 - p_k)^{\gamma} \log p_k, \quad \gamma = 2 \quad (7)$$

The Neutral class has been up-weighted by 1.4 while other classes have been set to 1.0, with the focal term concentrating gradient on hard misclassified examples.

5.5 Ensemble and Inference

At inference, the best validation checkpoints of the text-only, multimodal and combined models have been used to produce softmax probability vectors. These have been blended as per Equation (1) and the `argmax` of the weighted average has determined the predicted stance label.

5.6 Parameter Setting

Table 2 lists the key hyperparameters for training the multimodal systems. All models have used the AdamW optimizer with a text encoder learning rate of 2e-5, cosine scheduling, and gradient clipping at 1.0, ensuring stable training and reducing overfitting.

5.7 Tools and Reproducibility

All experiments have been implemented in PyTorch using the Hugging Face Transformers library. A fixed random seed of 42 has been set across Python, NumPy and PyTorch for reproducibility. Training has been conducted on a GPU with mixed-precision via `torch.amp` and a batch size of 8.

Model	Text LR	Head LR	Optimizer	Batch Size	Epochs
RoBERTa-base + CLIP ViT-B/32	2e-5	5e-5	AdamW	16	5+6
BERT-base + CLIP ViT-B/32	2e-5	2e-4	AdamW	16	5+6
Twitter-RoBERTa + CLIP ViT-L/14	**2e-5**	**5e-5**	**AdamW**	**8**	**5+8+4**

Table 2: Key hyperparameters for multimodal model training.

5.8 Evaluation Metrics

All models have been evaluated using macro-averaged F1, precision and recall, treating each class equally regardless of frequency. The best-performing validation checkpoint has been retained for each model.

6 Results and Discussion

6.1 Task: Multimodal Vaccine-Critical Meme Detection

Table 3 has presented the performance of all models on the VaxMeme dataset. Among text-only models, RoBERTa-Large has achieved the highest macro F1 of 0.8165, followed by Twitter-RoBERTa at 0.8125 and RoBERTa-base at 0.8009. Image-only models have underperformed compared to text-only models, with CLIP ViT-B/32 achieving the best visual F1 of 0.7479, ahead of CLIP ViT-L/14 at 0.7179 and DINOv2-base at 0.7034.

Among multimodal models, Twitter-RoBERTa fused with CLIP ViT-L/14 has achieved the highest multimodal F1 of 0.8079 and accuracy of 0.8105. The weighted ensemble has further improved this to a macro F1 of 0.8090 and accuracy of 0.8115, with a final test F1 of 0.8357. To address classification challenges with the Neutral class, a weighted focal loss with a 1.4× Neutral upweight has been applied.

We acknowledge that Table 3 does not isolate the gate's individual contribution from the choice of ViT-L/14 backbone or the cross-attention structure itself. A standard cross-attention baseline (same backbone, no gate) was not included due to compute constraints, and we leave this ablation to future work. However, the gate's theoretical motivation dynamically suppressing weak visual signals is supported by the LIME and Integrated Gradients analyses in Section 8, which confirm that the model attends to task-relevant cues rather than uniform image features.

7 Error Analysis

7.1 Confusion Matrix

Figure 2 has shown the confusion matrix for our best system (Twitter-RoBERTa + CLIP ViT-L/14). The model has correctly identified 237 Vaccine-Critical, 258 Neutral and 336 Pro-Vaccine samples. The Vaccine-Critical class has had the most misclassifications, with 61 samples misclassified as Neutral and 10 as Pro-Vaccine.

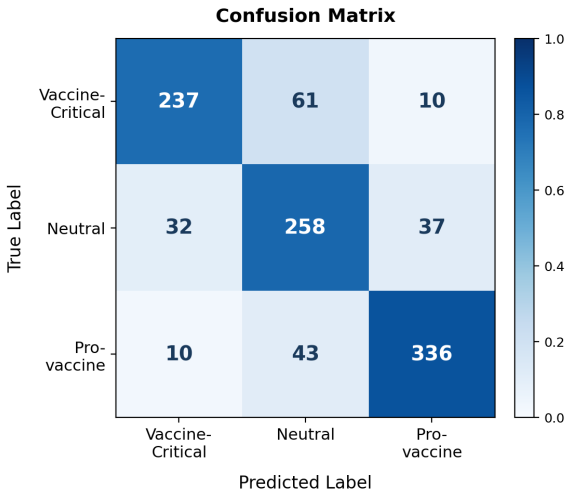


Figure 2: Confusion matrix for Twitter-RoBERTa + CLIP ViT-L/14

7.2 Failure Cases

An analysis of the misclassified samples has shown a clear pattern of errors across all three classes. The most difficult cases have involved the Vaccine-Critical class, where 61 samples have been misclassified as Neutral and 10 as Pro-Vaccine. This has suggested that the model has struggled to tell apart vaccine-critical content from ambiguous or sarcastic memes that look neutral on the surface.

The Neutral class has also shown confusion in both directions, with 32 samples misclassified as Vaccine-Critical and 37 as Pro-Vaccine. This has in-

Model	Classifier	P	R	F1	Acc
Unimodal (Text)	RoBERTa-base	0.8019	0.8003	0.8009	0.8047
Unimodal (Text)	Twitter-RoBERTa-base	0.8121	0.8138	0.8125	0.8145
Unimodal (Text)	RoBERTa-large	0.8174	0.8159	0.8165	0.8200
Unimodal (Image)	DINOv2-base	0.7033	0.7045	0.7034	0.7000
Unimodal (Image)	CLIP ViT-L/14	0.7191	0.7204	0.7179	0.7200
Unimodal (Image)	CLIP ViT-B/32	0.7496	0.7499	0.7479	0.7400
Multimodal	RoBERTa-base + CLIP ViT-B/32	0.7845	0.7834	0.7838	0.7881
Multimodal	BERT-base + CLIP ViT-B/32	0.7933	0.7926	0.7924	0.7969
Multimodal	Twitter-RoBERTa + CLIP ViT-L/14	0.8146	0.8058	0.8079	0.8105
Ensemble	**Twitter-RoBERTa + CLIP ViT-L/14**	**0.8132**	**0.8074**	**0.8090**	**0.8115**

Table 3: Performance comparison of unimodal, multimodal and ensemble classifiers on the validation set.

dicated that neutral memes have frequently shared image or text cues with both opposing stances, which has made boundary cases hard to classify correctly.

The Pro-Vaccine class has achieved the best recall of 86.38%, yet 43 samples have been misclassified as Neutral and 10 as Vaccine-Critical. These errors have mostly occurred in cases where pro-vaccine messaging has used mild or understated language, which has reduced the strength of the positive stance signal.

Overall, the confusion matrix has shown that most misclassifications have occurred between neighboring categories (Vaccine-Critical ↔ Neutral and Neutral ↔ Pro-Vaccine), rather than between the two opposing extremes (Vaccine-Critical ↔ Pro-Vaccine), which have accounted for only 20 of the 193 total misclassifications.

8 Explainability Analysis

We have applied two explainability techniques to our gated multimodal model for vaccine stance classification on the VaxMeme dataset:

- Integrated Gradients on token embeddings
- LIME superpixel explanations

Integrated Gradients has revealed token-level importance by computing attribution scores along the path from a baseline to the input embeddings. The method has consistently highlighted vaccine-critical keywords such as *antivax* and *side effect* as the strongest predictors, as has been shown in Figure 3. LIME has identified key visual regions by perturbing image superpixels while keeping the text input fixed, showing that symbolic imagery and text overlays have driven CLIP ViT-L/14 predictions, as has been shown in Figure 4. Together, both methods have confirmed that the model has learned task-relevant multimodal cues for vaccine stance classification.

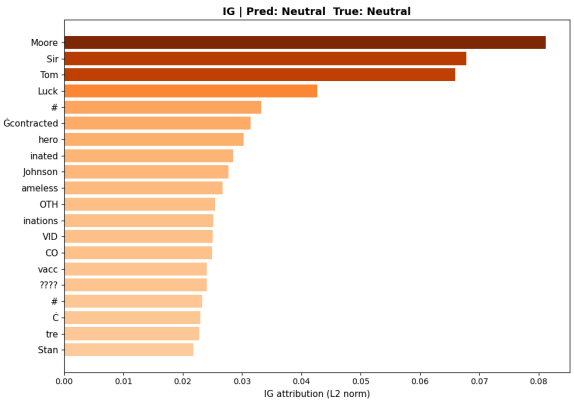


Figure 3: Integrated Gradients Token Attribution

9 Conclusion

Twitter-RoBERTa paired with CLIP ViT-L/14 under gated cross-modal attention has proven the strongest combination for vaccine-critical meme detection. Among the design choices, domain-specific pretraining, selective visual fusion and weighted focal loss have contributed the most to overall performance gains. The Neutral class has remained the most difficult category to classify correctly. Future work could explore larger vision-language models and multilingual training data to

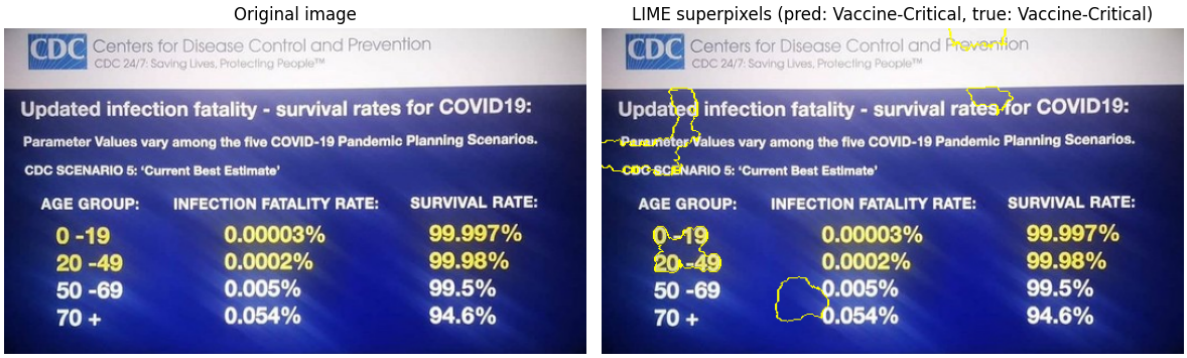


Figure 4: LIME Superpixel Importance Map

improve the system's robustness beyond English social media content.

Limitations

Despite strong overall results, the system has faced limitations due to CLIP's partial fine-tuning, which has caused the model to miss key visual cues in image-heavy memes. Furthermore, Twitter-RoBERTa has struggled with multilingual content and the Neutral class has remained the weakest performing category, which has reduced the system's applicability beyond Twitter-style content.

Ethical Considerations

All data has been drawn from the publicly available VaxMeme dataset and no personal data has been collected or stored. Our work has aimed to support public health monitoring and has been developed as a research prototype rather than a deployment-ready moderation tool. Any real-world use of this system should involve human oversight and domain expert review before application.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650. Association for Computational Linguistics.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

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)*.

Christos Koutlis, Manos Schinas, and Symeon Papadopoulos. 2023. MemeFier: Dual-stage modality fusion for image meme classification. *arXiv preprint arXiv:2304.02906*.

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*.

Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G. Dunn. 2023. A multimodal framework for the identification of vaccine critical memes on Twitter. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 706–714.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41.

Laxmi Thapa, Aryaman Jain, Lakshmojee Koduru, Surabhi Adhikari, Junaid Rashid, Jungeun Kim, Surendrabikram Thapa, and Usman Naseem. 2026a. Concept-grounded detection of vaccine misinformation in multimodal content using interpretable vision-language models. In *Companion Proceedings of the ACM on Web Conference 2026*.

Surendrabikram Thapa, Kritesh Rauniyar, Hariram Veeramani, Aditya Shah, Imran Razzak, and Usman Naseem. 2024. Did you tell a deadly lie? evaluating large language models for health misinformation identification. In *International Conference on Web Information Systems Engineering*, pages 391–405. Springer.

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. 2026b. Multimodal identification of vaccine content stance on social media. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

wenbin@EEUCA 2026: MoEs-VaxAgent, A Two-Stage Framework for Multimodal Vaccine Critical Meme Detection

Wenbin Shen
School of Cyber Science and Engineering
Nanjing University of Science and Technology
Nanjing, China
shenwenbin@njust.edu.cn

Abstract

Memes on social media have emerged as a crucial medium for disseminating vaccine-related viewpoints, yet their inherent irony, metaphor, and text-image misalignment pose significant challenges to automatic detection. In this paper, we propose MoEs-VaxAgent, a two-stage multimodal framework for vaccine critical meme detection. First, we design a dynamic routing Mixture-of-Experts module capable of adaptively capturing multi-granular semantic cues within memes. Second, to address hard samples located at the decision boundaries, we introduce an uncertainty-aware multi-agent rectification mechanism to perform a secondary detection on samples identified with low confidence in the first stage. In the EEUCA 2026 Shared Task on Multimodal Vaccine Critical Meme Detection, our system achieved a Macro F1-score of 0.8205, ranking 9th on the official leaderboard. Furthermore, we discuss various exploratory strategies evaluated during the competition and provide a detailed analysis of the model's performance.

1 Introduction

Internet memes have emerged as a highly influential medium for information dissemination within the public health sphere, demonstrating exceptional virality regarding topics such as COVID-19 vaccination. While memes can facilitate communication and encourage positive behaviors, they also serve as conduits for misinformation and skepticism. Memes frequently employ mechanisms such as image-text misalignment, irony, and deep cultural metaphors to convey stances (Kiela et al., 2020). These complex contextual associations and interactions between modalities pose severe challenges for automated detection.

Existing multimodal detection methods are broadly classified into two categories. The first category comprises discriminative detection methods based on multimodal features. These methods typically utilize pretrained models to extract multimodal features for classification (Wang et al., 2020; Naseem et al., 2024). Although these approaches have yielded promising results, they generally struggle to address the highly non-linear modal relationships inherent in memes and perform poorly on "metaphorical" hard samples. The second category involves Large Language Model (LLM) based agent detection methods (Hwang and Shwartz, 2023; Lin et al., 2024; Liu et al., 2025). While agents possess robust reasoning capabilities, performing comprehensive scans on massive social media datasets incurs prohibitive overhead and latency, making it difficult to meet the practical demands of large-scale public opinion monitoring.

To address these issues, we propose MoEs-VaxAgent, a two-stage classification framework. First, we design a dynamically routed Mixture-of-Experts module. This module integrates five heterogeneous experts and utilizes a Top-k gating mechanism to dynamically activate expert combinations. Second, to further enhance detection accuracy for hard samples, we introduce an uncertainty-aware agent correction mechanism. The system automatically identifies ambiguous samples with low confidence and delegates them to a text agent, a visual agent, and a judge agent for multi-perspective assessment to generate the final result.

The main contributions of this paper are summarized as follows:

- We propose MoEs-VaxAgent, a two-stage classification framework combining a Mixture-of-Experts model with multi-role agents.
- We evaluate our proposed MoEs-VaxAgent in the EEUCA 2026 Shared Task on Multimodal Vaccine Critical Meme Detection, achieving a Macro F1-score of 0.8205 and ranking 9th on the official leaderboard.

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 141–150

- We provide a comprehensive discussion of various exploratory strategies and conduct a detailed error analysis, offering practical insights into the challenges of multimodal stance detection.

2 Background

2.1 Task objective

Shared Task on Multimodal Vaccine Critical Meme Detection (VaxMeme) at EEUCA 2026 (Thapa et al., 2026b; Hürriyetoğlu et al., 2026) aims to develop models to automatically identify the stance of vaccine-related memes. Given that memes often convey information through image-text misalignment, irony, and metaphors, models require deep fusion of visual and textual modalities to capture fine-grained contexts. The task is defined as a three-class classification problem, adopting Macro F1-score as the primary ranking metric, and Accuracy, Precision, and Recall as auxiliary metrics.

2.2 Datasets

VaxMeme (Naseem et al., 2023; Thapa et al., 2026a; Bhandari et al., 2023) serves as the core benchmark dataset for this shared task. This dataset contains over 10,000 meme samples sourced from Twitter, with each sample consisting of an image and its corresponding embedded text or tweet text. It provides three fine-grained human-annotated categories, namely Pro-vaccine, Vaccine-critical, and Neutral. Following the official standardized partition, the dataset is divided into a training set comprising 8,195 samples, a validation set of 1,024 samples, and a test set of 1,025 samples.

Furthermore, competition rules permit participants to utilize external data to enhance model generalization or facilitate transfer learning. MM-CoVaR (Chen et al., 2021), a dataset regarding COVID-19 vaccine information within the field, can be employed for auxiliary research. Covering 2,593 news articles and 24,184 related tweets published between February 2020 and March 2021, its rich long-form narratives and detailed news reports provide domain background knowledge essential for comprehending short and highly context-dependent memes.

2.3 Related work

Multimodal Analysis of Vaccine-related Memes. Early research on vaccine-related public opinion primarily relied on pre-trained language models to capture textual sentiment (Zhang et al., 2020), or utilized domain-specific knowledge graphs to enhance the semantic understanding of vaccine-related tweets (Lovera et al., 2021). However, the inherent ironic nature of memes and the semantic misalignment between images and text limit the efficacy of unimodal approaches, leading research to gradually shift toward multimodal frameworks. MOMENTA (Pramanick et al., 2021) detects harmful memes through global and local perspectives, SeTa-Attn (Wang et al., 2020) employs a dual-attention mechanism specifically for modeling medical misinformation, and VaxMine (Naseem et al., 2024) reduces noise in user historical data via a collaborative mechanism. Furthermore, recent studies have also begun to extensively evaluate the potential and risks of using LLMs for identifying health misinformation (Thapa et al., 2024). Despite these advancements, most existing methods employ static fusion strategies, which struggle to adaptively weight the dynamically changing dominance of modalities across different samples, thereby restricting model performance when handling complex image-text dependencies.

Mixture-of-Experts in Classification. As an efficient conditional computation paradigm, the Mixture-of-Experts (MoEs) model achieves dynamic routing of input data through a gating network (Shazeer et al., 2017). In the multimodal domain, this dynamic mechanism is able to effectively address inter-modal heterogeneity, enabling the model to adaptively select the optimal inference path based on the semantic dominance within each sample. For instance, LIMoE (Mustafa et al., 2022) utilizes modality-specific experts to handle differences between images and text, while MMOE (Yu et al., 2024) designs specialized interaction experts to capture cross-modal relationships.

LLM-based Agents for Reasoning and Refinement. With the evolution of Large Language Models (LLMs), utilizing LLM-based agents for reasoning in complex computational social science tasks has emerged as a significant trend (Thapa et al., 2025). Unlike traditional classifiers, LLM-based frameworks such as Self-Refine (Madaan et al., 2023) and Multi-Agent Debate (Du et al., 2023) introduce multi-role interaction and iterative mechanisms, enabling multi-perspective scrutiny and correction of results. This paradigm performs exceptionally well when processing "hard samples" that involve irony, metaphors, or require deep cultural background.

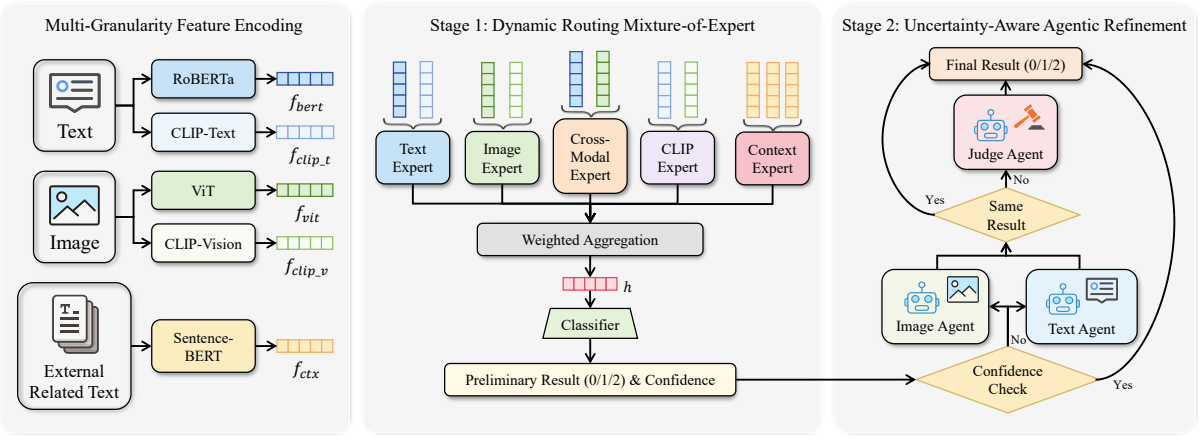


Figure 1: The overall architecture of the MoEs-VaxAgent. The framework consists of three parts: (1) **Multi-Granularity Feature Encoding**: utilizing RoBERTa, CLIP, ViT, and Sentence-BERT to extract comprehensive textual, visual, and external context features; (2) **Stage 1: Dynamic Routing Mixture-of-Experts**: which employs a Top-k gating mechanism to aggregate features from five heterogeneous experts (Text, Image, Cross-Modal, CLIP, and Context Experts) for preliminary classification; and (3) **Stage 2: Uncertainty-Aware Agentic Refinement**: where samples with low confidence are automatically delegated to a collaborative agent system (Text, Image, and Judge Agents) for final verification.

3 Methodology

3.1 Multi-Granularity Feature Encoding

To capture the rich multimodal semantics within the dataset, we employ a diverse set of pre-trained backbone networks for feature extraction. Specifically, we utilize RoBERTa (Liu et al., 2019) to extract textual semantic features f_{bert} and use ViT (Dosovitskiy, 2020) to capture visual features f_{vit} of the images. Meanwhile, to bridge the semantic gap between modalities, we utilize CLIP (Radford et al., 2021) to extract aligned representations for text and vision, denoted as f_{clip_t} and f_{clip_v}, respectively. Furthermore, we incorporate related text segments from the MMCoVaR dataset as external relevant knowledge and encode it using Sentence-BERT to obtain features f_{ctx}. Finally, these heterogeneous features are unified into a set $F = \{f_{bert}, f_{vit}, f_{clip\_t}, f_{clip\_v}, f_{ctx}\}$, serving as the input source for the subsequent Mixture-of-Experts module.

3.2 Dynamic Routing Mixture-of-Experts

To effectively integrate multi-granularity features, we design a MoEs module comprising five domain-specific experts. Each expert $E_i(\cdot)$ is constructed as an independent MLP, designed to map specific modal combinations into a unified latent space. We utilize feature vectors of different combinations, denoted as x_i, as inputs for the corresponding experts, as shown in Table 1.

Table 1: The input feature configurations for the five heterogeneous experts in the MoE module. The symbol $\oplus$ denotes the concatenation operation.

Expert Name	**Input** (x_i)
Text Expert	$x_1 = f_{bert} \oplus f_{clip\_t}$
Image Expert	$x_2 = f_{vit} \oplus f_{clip\_v}$
Cross-Modal Expert	$x_3 = f_{bert} \oplus f_{vit}$
Alignment Expert	$x_4 = f_{clip\_t} \oplus f_{clip\_v}$
Context Expert	$x_5 = f_{ctx}$

We employ a learnable gating network as a dynamic routing mechanism to calculate the activation weights for each expert. To reduce computational redundancy and focus on the most salient feature perspectives, we adopt a Top-k strategy (setting $k = 2$ in this study) to activate only the two experts with the highest scores. The final fused representation h is obtained by the weighted sum of the outputs from the activated experts, formulated as $h = \sum w_i E_i(x_i)$. Here, w_i represents the normalized gating score. This fused representation is subsequently fed into a classifier for final stance prediction.

3.3 Uncertainty-Aware Agentic Refinement

While the Mixture-of-Experts model provides a robust baseline for feature integration, it still faces limitations when handling ambiguous samples containing deep metaphors or irony. To address this, we design an uncertainty-aware multi-agent refine-

Table 2: Performance comparison on the EEUCA 2026 Shared Task leaderboard (Top 20).

Rank	Participant	Evaluation Indicators			
		F1 Macro	Accuracy	Precision	Recall
1	lili12-637947	0.8494	0.8517	0.8494	0.8517
2	wangxiuxian-637268	0.8389	0.8420	0.8386	0.8409
3	rishta_19-611897	0.8357	0.8390	0.8383	0.8359
4	_alexcristea-636983	0.8340	0.8380	0.8338	0.8351
5	sumaiya_110-594217	0.8332	0.8361	0.8345	0.8340
6	anchy-637928	0.8308	0.8341	0.8309	0.8309
7	myname-637930	0.8308	0.8341	0.8309	0.8309
8	quasar-637336	0.8306	0.8322	0.8331	0.8324
9	**wenbin-634065 (Ours)**	**0.8205**	**0.8244**	**0.8205**	**0.8218**
10	naturia_beast-636958	0.8201	0.8244	0.8212	0.8209
11	vinaybabu-637935	0.8184	0.8215	0.8216	0.8190
12	ratpier-637076	0.8150	0.8176	0.8170	0.8161
13	yjwong1999-494691	0.8122	0.8137	0.8189	0.8141
14	linus-637363	0.8105	0.8137	0.8106	0.8123
15	havnis-636808	0.8067	0.8117	0.8080	0.8083
16	alishba-wazir-604227	0.8067	0.8088	0.8132	0.8071
17	zmin123-553584	0.7997	0.8039	0.8005	0.8013
18	lin123-637530	0.7994	0.8039	0.7992	0.8007
19	barkion-636765	0.7976	0.7990	0.8080	0.7986
20	merrli-636903	0.7972	0.7990	0.8058	0.7982

ment mechanism. First, the system identifies "hard samples" with low MoE prediction confidence based on class-specific confidence thresholds (set to 0.5 in this study). Subsequently, we construct a Text Agent and a Visual Agent using Large Language Models to independently analyze the text and images within the dataset. To resolve potential conflicts arising from uni-modal perspectives, we introduce a "Divergence-Arbitration" strategy. If the predictions of both agents are consistent, they are directly adopted; otherwise, a Judge Agent is activated to synthesize the multimodal context for a final verdict, thereby achieving precise correction for long-tail complex samples.

4 Experimental Setup

4.1 Data Split and Augmentation

This study utilizes the VaxMeme dataset provided by the official EEUCA 2026 Shared Task. The dataset partition comprises a training set of 8,195 samples, a validation set of 1,024 samples, and a test set of 1,025 samples. In the data preprocessing phase, we removed all URL links from the text to reduce noise and concatenated the `post_text` with the `image_text` to form a complete text input.

Furthermore, we employ the MMCoVaR dataset for retrieval augmentation. To address the issue of excessive length in MMCoVaR source texts, we first split the texts by newline characters and filtered out segments with lengths less than 100 or greater than 1000, ensuring the semantic integrity and moderate length of the context segments. Upon constructing this external knowledge base, we performed semantic retrieval for each VaxMeme sample, selecting the Top-k (set to $k = 3$ in this study) text segments with the highest similarity as relevant knowledge to be input into the model alongside the original image and text.

4.2 Implementation Details

Experiments were implemented based on the PyTorch framework on NVIDIA GPUs. We initialize the text, visual, and alignment encoders using RoBERTa `[roberta-base]`[1], ViT `[google/vit-base-patch16-224]`[2], and CLIP `[openai/clip-vit-base-patch32]`[3], respectively. Additionally, we utilize Sentence-BERT `[sentence-transformers/all-MiniLM-L6-v2]`[4] to process external related knowledge.

During the training phase, the model adopts the AdamW optimizer with an initial learning rate set to 1×10^{-5} and a weight decay of 1×10^{-3}. We employ a Cosine Annealing strategy to adjust the learning rate, with a minimum learning rate of 1×10^{-6}. The batch size is set to 64, the number of epochs is 50, and label smoothing ($\epsilon = 0.1$) is used to prevent overfitting.

[1] huggingface.co/roberta-base
[2] huggingface.co/google/vit-base-patch16-224
[3] huggingface.co/openai/clip-vit-base-patch32
[4] huggingface.co/sentence-transformers/all-MiniLM-L6-v2

In the multi-agent refinement phase, we invoke the Tongyi Qianwen model via the DashScope API, employing Qwen `[qwen-plus]` as the Text Agent and Qwen-VL `[qwen-vl-max]` as both the Visual Agent and the Judge Agent to handle the reasoning and arbitration of low-confidence samples.

5 Experimental Results

5.1 Leaderboard Results

We submitted the predictions of MoEs-VaxAgent to the official evaluation platform of the EEUCA 2026 Shared Task on Multimodal Vaccine Critical Meme Detection. The final results are presented in Table 2, where we ranked 9th on the official leaderboard with a Macro F1-score of 0.8205.

5.2 Benchmark Results

Given that the ground truth labels for the test set are not publicly available, all our baseline comparison experiments were conducted on the validation set. To evaluate the performance of existing models in vaccine meme detection, we tested multiple groups of mainstream models, including uni-modal text encoders (BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019)), visual encoders (ResNet (He et al., 2016), ViT (Dosovitskiy, 2020)), and multimodal pre-trained models (CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023)). The detailed results are illustrated in Figure 2.

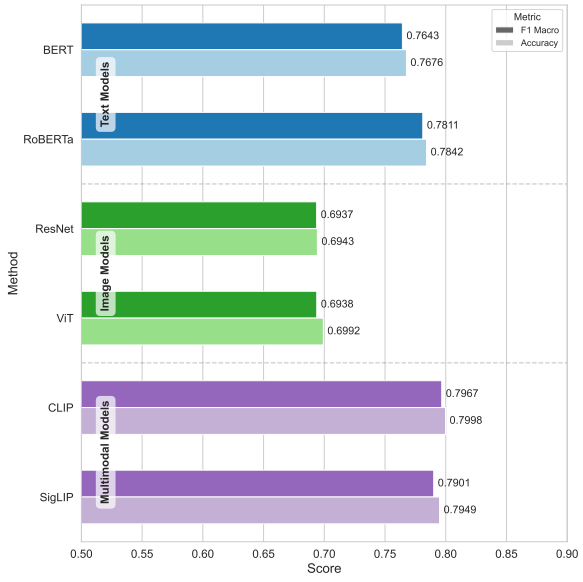


Figure 2: Performance comparison of various baseline models on the validation set.

Table 3: Ablation study results on the validation set. We report Macro-F1 (F1), Accuracy (Acc), Precision (Prec), and Recall (Rec) score.

Method	F1	Acc	Prec	Rec
w/o CLIP	0.786	0.788	0.786	0.785
w/o Context	0.799	0.802	0.798	0.799
Text Only	0.797	0.802	0.798	0.800
Image Only	0.727	0.729	0.729	0.726
Stage 1 Only	0.814	0.816	0.813	0.815
MoEs-VaxAgent (Full)	**0.825**	**0.828**	**0.825**	**0.826**

5.3 Ablation Study

To validate the effectiveness of the key components within MoEs-VaxAgent, we conducted a series of ablation experiments on the validation set, as presented in Table 3. The results indicate that methods based on multimodal features outperform uni-modal baselines. Specifically, while text features play a dominant role in stance determination, visual features provide valuable complementary information. The absence of CLIP features or external context leads to a decline in model performance, demonstrating the necessity of cross-modal aligned representations and domain background knowledge for enhancing classification effectiveness. Furthermore, the complete model, incorporating the uncertainty-aware multi-agent refinement mechanism, achieves further predictive improvements over the Stage 1, thereby validating the effectiveness of the overall framework.

6 Discussion

6.1 Exploration of Strategies

To explore the upper bounds of performance, we extensively evaluated a variety of mainstream strategies before finalizing the proposed architecture. Although certain methods yielded near-optimal results (with the highest achieving a Macro-F1 of 0.8147 and an Accuracy of 0.8185), none were able to breach the performance bottleneck of 0.815 Macro-F1. We categorize these exploratory attempts into the following three groups (listing only primary methodologies and omitting minor variations):

Data-Centric Strategies. Focusing on data quality, we investigated various data filtering and augmentation schemes. Counter-intuitively, results indicated that so-called "noisy" data often constitute critical decision boundaries; removing them compromised data diversity and degraded performance:

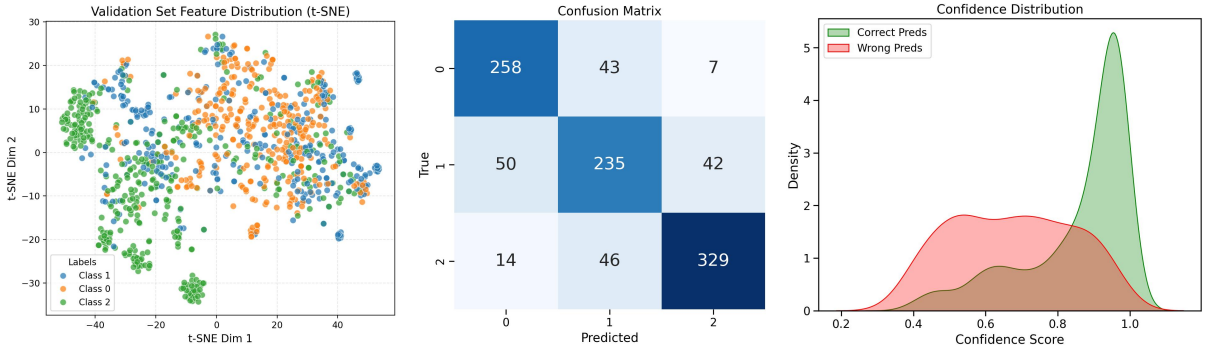


Figure 3: Visualization of model performance on the validation set. The figure displays the t-SNE feature distribution (Left), the Confusion Matrix (Middle), and the Confidence Distribution (Right) distinguishing between correct and incorrect predictions.

- *Low-quality Data Removal:* Attempting to clean and remove data deemed low-quality from the training and validation sets.
- *Greedy Data Selection:* Splitting the data into K folds and using a greedy strategy to dynamically select subsets that improve performance.
- *Pseudo-labeling Self-training:* Running two rounds of models, utilizing high-confidence predictions from the first round as pseudo-labels for secondary training.

Model Architecture and Ensemble Variants. We initially attempted to enhance the robustness of discriminative models through ensemble learning and architectural adjustments. However, we found that while these methods increased inference costs, they failed to fundamentally address the alignment issues of complex multimodal semantics:

- *Backbone Replacement:* Substituting different pre-trained backbones to seek better feature representations, particularly those explicitly fine-tuned on pandemic-related data.
- *Integration of Deep Learning and Machine Learning:* Extracting deep learning features and feeding them into traditional machine learning classifiers such as XGBoost, KNN, or Random Forest.
- *Task Decomposition and Fusion:* Transforming the three-class problem into three "One-vs-Rest" binary classifiers trained separately, followed by weighted fusion.

Optimization Objectives and Training Strategies. At the optimization level, we attempted to improve the model's ability to learn hard samples by adjusting loss functions and introducing auxiliary tasks. However, experiments showed that mere adjustments to optimization objectives were insufficient to bridge the cognitive gap in irony detection:

- *Loss Function Improvement:* Introducing Focal Loss and Supervised Contrastive Loss to address class imbalance.
- *Fine-tuning Strategies:* Attempting to partially unfreeze feature extractors for fine-tuning, as well as adjusting hyperparameters like learning rates and model dimensions.
- *Multi-task Learning:* Incorporating Domain Prediction as an auxiliary task optimized jointly with the main classification task to reinforce feature separability.

6.2 Analysis of Performance

We conducted a visual analysis of the model's performance in the first stage on the validation set, as illustrated in Figure 3. The t-SNE scatter plot and confusion matrix collectively reveal that the feature boundaries for the Pro-vaccine category are distinct, whereas significant feature entanglement and mutual misclassification exist between the Vaccine-critical and Neutral categories. Furthermore, the confidence distribution plot indicates that correct predictions are highly concentrated within high-confidence intervals, while erroneous predictions are primarily distributed across low-to-medium confidence ranges. This statistical phenomenon underpins our "uncertainty-aware" strategy, suggesting that filtering "hard samples" located at ambiguous boundaries via confidence thresholds and delegating them to Agents for refinement represents an optimal balance between computational cost and error correction efficiency.

6.3 Cost-Benefit Analysis of Stage 2

To evaluate the cost and benefit of the Agent stage, we analyzed the validation set consisting of 1,025 samples. The results show that only 149 samples (about 14.5%) triggered the Agent refinement. This means that the fast model in the first stage efficiently processed over 85% of the data, keeping the overall system latency relatively low. Among the 149 samples that entered the second stage, the Text and Visual Agents produced the same prediction for 76 samples (about 51.0%), which were then directly adopted. Only the remaining 73 samples with diverging predictions activated the Judge Agent for a final decision. These figures indicate that our Agent stage can achieve performance improvements at a relatively low additional cost.

7 Conclusion

In this study, we propose MoEs-VaxAgent, a multimodal classification framework designed to address the complex semantic challenges inherent in vaccine memes. By integrating a dynamic routing-based Mixture-of-Experts module with an uncertainty-aware multi-agent refinement mechanism, our approach not only effectively captures multi-granularity modal features but also leverages the reasoning capabilities of Large Language Models to successfully resolve hard samples situated at decision boundaries. Ranking 9th in the EEUCA 2026 Shared Task demonstrates the effectiveness of our framework in handling high-context and ironic memes. We also document our exploratory strategies and error analysis to share our practical experiences.

Limitations

Overconfident Misclassification. Our multi-agent refinement mechanism relies strictly on the uncertainty threshold defined in the first stage. If the MoE model assigns an excessively high confidence score to an incorrect prediction, the sample will bypass the refinement mechanism and be directly output as an error. The current system still has room for optimization regarding confidence calibration; relying solely on Softmax probabilities as a metric for uncertainty may lack robustness.

Inference Latency and Computational Cost. Although we adopted a two-stage strategy to avoid utilizing the Agent for every sample, the feature extraction process necessitates the concurrent execution of multiple backbone models, such as RoBERTa, ViT, and CLIP. Additionally, the second stage relies on API calls to external LLMs, which introduces inevitable inference latency and computational overhead. Consequently, the current framework is more suitable for offline analysis and may face challenges in streaming media monitoring scenarios that demand high real-time performance.

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.

Mingxuan Chen, Xinqiao Chu, and KP Subbalakshmi. 2021. Mmcovar: multimodal covid-19 vaccine focused data repository for fake news detection and a baseline architecture for classification. In *Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 31–38.

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.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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)*.

EunJeong Hwang and Vered Shwartz. 2023. MemeCap: A dataset for captioning and interpreting memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages

1433–1445, Singapore. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM Web Conference 2024*, pages 2359–2370.

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*.

Ziyan Liu, Chunxiao Fan, Haoran Lou, Yuexin Wu, and Kaiwei Deng. 2025. MIND: A multi-agent framework for zero-shot harmful meme detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 923–947, Vienna, Austria. Association for Computational Linguistics.

Fernando Andres Lovera, Yudith Coromoto Cardinale, and Masun Nabhan Homsi. 2021. Sentiment analysis in twitter based on knowledge graph and deep learning classification. *Electronics*, 10(22):2739.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback, 2023. *URL https://arxiv.org/abs/2303.17651*.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. 2022. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576.

Usman Naseem, Adam G Dunn, Matloob Khushi, and Jinman Kim. 2024. Vaccine misinformation detection in x using cooperative multimodal framework. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4034–4042.

Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G Dunn. 2023. A multimodal framework for the identification of vaccine critical memes on twitter. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 706–714.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Laxmi Thapa, Aryaman Jain, Lakshmojee Koduru, Surabhi Adhikari, Junaid Rashid, Jungeun Kim, Surendrabikram Thapa, and Usman Naseem. 2026a. Concept-grounded detection of vaccine misinformation in multimodal content using interpretable vision-language models. In *Companion Proceedings of the ACM on Web Conference 2026*.

Surendrabikram Thapa, Kritesh Rauniyar, Hariram Veeramani, Aditya Shah, Imran Razzak, and Usman Naseem. 2024. Did you tell a deadly lie? evaluating large language models for health misinformation identification. In *International Conference on Web Information Systems Engineering*, pages 391–405. Springer.

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. 2026b. Multimodal identification of vaccine content stance on social media. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Zuhui Wang, Zhaozheng Yin, and Young Anna Argyris. 2020. Detecting medical misinformation on social media using multimodal deep learning. *IEEE journal of biomedical and health informatics*, 25(6):2193–2203.

Haofei Yu, Zhengyang Qi, Lawrence Keunho Jang, Russ Salakhutdinov, Louis-Philippe Morency, and Paul Pu Liang. 2024. Mmoe: Enhancing multimodal models with mixtures of multimodal interaction experts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10006–10030.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

Li Zhang, Haimeng Fan, Chengxia Peng, Guozheng Rao, and Qing Cong. 2020. Sentiment analysis methods for hpv vaccines related tweets based on transfer learning. In *Healthcare*, volume 8, page 307. MDPI.

Table 4: The specific prompt templates designed for the multi-agent refinement mechanism.

Agent Role	Prompt Template
Text Agent	You are an expert in public health and social media analysis. Your task is to classify the stance of COVID-19 vaccination based on the text extracted from a meme. Input: Text: "{text}" Labels and Detailed Definitions: **0: Vaccine critical** • *Conspiracy*: Implies malicious intent by "authorities" or "pharmaceutical companies", or contains content related to "bioweapons", "crimes against humanity", etc. • *Co-opted Slogans*: Uses human rights or feminist slogans to oppose vaccines, shifting the focus from health to "resistance against control". • *Malicious Interpretation*: Shares news and adds comments implying the vaccine is a lie or ineffective. • *Natural Immunity*: Mocks the necessity of vaccines, claiming natural immunity is better. **1: Neutral** • *Raw News and Data*: Shares news headlines, charts, etc., without explicit personal commentary. • *Relevant Information*: Job postings, queuing situations, or statements like "I am waiting for my turn" without an obvious emotional tone. • *Criticism but Unrelated to Vaccines*: Complaining about "lockdowns", "censorship", or attacking the mandate system rather than claiming the vaccine itself is toxic, is usually neutral. **2: Pro-vaccine** • *Social Rewards*: Links vaccination to returning to normal life, dating, or travel. • *Mocking Anti-vaxxers*: Memes that satirize conspiracy theorists. • *Education and Progress*: Debunks rumors, explains definitions, or celebrates high vaccination rates.
Visual Agent	You are an expert in analyzing internet memes and visual rhetoric. Your task is to classify the stance of COVID-19 vaccination based on the visual content and text visible in the image. Input: Image: "{image}" Labels and Detailed Definitions: **0: Vaccine critical** • *Conspiracy Images*: Depicts sinister images of "authorities" or "pharmaceutical companies", or visually implies "depopulation", "gene therapy", or "bioweapons". • *Visual Metaphors*: Uses visual elements to associate vaccination with negative, oppressive concepts (e.g., control, submission). • *Screenshots with Malicious Text Overlays*: News screenshots accompanied by text overlays visually implying the vaccine is a hoax. **1: Neutral** • *Untampered Screenshots/Charts*: Pure news headlines, charts, or data screenshots without visual tampering or conspiracy markers. • *Simple Relevant Images*: Ordinary pictures of clinic lines, job postings, or vaccine vials. **2: Pro-vaccine** • *Positive Lifestyle Images*: Visually links vaccines to returning to normal life (travel, social events, hugging). • *Images Mocking Anti-vaxxers*: Visually satirizes conspiracy theories. • *Education/Milestones*: Infographics celebrating vaccination milestones or explaining how vaccines work.
Judge Agent	You are the final judge. The previous two experts (Text Expert and Visual Expert) disagreed. Your task is to make the final decision on the stance of COVID-19 vaccination. Input: Text: "{text}" Image: "{image}" Labels and Detailed Definitions: • **0: Vaccine critical.** Contains conspiracy theories (textual or visual implication of malicious authorities, bioweapons, etc.), co-opted slogans, malicious interpretation/tampering of news screenshots, or promotes natural immunity. • **1: Neutral.** Pure news/data screenshots, unbiased logistical information like queuing/recruitment, or solely criticizing mandate policies/censorship systems without attacking the vaccine itself. • **2: Pro-vaccine.** Promotes social rewards brought by vaccination (returning to normal life), visually or textually mocks anti-vaxxers, or educates/celebrates vaccine progress.

thaulab@EEUCA 2026: Who Said What to Whom? A Targeting-Aware Neural-Symbolic Pipeline for Gaming Toxicity Detection

Anmol Guragain[*], **Marcos Estecha Garitagoitia,**
Luis Fernando D'Haro Enríquez, Ricardo Córdoba
ETSI de Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain
[*]anmol.g@upm.es (Corresponding author)

Abstract

This paper describes our system for the EEUCA 2026 Shared Task on toxicity classification in gaming chat. We implement a three-stage pipeline combining an ensemble of two compact transformers (DeBERTa-v3-base, 184M; XLM-RoBERTa-base, 278M) with a Linguistically-Informed Mediator (LIM) that resolves inter-model disagreements through corpus-backed lexical normalization, class-conditional unigram scoring, multilingual profanity detection, and agentive targeting analysis grounded in speech act theory. The LIM specifically targets the minority classes (Hate & Harassment, Threats, and Extremism), which are the most safety-critical categories in real-world gaming moderation. To address the extreme class imbalance (1,450:1 Non-toxic to Extremism ratio), we introduce a two-stage data augmentation strategy using only the provided training data. Our system achieves a Macro F1 of 0.6441 and accuracy of 0.9062 on the official test set, ranking 3rd in Macro F1 and 1st in accuracy among all teams. The proposed pipeline is domain-portable: adapting to other gaming platforms requires substituting only the game-specific entity lexicon. Code is publicly available at https://github.com/Anmol2059/thaulab_EEUCA.

1 Introduction

Online gaming platforms host millions of real-time text interactions daily, and toxic behavior in these environments has been linked to serious consequences including cyberbullying, psychological harm, and player attrition (Parihar et al., 2021). A recent systematic review of 64 studies confirms that cyberbullying in multiplayer games is associated with anxiety, depression, and social withdrawal (Hu et al., 2025), and empirical evidence shows that toxic behavior propagates virally among teammates (exposure to toxic teammates increases a player's own toxicity likelihood by up to 30×), amplifying its reach when left undetected (Morrier et al., 2024).

The EEUCA 2026 Shared Task on Gaming Toxicity (Thapa et al., 2026; Hürriyetoğlu et al., 2026) introduces a six-class classification benchmark derived from World of Tanks chat logs (Naseem et al., 2025), annotated following the directed/undirected hate speech framework of Bhandari et al. (2023). The dataset poses three key challenges: extreme class imbalance (81.0% Non-toxic vs. 0.06% Extremism), multilingual content spanning 10+ languages, and domain-specific lexical ambiguity where violent vocabulary ("kill", "destroy") carries non-violent illocutionary force.

We implement a three-stage system combining neural ensemble classification with a rule-based Linguistically-Informed Mediator (LIM), following evidence that logical rules provide complementary signal to neural hate speech classifiers (Clarke et al., 2023; Awasthi et al., 2020). Our contributions are: (1) a two-stage augmentation strategy (confusion-pair-driven and contrastive boundary generation) that improves Macro F1 by +9.7% relative using only the provided data; (2) a LIM module grounded in speech act theory (Austin, 1962; Searle, 1969) that resolves ensemble disagreements through four interpretable, corpus-backed components; (3) empirical evidence that even multilingual transformers exhibit residual blind spots on domain-specific non-Latin profanity (22.6% of Hate & Harassment contains Cyrillic); and (4) demonstration that general-purpose toxicity models (toxic-bert) fail catastrophically in the gaming domain (Macro F1 = 0.3154), showing that gaming chat is a distinct linguistic register that requires domain-specific handling.

Related work. Recent NLP approaches to gaming toxicity include domain-adaptive pretraining of RoBERTa with match metadata for DOTA 2 and Call of Duty (Schurger-Foy et al., 2025), and hybrid

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 151–160

architectures combining LLM-generated embeddings with lightweight classifiers for Twitch moderation (Ansari et al., 2026). For class imbalance in hate speech detection, Zhang et al. (2024) show that focal loss (Lin et al., 2017) consistently yields peak performance, motivating our loss function choice. LLM-based data augmentation has proven effective for hate speech minority classes (Li et al., 2026), supporting our two-stage augmentation strategy (§2.1). The GameTox dataset (Naseem et al., 2025) additionally provides intent and slot filling annotations, but these labels were not released for the shared task, limiting participants to the six-class toxicity schema. Annotation disagreement is a recognized challenge in hate speech classification (Dehghan et al., 2025; Bhandari et al., 2023); we quantify its extent in this dataset in §D.

2 Task, Dataset, and Augmentation

The shared task (Thapa et al., 2026; Hürriyetoğlu et al., 2026) requires classifying World of Tanks chat into six categories: **Non-toxic** (0), **Insults** (1), **Other Offensive** (2), **Hate & Harassment** (3), **Threats** (4), and **Extremism** (5), evaluated by Macro F1. Table 1 shows the class distribution; the dataset contains 42,959 training, 5,367 validation, and 5,375 test samples with extreme imbalance (Non-toxic to Extremism ratio of 1,450:1). Messages are very short (median 2–4 tokens) and 6.9% contain Cyrillic script, rising to 22.6% in Hate & Harassment.

2.1 Two-Stage Augmentation

We augment minority classes without external data using a two-stage strategy (Figure 1; prompt templates in Appendix C).

Stage A uses a seed model (M0, DeBERTa-v3-base trained on original data) to identify confused class pairs: for each validation sample, we record the two highest-probability classes from M0 and flag the pair as a confusion boundary when the second-highest probability exceeds 0.15 (set empirically), indicating non-trivial model uncertainty between the two classes (Swayamdipta et al., 2020). This threshold was selected based on validation-set Macro F1 evaluated across candidate values $\{0.10, 0.15, 0.20, 0.25\}$; 0.15 maximized minority-class recall without introducing excess noise into the augmentation pool, as lower values produced near-duplicate confusion pairs while higher values missed meaningful boundary cases. Claude Opus 4.6 then generates synthetic samples targeting these confusion boundaries. This yields augmentation for Other Offensive, Hate & Harassment, and Threats.

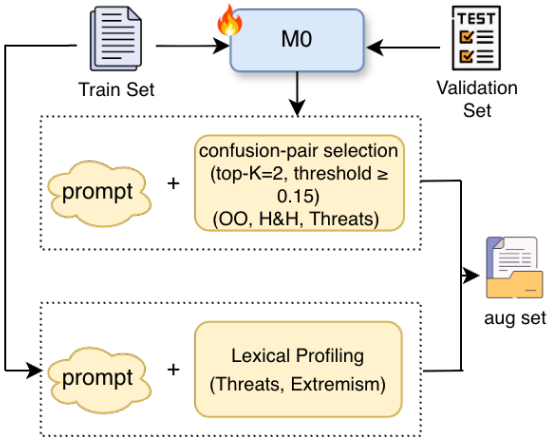


Figure 1: Two-stage augmentation. Stage A: confusion-pair-driven. Stage B: contrastive boundary with lexical profiling.

Class	Original	Added	Final
Non-toxic	34,797	0	34,797
Insults	5,925	0	5,925
Other Offensive	1,874	34	1,908
Hate & Harassment	279	155	434
Threats	60	235	295
Extremism	24	207	231
Total	42,959	631	43,590

Table 1: Training set class distribution before and after augmentation.

Stage B addresses Extremism (n=24) and supplemental Threats (n=60), which are too rare for confusion-pair analysis. We apply contrastive boundary augmentation: (1) mine class-discriminative tokens at $\geq 5\times$ frequency ratio using $P(c \mid w)$, the fraction of training messages containing word w that belong to class c (the same statistic used in the LIM's unigram scoring, §3.3.2); (2) generate cross-lingual variants and unmask leet-speak (e.g., `naz1`→`nazi`); (3) verify that generated samples fall within valid similarity bounds to real training data via cosine similarity in a TF-IDF subspace restricted to the mined discriminative vocabulary.

Table 1 shows the result: 631 synthetic samples, improving Macro F1 by +9.7% relative over M0.

3 System Architecture

Our system is a three-stage pipeline (Figure 2); hyperparameters for all models are in Appendix A.

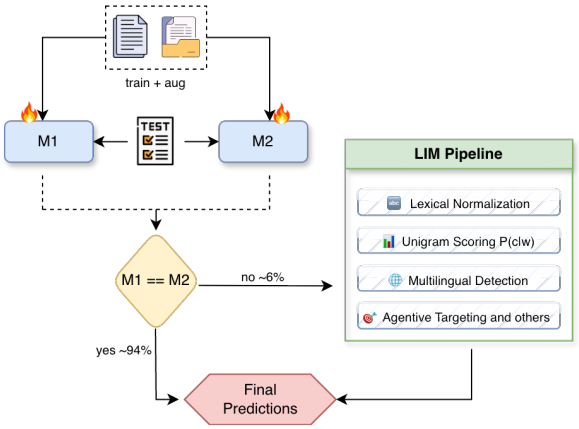


Figure 2: System pipeline. M1 (DeBERTa) and M2 (XLM-R) produce predictions; agreements (∼94%) are accepted, disagreements (∼6%) are refined by the LIM.

Model	F1	Acc	Prec	Rec
M0: DeBERTa (seed)	.5234	.8945	.5463	.5156
M1: DeBERTa (aug.)	.5742	.8979	.6042	.5538
M2: XLM-RoBERTa	.5613	.8910	.5670	.5730
M3: BERT-base	.5439	.8947	.5252	.5811
M4: toxic-bert	.3154	.7518	.3187	.5497

Table 2: Individual model results. M0/M1 share the DeBERTa architecture (seed vs. augmented). M1 and M2 form the final ensemble.

3.1 Stage 1: Model Exploration and Selection

We explored four base-sized transformers, all trained with focal loss (Lin et al., 2017) (γ=2.0, α=None). Table 2 summarizes results. We select M1 (DeBERTa-v3-base (He et al., 2023), 184M, highest F1 after augmented training) and M2 (XLM-RoBERTa (Conneau et al., 2020), 278M, complementary multilingual coverage) for the ensemble. M3 (BERT-base (Devlin et al., 2019), 110M) served as a development baseline. M4 (toxic-bert (Hanu and Unitary team, 2020), 110M, frozen backbone + MLP) achieved only 0.3154 Macro F1 despite toxicity-specific pre-training, demonstrating that general-domain toxicity representations do not transfer to gaming contexts, where violent vocabulary is routinely non-toxic and multilingual slang is pervasive. This is part of what the LIM's domain-specific linguistic rules are designed to address.

3.2 Stage 2: Agreement-Based Fusion

When M1 and M2 agree (∼94%), we accept the consensus. For the ∼6% disagreements, we adopt the prediction with higher softmax probability as the initial estimate and route to the LIM.

3.3 Stage 3: Linguistically-Informed Mediator

The LIM refines disagreement predictions through four sequential components. It combines neural and symbolic processing: the ensemble captures distributional semantics, while the LIM encodes domain-specific linguistic facts that neural models cannot reliably learn from limited minority-class data. Every LIM decision traces back to a specific rule and corpus statistic, making it auditable.

3.3.1 Corpus-Backed Lexical Normalization

We normalize test messages (lowercase, strip punctuation, collapse expressive lengthening (Brody and Diakopoulos, 2011): "hahaha"→"haha") and perform exact-match lookup against train+val. Matches with $\geq$ 2 occurrences and $\geq$ 60% majority agreement adopt the majority label. These conservative thresholds directly reflect the annotation noise quantified in §D.

3.3.2 Class-Conditional Unigram Scoring

Inspired by the token-level analysis of Naseem et al. (2025), we compute $P(c \mid w) = \frac{n(w,c)}{n(w)}$ for each word w in the training vocabulary, where c is a class label, $n(w, c)$ is the number of messages containing w in class c, and $n(w)$ is the total count, effectively a unigram Naïve Bayes estimate. Words exceeding a precision threshold $P(c \mid w) \geq 0.80$ with sufficient support ($n(w) \geq 5$) serve as high-confidence minority-class indicators (Wiegand et al., 2018). For instance, identity-based slurs consistently map to H&H (P=1.00), while "kys" maps to Threats and leet-speak variants like "naz1" to Extremism. Overrides apply only toward safety-critical classes (3–5: H&H, Threats, Extremism), prioritizing precision to avoid false escalation.

3.3.3 Multilingual Profanity Detection

While M2 (XLM-RoBERTa) handles multilingual tokenization, our validation analysis revealed that *both* M1 and M2 still misclassify domain-specific non-Latin profanity, particularly terms rare even in XLM-R's 100-language pre-training. We applied the same statistics to non-Latin tokens, flagging words where $P(\text{toxic} \mid w) = 1 - P(\text{Non-toxic} \mid w) \geq 0.80$ but both models predicted Non-toxic. This yielded an empirically-validated multilingual lexicon organized by language family: East Slavic (Russian, Ukrainian), West Slavic (Polish, Czech), and other (Turkish, Hungarian, German). Reclas-

sification follows targeting: player-directed → Insults; game-directed → Other Offensive.

3.3.4 Agentive Targeting and Pragmatic Refinement

Drawing on speech act theory (Austin, 1962; Searle, 1969), we formalize the targeting function $\tau(m)$. Let T denote the set of tokens flagged as toxic by the preceding LIM components (unigram scoring and multilingual detection). For a message m containing a toxic token $t \in T$:

$$\tau(m) = \begin{cases} \textsc{Other-dir} & \text{if } \exists\, p \in P_2 : p \prec t \\ \textsc{Self-dir} & \text{if } \exists\, p \in P_1 : p \prec t \\ \textsc{Entity-dir} & \text{if } \exists\, e \in E : e \prec t \\ \textsc{Untargeted} & \text{otherwise} \end{cases} \tag{1}$$

where P_2/P_1 are second/first-person pronoun sets (English and Russian), E is a Game-Specific Entity (GSE) lexicon covering vehicles (400+ tanks), mechanics (*rng*, *arty*, *cap*), map locations (*Himmelsdorf*, *hill*, *banana*), and game roles (*light*, *heavy*, *TD*), and $\prec$ denotes linear precedence in the message. Table 3 maps targeting types to labels.

$\tau(m)$	Signal	$\hat{y}$
Other-dir	$P_2 + T$ (*you* + insult)	Insults
Self-dir	$P_1 + T$ (*I* + insult)	Non-toxic
Entity-dir	$E + T$ (*arty* + profanity)	Other Off.
Untargeted	T only	Non-toxic

Table 3: Targeting function $\tau(m)$. E = GSE lexicon. Based on speech act theory (Searle, 1969).

This component also applies censored-text recovery ([GSE] + [***] → Non-toxic) and implicit word sense disambiguation: "kill that Tiger" (GSE → Non-toxic) vs. "kill yourself" (person → Threats) (Firth, 1957).

4 Results and Discussion

Table 4 shows incremental results on the official test set. The ensemble improves over the best single model through complementary coverage, and the LIM further refines the ∼6% disagreements, with the largest contribution from unigram scoring (Table 4). The LIM's impact is concentrated in safety-critical minority classes, where high-precision corrections ensure that identity-based hate, threats, and extremist content are not missed by the neural ensemble.

Our system ranks 3rd in Macro F1 (0.6441) but achieves the **highest accuracy** (0.9062) among all participating teams, indicating the fewest total errors; the F1 gap to the top-ranked systems (0.7041, 0.6725) is concentrated in minority class recall. **Annotation noise** accounts for part of this ceiling: 340 unique messages carry conflicting labels across 7,416 training samples (17.3%), with the Non-toxic ↔ Insults boundary alone responsible for 6,455 conflicting samples, reflecting the same ambiguity between a playful insult and a genuine attack that makes this boundary the hardest to learn. **Multilingual blind spots** persist even with XLM-RoBERTa: domain-specific Cyrillic profanity is concentrated at 22.6% of H&H messages (3.5× the dataset average), reflecting the prevalence of Russian and Ukrainian identity-based slurs that fall outside standard multilingual pre-training corpora. **Domain gap**: the toxic-bert result (F1 = 0.3154 vs. 0.5439 for vanilla BERT) shows that Twitter/Reddit toxicity pre-training actively hurts gaming performance by associating GSE terms with toxicity.

System	F1	Acc	Prec	Rec
Best single (M1)	.5742	.8979	.6042	.5538
M1+M2 ensemble	.6032	.9059	.5713	.6579
+ Lex. norm.	.6107	.9057	.5782	.6591
+ Unigram scoring	.6221	.9044	.5964	.6559
+ Multilingual	.6256	.9047	.5970	.6626
+ Targeting & ref.	**.6441**	**.9062**	**.6334**	**.6601**

Table 4: Incremental ablation. Top: models. Bottom: LIM components applied to ∼6% disagreements. Full LIM includes targeting, censored-text recovery, and boundary enforcement.

We presented **thaulab**'s system for the EEUCA 2026 GameTox Shared Task, achieving Macro F1 of 0.6441 (3rd) and the highest accuracy (0.9062) using exclusively base-sized models and no external data. The pipeline is adaptive by design: augmentation targets the boundaries the model struggles with, and the LIM concentrates corrections on Extremism, Threats, and H&H, categories where misclassification causes real harm beyond any leaderboard metric. The three-stage framework generalizes to other gaming platforms with only GSE lexicon substitution. Key findings: (1) symbolic mediation on ensemble disagreements improves safety-critical minority-class detection; (2) multilingual transformers retain blind spots on domain-specific profanity; (3) agentive targeting distinguishes toxic intent from benign game communication; (4) general toxicity models fail in gaming contexts.

Limitations

The LIM relies on simple symbolic methods (lexicon lookup, unigram statistics, pronoun-based targeting) rather than learned components. While this trades off flexibility for explainability, since every correction traces back to a specific corpus statistic or linguistic pattern, it likely leaves performance on the table. We set strict thresholds throughout (e.g., $P(c \mid w) \geq 0.80$, majority agreement $\geq 60\%$) and have not yet explored relaxing them per class or per confidence region. The four LIM components are applied in a fixed sequential order without searching over alternative orderings, and the LIM currently operates only on the $\sim$6% of samples where M1 and M2 disagree; extending symbolic mediation to the full prediction set could improve coverage but risks false positives at the noisy majority-class boundaries. On the augmentation side, all synthetic data was generated using Claude Opus 4.6, a safety-aligned model that may undergenerate realistic toxic content due to its content filtering; less restricted models (e.g., Grok) could potentially yield more naturalistic minority-class samples, particularly for Extremism where only 24 original samples exist. Finally, the multilingual lexicon is manually curated and incomplete for unrepresented languages, the GSE lexicon is specific to World of Tanks, targeting analysis cannot resolve implicit targeting or zero pronouns in pro-drop languages, and we could not experiment with larger backbone models due to time and compute constraints.

Ethics Statement

This work classifies toxic language including slurs and profanity in multiple languages. Examples are reported solely for scientific reproducibility. Our multilingual profanity lexicon is compiled for research purposes only.

Acknowledgements

This work is supported by the project BRAINS (PID2024-155948OB-C52), funded by MCIN/AEI/10.13039/501100011033 and, as appropriate, by "ERDF A way of making Europe" and by the "European Union"; and by the Program *Actividades I+D Procesos Humanos y Sociales* of the Comunidad de Madrid, Spain, under Project PHS2024/PH-HUM-52 (Innovatrad-CM).

References

Baktash Ansari, Elias Martin, and Afra Mashhadi. 2026. ToxiTwitch: Toward emote-aware hybrid moderation for live streaming platforms. *Preprint*, arXiv:2601.15605.

John L. Austin. 1962. *How to Do Things with Words*. Oxford University Press.

Abhijeet Awasthi, Sabyasachi Ghosh, Rasna Goyal, and Sunita Sarawagi. 2020. Learning from rules generalizing labeled exemplars. In *International Conference on Learning Representations*.

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.

Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: Using word lengthening to detect sentiment in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 562–570.

Christopher Clarke, Matthew Hall, Gaurav Mittal, Ye Yu, Sandra Sajeev, Jason Mars, and Mei Chen. 2023. Rule by example: Harnessing logical rules for explainable hate speech detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 364–376.

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. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Somaiyeh Dehghan, Mehmet Umut Sen, and Berrin Yanikoglu. 2025. Dealing with annotator disagreement in hate speech classification. *Preprint*, arXiv:2502.08266.

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.

John R. Firth. 1957. A synopsis of linguistic theory, 1930–1955. *Studies in Linguistic Analysis*, pages 1–32.

Laura Hanu and Unitary team. 2020. Detoxify. https://huggingface.co/unitary/toxic-bert.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Yunhao Hu, Sophie Evelyn, and Elizabeth M. Clancy. 2025. Player versus player: A systematic review of cyberbullying in multiplayer online games. *Computers in Human Behavior*.

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)*.

Peiran Li, Jan Fillies, and Adrian Paschke. 2026. ToxiGAN: Toxic data augmentation via LLM-guided directional adversarial generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

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.

Jacob Morrier, Amine Mahmassani, and R. Michael Alvarez. 2024. Uncovering the viral nature of toxicity in competitive online video games. *Preprint*, arXiv:2410.00978.

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.

Adrien Schurger-Foy, Rafal Dariusz Kocielnik, Caglar Gulcehre, and R. Michael Alvarez. 2025. Context-aware toxicity detection in multiplayer games: Integrating domain-adaptive pretraining and match metadata. *Preprint*, arXiv:2504.01534.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293.

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)*.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words–a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056.

Yaqi Zhang, Viktor Hangya, and Alexander Fraser. 2024. A study of the class imbalance problem in abusive language detection. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*. Association for Computational Linguistics.

A Hyperparameter Configuration

Table 5 lists the model-specific training configuration for all four architectures. The pre-trained checkpoints are `microsoft/deberta-v3-base` (M0/M1), `xlm-roberta-base` (M2), `bert-base-uncased` (M3), and `unitary/toxic-bert` (M4). All models share the same base settings: focal loss with focusing parameter $\gamma = 2.0$ and no class balancing ($\alpha =$ None), maximum sequence length of 64 tokens, batch size 32, 5 training epochs, AdamW optimizer with weight decay 0.01, and random seed 42.

While M2 and M3 natively instantiate in single-precision (`float32`), the DeBERTa-v3 checkpoints (M0/M1) are natively stored in half-precision (`float16`). We observed that fine-tuning DeBERTa-v3 in `float16` resulted in catastrophic gradient collapse (`NaN` loss) during the initial training steps, a known instability caused by arithmetic overflow within DeBERTa's Disentangled Attention matrices, where intermediate activation values exceed the `float16` maximum representable limit. To resolve this, we explicitly upcast the DeBERTa weights to `float32` during initialization, providing sufficient numerical stability for the attention mechanism to converge. We additionally tested $\gamma = 2.5$, dynamic α (inverse class frequency), and a two-stage hierarchical approach (binary toxic/non-toxic classification followed by fine-grained 6-class prediction within the toxic branch). All alternatives yielded marginal differences (Δ Macro F1 < 0.005), so we standardized the simplest configuration for reproducibility across all architectures.

B LIM Component Details

Table 6 lists all LIM thresholds, selected on the validation set and held fixed during test evaluation. The lexical normalization majority threshold was set at 60% rather than 50% because lower values introduced false corrections at the noisy Insults $\leftrightarrow$ Other Offensive boundary. The unigram precision cutoff of $P \geq 0.80$ was chosen because at $P \geq 0.70$, ambiguous terms (e.g., "monkey" at $P(\text{H\&H})=0.75$) triggered false positives; raising to 0.80 retains only unambiguous high-precision tokens. These thresholds are intentionally strict for the competition setting and can be relaxed for higher-recall deployment.

Throughout the LIM, we enforce annotation-guideline boundaries (Naseem et al., 2025): identity-based slurs $\rightarrow$ H&H; 2nd person + non-identity insult $\rightarrow$ Insults; profanity without personal targeting $\rightarrow$ Other Offensive; game callouts and GSE terms $\rightarrow$ Non-toxic; directed violence + personal target $\rightarrow$ Threats; political ideology and recruitment $\rightarrow$ Extremism.

C Augmentation Pipeline

Toxic vocabulary mining. Class-conditional unigram probabilities $P(c \mid w)$ are computed as described in §3.3.2. For augmentation, we additionally flag *class-discriminative tokens* using the frequency ratio $\frac{n(w,c)/N_c}{n(w)/N} \geq 5$, where N_c and N are class and corpus sizes respectively; this yields a focused toxic vocabulary substantially smaller than the full $\sim$30K vocabulary, defining the TF-IDF subspace for similarity gating below.

Before computing any statistics, all text undergoes leet-speak normalization to unmask common obfuscation patterns prevalent in gaming chat. The character substitution mappings are: `0`$\rightarrow$`o`, `1`$\rightarrow$`i`, `3`$\rightarrow$`e`, `4`$\rightarrow$`a`, `5`$\rightarrow$`s`, `7`$\rightarrow$`t`, `@`$\rightarrow$`a`, `$`$\rightarrow$`s`. This normalization is applied consistently in both the augmentation pipeline (for seed term selection and similarity verification) and the LIM (for unigram scoring and multilingual detection at inference time).

Cosine similarity gating in the toxic subspace. To verify that generated samples are linguistically consistent with real training data, we project both real and synthetic utterances into a *toxic-only TF-IDF subspace*. Rather than computing TF-IDF vectors over the full $\sim$30K vocabulary (which produces extremely sparse, high-dimensional vectors for short gaming messages of 2–4 tokens), we restrict the vocabulary to only the mined class-discriminative terms. This projection substantially reduces dimensionality and eliminates the sparsity problem inherent in full-vocabulary TF-IDF for short texts. Cosine similarity between each generated sample and its nearest real training neighbor in this subspace serves as a geometric filter: samples that fall below a minimum similarity threshold are rejected as out-of-distribution, while samples above a maximum threshold are rejected as near-duplicates of existing training data. This dual-threshold approach ensures that generated samples are close enough to the training distribution to be realistic, yet sufficiently novel to provide genuine augmentation value.

Parameter	M0/M1 (DeBERTa-v3)	M2 (XLM-RoBERTa)	M3 (BERT)	M4 (toxic-bert)
Total Parameters	184M	278M	110M	110M
Trainable Parameters	184M (full)	278M (full)	110M (full)	∼200K (MLP only)
Training Paradigm	Full fine-tune	Full fine-tune	Full fine-tune	Frozen backbone + MLP
Learning Rate	1×10^{-5}	2×10^{-5}	2×10^{-5}	1×10^{-3} (MLP)
Warmup Steps	10% of total	0	0	0
Weight Initialization	`float32` (upcast)	`float32` (native)	`float32` (native)	`float32` (native)

Table 5: Model-specific training hyperparameters. M0 and M1 share the identical configuration; M0 is trained on the original training set, M1 on the augmented set.

Component	Parameter	Value
Lex. Norm.	Min occurrences	≥ 2
	Majority threshold	$\geq 60\%$
	Normalization	Lower, strip, dedup (≥ 3)
Unigram	$P(c \mid w)$ cutoff	≥ 0.80
	Min support $n(w)$	≥ 5
	Direction	Minority ↑ only
	Freq. ratio	$\geq 5\times$ baseline
Multilingual	Languages	10+
	Mining criterion	$P(\text{toxic} \mid w) \geq 0.80$ + both models wrong
	Reclassification	Targeting-sensitive
Targeting	Pronoun sets	EN, RU
	GSE lexicon E	Vehicles (400+), mechanics, maps, roles
	Censored pattern	[GSE] + [***]

Table 6: LIM thresholds. All selected on the validation set.

Generation API configuration. All synthetic samples were generated using the Claude Opus 4.6 API (`claude-opus-4-6-20250514`). We used a temperature of `1.0` to encourage lexical diversity across generated samples, `max_tokens = 2048`, and `top_p = 1.0` (no nucleus truncation). No additional system-level parameters were set beyond the defaults; the full generation behavior is governed solely by the prompt templates below. These settings are fixed across both Template A and Template B calls to ensure reproducibility.

Template A: Confusion-pair-driven generation. For classes identified through the seed model's (M0) prediction uncertainty, we provide the language model with the target class definition, representative seed examples from the training set, and the specific confused class pair that the model struggles with:

```
You are a data augmentation assistant
for a toxicity classification dataset
derived from World of Tanks in-game
chat. Your task is to generate realistic
synthetic chat messages for a specific
toxicity class.

Target class: {CLASS_NAME}
Class definition: {CLASS_DEFINITION}

The class definitions follow the
annotation guidelines from the GameTox
dataset (Naseem et al., 2025):
- Hate and Harassment: Identity-based
hate or harassment (racism, sexism,
homophobia)
- Threats: Threats of violence, physical
safety, terrorism, or doxxing
- Extremism: Extremist views,
grooming/recruitment for extremist
groups
- Insults and Flaming: Insults or
attacks not based on identity
- Other Offensive: Offensive content
not covered by the above categories
- Non-toxic: Neutral game communication

Seed examples from the training data:
{SEED_EXAMPLES}

Confused with: {CONFUSED_CLASS}
(our classifier frequently confuses
{CLASS_NAME} with {CONFUSED_CLASS})

Requirements:
1. Generate exactly 20 new chat messages
that CLEARLY belong to {CLASS_NAME} and
NOT to {CONFUSED_CLASS}.
2. Each message should be 1-8 words long
(typical length in game chat).
3. Include common gaming abbreviations,
slang, and informal spelling.
4. Include multilingual variants where
appropriate (Russian, Polish, Turkish,
German).
5. Each message must be unambiguously
classifiable by a human annotator
following the guidelines above.
6. Do NOT repeat or closely paraphrase
the seed examples.
7. Output one message per line with no
numbering or formatting.
```

Template B: Contrastive boundary augmentation. For extreme minority classes (Extremism with only $n = 24$ training samples, and supplemental Threats with $n = 60$) that are too rare to appear reliably in confusion-pair analysis, we provide discriminative keywords mined from the training set along with explicit instructions to generate boundary-proximal samples:

```
You are a data augmentation assistant for
```

```
a toxicity classification dataset from
World of Tanks in-game chat.

Target class: {CLASS_NAME}
Class definition: {CLASS_DEFINITION}
Adjacent (easily confused) class:
{ADJACENT_CLASS}
Adjacent class definition:
{ADJACENT_DEFINITION}

Discriminative keywords for {CLASS_NAME}
(statistically mined from training data,
P(class|word) >= 0.80):
{HIGH_P_KEYWORDS}

Existing training examples of
{CLASS_NAME}:
{SEED_EXAMPLES}

Requirements:
1. Generate exactly 20 new messages
that belong to {CLASS_NAME}.
2. CRITICAL: Messages must be
CLOSE to the decision boundary with
{ADJACENT_CLASS}. They should be
challenging to classify, but still
clearly {CLASS_NAME} according to the
annotation guidelines.
3. Include cross-lingual variants
(Russian, Polish, Turkish, German).
4. Vary message length (1-8 words).
5. Each message should be
distinguishable from {ADJACENT_CLASS}
ONLY by the specific class-defining
linguistic feature (e.g., identity
targeting for H&H vs. skill targeting
for Insults, or political ideology for
Extremism vs. identity hate for H&H).
6. Do NOT repeat seed examples.
7. Output one message per line with no
numbering.
```

Generation results and quality control. Template A yielded 34 Other Offensive, 155 Hate & Harassment, and a portion of the Threats samples. Template B yielded all 207 Extremism samples and supplemental Threats samples, for a combined total of 631 synthetic samples. All generated samples underwent three quality control steps: (1) cosine similarity gating in the toxic TF-IDF subspace to reject out-of-distribution and near-duplicate generations; (2) exact and near-duplicate removal against the original training set to prevent data leakage; (3) manual spot-checking of a random 10% subset for label consistency with the annotation guidelines. Class definitions in both templates were drawn directly from the annotation guidelines of Naseem et al. (2025).

D Annotation Noise Analysis

A key challenge in the GameTox dataset is annotation inconsistency at class boundaries. We identify 340 unique normalized messages that appear with conflicting labels across their multiple occurrences in the training set, collectively affecting 7,416 individual training samples (17.3% of the dataset). This inconsistency arises because identical text appears in different game sessions and receives different annotations each time. For instance, a player typing "wtf" in one match may be reacting to an unfair death (Other Offensive), while in another match the same message is interpreted as a neutral exclamation (Non-toxic). This is not a failure of multiple annotators disagreeing on a single instance; rather, it reflects the genuine context-dependence of short gaming messages.

Table 7 shows representative examples. The column n indicates the total number of times that normalized message appears in the training set across all game sessions. The label distribution shows the percentage of those n occurrences assigned to each class.

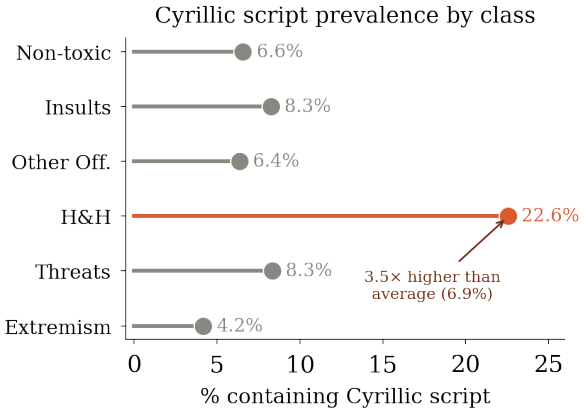


Figure 3: Cyrillic-script prevalence by toxicity class. Hate & Harassment contains 3.5× more Cyrillic content than the dataset average (6.9%), indicating that non-Latin-script profanity is structurally concentrated in the most severe toxicity category.

Message	n	Label Distribution
gg	2,703	NT: 99.9%, Ins: 0.1%
wtf	208	OO: 77.4%, NT: 22.1%, Ins: 0.5%
cap	192	NT: 96.4%, OO: 2.1%, Ins: 1.6%
arty	189	NT: 97.4%, Ins: 2.6%
idiot	101	Ins: 94.1%, NT: 5.0%, OO: 1.0%
ffs	55	OO: 80.0%, NT: 16.4%, Ins: 3.6%

Table 7: Annotation noise examples. n = total occurrences in training data. The same normalized text receives different labels across game sessions. NT = Non-toxic, Ins = Insults, OO = Other Offensive.

Figure 4 visualizes the magnitude of annotation conflicts across all class pairs. Each bubble represents a pair of classes; the bubble size and color intensity are proportional to the num-

ber of training samples where the same message receives labels from both classes. The Non-toxic ↔ Insults boundary dominates with 6,455 conflicting samples, reflecting the fundamental ambiguity between a playful insult and a genuine attack in gaming chat. The Non-toxic ↔ Other Offensive boundary (1,528 samples) and the Insults ↔ Other Offensive boundary (958 samples) are the next most noisy. Notably, minority class boundaries (involving H&H, Threats, or Extremism) exhibit minimal noise, because the linguistic signals for these classes (identity-based slurs, directed violence, political ideology) are more distinctive and less context-dependent.

This noise pattern directly informs the LIM design: we use conservative thresholds ($\geq 60\%$ majority agreement) for the noisy majority-class boundaries, while applying more aggressive corrections for minority classes where annotation agreement is near-unanimous.

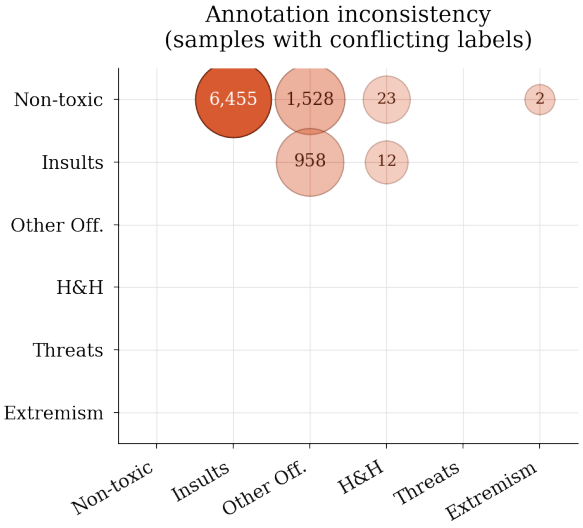


Figure 4: Annotation inconsistency across all class pairs. Bubble size reflects the total number of training samples where identical normalized messages receive conflicting labels from the two classes. The Non-toxic ↔ Insults boundary dominates at 6,455 conflicting samples, illustrating the context-dependent nature of short gaming messages.

E Multilingual Content Analysis

Figure 3 shows the proportion of Cyrillic-script messages by toxicity class. The 22.6% concentration in Hate & Harassment (compared to a 6.9% dataset average, a 3.5× difference) reflects the prevalence of Russian and Ukrainian identity-based profanity systems, collectively known as *mat*, which include some of the strongest and most targeted slurs in the Slavic language family. Beyond Cyrillic, the dataset contains content in Polish and Czech (0.30% of training data), Turkish (0.31%), and Hungarian (0.30%). In the test set, 389 messages (7.2%) contain Cyrillic script, 22 contain Turkish characters, and 17 contain Polish/Czech characters.

Why XLM-RoBERTa is insufficient. A natural question is why the LIM's multilingual lexicon is needed given that M2 (XLM-RoBERTa) is pre-trained on 100 languages including Russian, Ukrainian, Polish, and Turkish. The answer lies in the distinction between *general-vocabulary* multilingual competence and *domain-specific* profanity detection. XLM-RoBERTa's pre-training corpus (CommonCrawl) contains formal and semi-formal text, but underrepresents the specific register of gaming chat profanity: context-dependent slurs that are used as identity-based attacks in one context and as general frustration in another, obfuscated forms of profanity, and compound insults that combine multiple languages within a single utterance. Our validation analysis confirmed this empirically: we identified tokens where $P(\text{toxic} \mid w) = 1 - P(\text{Non-toxic} \mid w) \geq 0.80$ in the training data, yet *both* M1 (DeBERTa) and M2 (XLM-RoBERTa) predicted Non-toxic on the validation set. The LIM's multilingual lexicon targets precisely these residual blind spots, not as a replacement for XLM-RoBERTa's multilingual capacity, but as a domain-specific complement to it.

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.

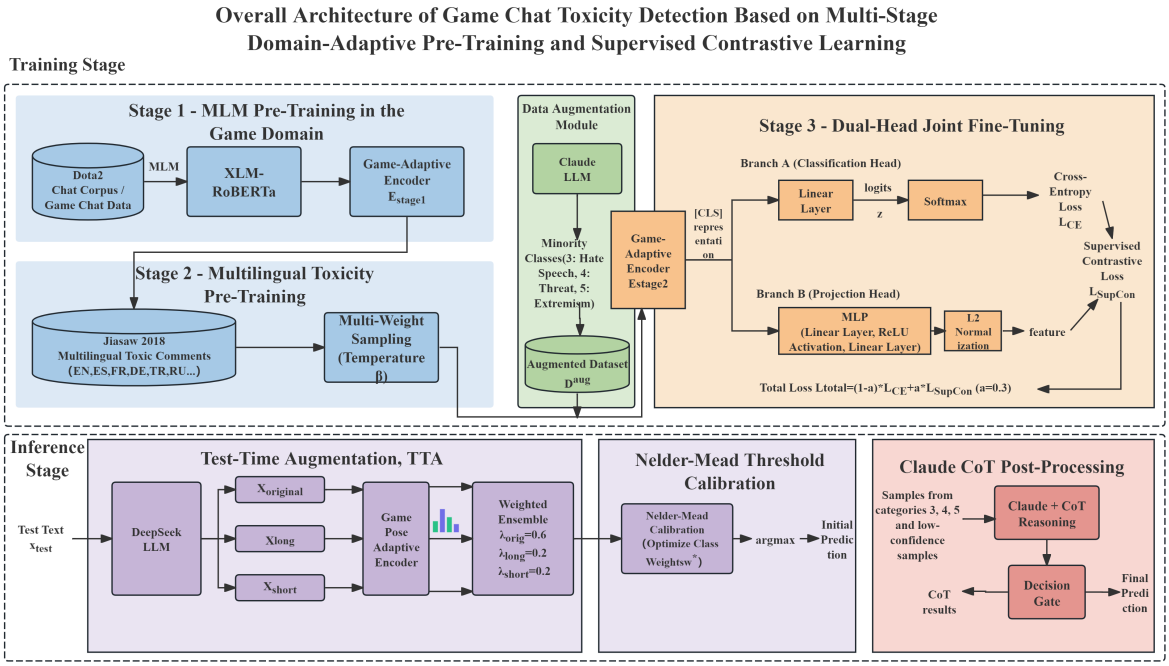


Figure 1: Overall architecture of our three-stage progressive training framework. The pipeline covers domain adaptive pre-training (Stage 1), multilingual toxicity transfer fine-tuning (Stage 2), SCL-enhanced end-to-end fine-tuning (Stage 3), and targeted optimization strategies.

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:
```

CSECU-Learners@EEUCA 2026: Vaccine Critical Memes Identification using Two-Stage Early Fusion of Transformers

Monir Ahmad
Department of Computer Science and Engineering
University of Chittagong,
Chattogram-4331, Bangladesh
ahmad.csecu@gmail.com

Md. Saif Uddin
Department of Computer Science and Mathematics
Bangladesh Agricultural University,
Mymensingh-2202, Bangladesh
saifuddin.csm@bau.edu.bd

Abstract

Memes have emerged as a fast and influential way to share information online, particularly during major public health events like COVID-19 vaccination. While they can support awareness and encourage positive behavior, they are also widely used to spread misinformation and vaccine-critical views. These messages are often expressed through sarcasm and implicit meaning, which makes automatic detection difficult. To tackle this problem, EEUCA 2026 introduces a shared task based on the VaxMeme dataset for multimodal vaccine critical meme detection. The task encourages us to design models that can jointly understand both image and text, capturing the underlying context more effectively. In this work, we present our approach to this task by proposing a two-stage early fusion framework that integrates multiple transformer-based encoders. We train our model using focal loss to give more attention to difficult samples. Our experimental results show that our method performs competitively in the shared task, demonstrating its effectiveness for this problem.

1 Introduction

The rapid growth of social media has transformed how information is created, shared, and consumed, with memes emerging as one of the most influential forms of communication. Memes, which typically combine images and short textual elements, are highly engaging due to their humorous, sarcastic, and easily shareable nature (Pramanick et al., 2021a).

However, this same virality makes them a powerful vehicle for spreading misleading or harmful narratives (Wang et al., 2020). In the context of public health, vaccine-critical memes have become particularly concerning, as they can promote misinformation, reinforce vaccine hesitancy, and negatively influence public perception toward immunization efforts. Given the demonstrated relationship between online exposure and real-world attitudes, the automatic detection of such content is crucial for enabling timely interventions and supporting public awareness campaigns (Wang et al., 2020). The organizer of EEUCA 2026 (Hürriyetoğlu et al., 2026) proposes a shared task to support multimodal vaccine-critical meme detection (Thapa et al., 2026b). The task allows participants to develop models that jointly leverage both visual and textual representations to capture the global and local contextual cues embedded in memes.

Early research in this domain has predominantly focused on text-based analysis of social media content (Zhang et al., 2020; Naseem et al., 2021). These approaches leverage traditional machine learning models and, more recently, transformer-based architectures such as BERT (Devlin et al., 2019) to classify vaccine-related opinions and sentiments. While these methods have shown promising results, they are inherently limited in their ability to capture the full meaning of memes. To address these shortcomings, recent studies have explored multimodal approaches that jointly analyze textual and visual information (Pramanick et al., 2021b; Volkova et al., 2019). These methods have demonstrated improved performance across various tasks, including fake news detection, hateful meme identification, and misinformation analysis. Many models focus primarily on either global or local feature representations, without effectively combining both. Another critical limitation is the scarcity of publicly available, well-annotated multimodal datasets for vaccine critical content detection.

Our submitted system, CSECU-Learners addresses this challenge through a multi-encoder two-stage early fusion architecture. Specifically, we employ a Twitter-domain RoBERTa (Barbieri et al., 2020) to encode the textual content of each post, a Vision Transformer (ViT) (Dosovitskiy et al., 2020) to capture visual features from the

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 169–176

meme image, and a Vision-and-Language Transformer (ViLT) (Kim et al., 2021) to obtain joint cross-modal representations. In Stage 1, the pooler outputs of RoBERTa and ViT are combined via performance-weighted summation. In Stage 2, this visual contextualized representation is concatenated with the ViLT pooler output that is passed to a linear classification layer. To mitigate the effect of class imbalance in the training corpus, we adopt Focal Loss (Lin et al., 2017).

We structure the remainder of this paper as follows. Section 2 introduces the proposed approach for the EEUCA-2026 multimodal vaccine-critical meme detection task. Section 3 outlines the experimental design, including implementation details and parameter settings. The results and their analysis are presented in Section 4. Finally, we conclude the paper by highlighting future research directions in Section 5, followed by a discussion of the limitations of the proposed method.

2 System Overview

This section provides an overview of our proposed system for the shared task on multimodal identification of vaccine critical content on social media at EEUCA 2026. Figure 1 presents a high-level illustration of our proposed system.

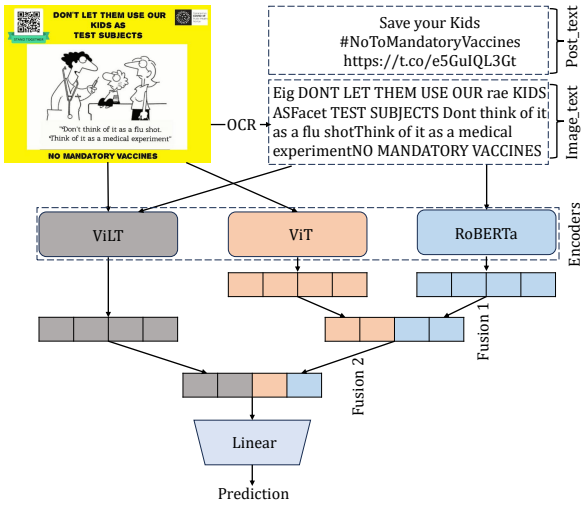


Figure 1: Overview diagram of our proposed system for EEUCA 2026: the shared task on multimodal identification of vaccine critical content on social media.

A social media post typically consists of two distinct information sources: the accompanying caption (post text) and the text embedded within the image itself (image text), the latter of which is recovered via Optical Character Recognition (OCR). We design a multi-encoder fusion architecture that jointly processes the visual and textual signals extracted from each post.

At the encoder stage, three pre-trained transformer-based models operate in parallel. Vision-and-Language Transformer (ViLT) (Kim et al., 2021) processes the raw image alongside its associated textual content, leveraging its native cross-modal attention to learn joint image–text representations. Vision Transformer (ViT) (Dosovitskiy et al., 2020), by contrast, focuses exclusively on the visual content of the post, while the Robustly Optimized BERT Pretraining Approach (RoBERTa) (Liu et al., 2019) handles the purely linguistic dimensions of the task.

We subsequently consolidate the representations produced by these three encoders through a two-stage fusion strategy. In the first fusion stage, the contextualized pooler output from ViT and the contextualized pooler output from RoBERTa are fused to form a combined multimodal representation. In the second fusion stage, this combined representation is merged with the holistic vision–language embedding produced by ViLT, yielding a unified feature vector that integrates all three perspectives. This final fused representation is passed through a linear classification layer, which produces logits (unnormalized scores) from which our model derives its final prediction.

2.1 Encoder Models

We fine-tune ViLT to obtain contextualized multimodal representations, employ ViT to capture visual information from the given image, and utilize RoBERTa to extract contextualized textual feature representations.

2.1.1 RoBERTa

For text encoding, we employ a RoBERTa-base model (Liu et al., 2019) fine-tuned on Twitter data[1]. It was originally introduced by Barbieri et al. (Barbieri et al., 2020) as part of the TweetEval benchmark. Unlike general-domain language models, this checkpoint was trained on roughly 58 million tweets, making it particularly sensitive to the informal grammar, hashtags, URLs, and emotionally charged phrasing that characterize vaccine-related discourse on social media. Given an input sequence, the model produces a contextualized

[1] https://huggingface.co/cardiffnlp/twitter-roberta-base

[CLS] pooler output, which serves as the textual representation fed into the fusion stage.

2.1.2 ViT

For visual encoding, we adopt the Vision Transformer (ViT) (Dosovitskiy et al., 2020) in its base configuration with 32×32 patch size, pre-trained on ImageNet-21k[2]. ViT treats an input image as a sequence of fixed-size non-overlapping patches, projects each patch into a linear embedding, and processes the resulting sequence through a standard transformer encoder. The larger patch size reduces sequence length and computational cost while retaining sufficient visual detail for meme-style social media images. It typically carries meaning through coarse layout and salient objects rather than fine-grained texture. The [CLS] pooler output of ViT is used as the visual representation at the first fusion stage.

2.1.3 ViLT

To capture cross-modal interactions directly at the encoder level, we incorporate the Vision-and-Language Transformer (ViLT) (Kim et al., 2021). We specifically utilize the base model pre-trained with masked language modeling on image–text pairs[3]. Unlike pipeline approaches that extract visual features with a separate object detector before cross-modal fusion, ViLT encodes image patches and text tokens within a single unified transformer. This approach enables direct attention between visual and linguistic elements at every layer. The model receives the raw post image together with the text as inputs, and its pooler output provides a holistic vision–language representation that complements the independently encoded visual and textual streams at the second fusion stage.

2.2 Two-Stage Early Fusion

Rather than relying on a single encoder to capture all modality-specific signals, we propose a two-stage early fusion strategy that progressively consolidates the representations from RoBERTa, ViT, and ViLT into a unified feature vector for downstream classification.

2.2.1 Stage 1: Weighted Fusion of Visual and Textual Representations

In the first stage, we combine the pooler outputs of RoBERTa and ViT through a performance-aware weighted summation followed by a $\tanh$ activation. Let $\mathbf{p}_R \in \mathbb{R}^{768}$ and $\mathbf{p}_V \in \mathbb{R}^{768}$ denote the pooler outputs of RoBERTa and ViT, respectively.

To determine the fusion weights, we rank the two encoders by their individual performance on the validation set, assigning an order number k to each model such that the better-performing model receives $k = 1$ (Du et al., 2022). Since RoBERTa outperforms ViT on the validation set, RoBERTa is assigned $k_R = 1$ and ViT is assigned $k_V = 2$. The weight for each encoder is then computed as:

$$w_k = \frac{1}{\sqrt{k}}, \quad k \in \{1, 2\} \tag{1}$$

This yields $w_R = 1/\sqrt{1} = 1.000$ for RoBERTa and $w_V = 1/\sqrt{2} \approx 0.707$ for ViT. The weighted sum is then computed as:

$$\mathbf{f}_1 = w_R \, \mathbf{p}_R + w_V \, \mathbf{p}_V \tag{2}$$

where $\mathbf{f}_1 \in \mathbb{R}^{768}$ is the resulting visual-contextualized representation that jointly encodes textual semantics and visual content while preserving the relative contribution of each encoder according to its predictive capability.

2.2.2 Stage 2: Fusion with Cross-Modal ViLT Representation

In the second stage, we incorporate the joint image–text representation produced by ViLT. Let $\mathbf{p}_L \in \mathbb{R}^{768}$ denote the pooler output of ViLT. A key limitation of ViLT is that it accepts a maximum input sequence length of 40 tokens. However, vaccine-critical content on social media is often considerably longer. Specifically, in the VaxMeme training set, approximately 63% of non-empty image texts and 72% of non-empty post texts exceed this 40-token limit when tokenized using the `cardiffnlp/twitter-roberta-base` tokenizer. Consequently, ViLT alone cannot fully encode the linguistic content of such posts, whereas the Stage 1 fusion — which employs RoBERTa — does not suffer from this constraint.

To complement the full-sequence textual encoding from Stage 1 with the cross-modal attention capability of ViLT, we concatenate $\mathbf{f}_1$ and $\mathbf{p}_L$ as follows:

$$\mathbf{f}_2 = \text{Concat}(\mathbf{f}_1, \, \mathbf{p}_L) \tag{3}$$

where $\mathbf{f}_2 \in \mathbb{R}^{1536}$ is the final fused representation obtained by concatenating the 768-dimensional visual-contextualized output from Stage 1 with the

[2] https://huggingface.co/google/vit-base-patch32-224-in21k

[3] https://huggingface.co/dandelin/vilt-b32-mlm

768-dimensional ViLT pooler output. This unified representation is subsequently passed to the linear classification layer for prediction.

2.3 Classification

The unified representation $\mathbf{f}_2 \in \mathbb{R}^{1536}$ obtained from Stage 2 fusion is fed into a single linear feed-forward layer that maps the fused embedding to the output space. Formally, the unnormalized class scores (logits) are computed as:

$$\hat{\mathbf{y}} = \mathbf{f}_2 \, \mathbf{W}^\top + \mathbf{b} \tag{4}$$

where $\mathbf{W} \in \mathbb{R}^{n \times d}$ is the weight matrix and $\mathbf{b} \in \mathbb{R}^n$ is the bias vector of the linear layer, with $d = 1536$ being the dimensionality of the fused representation from Stage 2 and n denoting the number of target classes. The final predicted class label $\hat{y}$ is determined by selecting the class corresponding to the maximum logit.

2.4 Focal Loss

Training on real-world social media datasets often involves skewed class distributions, where certain categories are substantially under-represented relative to others. Standard cross-entropy loss (Zhang and Sabuncu, 2018) tends to be dominated by the more frequent, easily classified examples, which can impede the model from learning discriminative patterns for minority or harder instances. To address this, we adopt Focal Loss (Lin et al., 2017).

Let t denote the index of the ground-truth class for a given input sample, and let $\hat{\mathbf{y}} \in \mathbb{R}^n$ be the logit vector. The predicted probability for the true class p_t is obtained via the softmax function. Focal loss addresses limitation of cross-entropy loss by augmenting the cross-entropy term with a modulating factor $(1 - p_t)^\gamma$:

$$\mathcal{L}_{\text{FL}}(p_t) = -(1 - p_t)^\gamma \log(p_t) \tag{5}$$

where $\gamma \geq 0$ is the focusing parameter that governs the rate at which well-classified examples are down-weighted. When $p_t \to 1$, the modulating factor $(1 - p_t)^\gamma \to 0$, effectively reducing the loss contribution of confidently correct predictions. Conversely, when the model misclassifies a sample and p_t remains small, the factor approaches unity, preserving the full loss signal for that instance.

3 Experimental setup

3.1 Dataset Overview

To evaluate our proposed framework on the shared task on multimodal identification of vaccine critical content on social media at EEUCA 2026, we use the annotated benchmark dataset provided by the task organizers (Thapa et al., 2026a). The dataset builds upon the VaxMeme corpus (Naseem et al., 2023), with an annotation schema shared with CrisisHateMM (Bhandari et al., 2023). Each instance in the dataset includes three components: the image, the text extracted from the image using OCR, and the accompanying post caption. In our approach, we utilize the image and the post text, while excluding the OCR-extracted image text, as it is often empty for most samples and its incorporation with the post text leads to performance degradation on the validation set. Each sample is assigned one of three class labels: *Vaccine Critical*, *Neutral*, or *Pro-vaccine*. The distribution of samples across the training, validation, and test splits is summarized in Table 1.

Class Label	Train	Validation	Test
Vaccine Critical	2,535	308	314
Neutral	2,461	327	316
Pro-vaccine	3,199	389	395
Total	8,195	1,024	1,025

Table 1: Distribution of data samples across splits in the shared task dataset.

The training set comprises 8,195 samples in total, with *Pro-vaccine* being the most frequent class at 3,199 instances (39.03%), followed by *Vaccine Critical* at 2,535 instances (30.93%) and *Neutral* at 2,461 instances (30.03%). A similar trend is observed in the validation set of 1,024 samples, where *Pro-vaccine* again constitutes the largest proportion at 389 instances (37.99%), with *Neutral* and *Vaccine Critical* accounting for 31.93% (327 instances) and 30.08% (308 instances), respectively. The test set contains 1,025 samples.

Across both the training and validation splits, the dataset exhibits a moderate degree of class imbalance. The *Pro-vaccine* class consistently outnumbers the other two categories, with a ratio of approximately 1.30:1 and 1.26:1 relative to *Vaccine Critical* and *Neutral* in the training set, respectively. While this imbalance is not extreme, it is sufficient to bias a naively trained classifier towards the majority class, potentially at the expense of cor-

rectly identifying *Vaccine Critical* content. This motivates our adoption of Focal Loss (Section 2.4), which mitigates the adverse effect of class imbalance by down-weighting the loss contribution of over-represented, easily classified samples during training.

3.2 Parameter Settings

This section describes the experimental setup for our submission to the shared task at EEUCA 2026. We fine-tune the Twitter-RoBERTa, ViT, and ViLT models available through the Hugging Face Transformers library (Wolf et al., 2020) using a Kaggle notebook[4] equipped with an NVIDIA Tesla T4 GPU. To ensure reproducibility across experimental runs, the random seed is fixed at 66 throughout all experiments. We employ the AdamW optimizer (Loshchilov and Hutter, 2017) for parameter updates, and model selection is performed by saving the checkpoint that achieves the best macro-averaged F1 score on the validation set. The optimal hyperparameter values identified through our experiments are summarized in Table 2. The maximum input sequence length is set to 256 tokens for RoBERTa. The focusing parameter of Focal Loss is set to $\gamma = 1$, which provides a mild emphasis on harder, misclassified samples.

Hyperparameter	Optimal Value
Train batch size	8
Test batch size	8
Learning rate	3e-5
Random seed	66
Max sequence length	256
Dropout probability	0.3
Number of train epochs	3
Focusing parameter (γ)	1

Table 2: Optimal hyperparameter configuration used in our experiments.

3.3 Evaluation Measures

To assess the effectiveness of the systems proposed by participants, the organizers use the macro-averaged F1 score (Sokolova and Lapalme, 2009) as the primary evaluation metric. This evaluation metric is well-suited for datasets with long-tail class distributions, as it treats all classes equally. By taking the harmonic mean of precision and recall for each class and then averaging the results, it offers a balanced view of overall model performance.

[4] https://www.kaggle.com

This formulation ensures that minority classes such as *Vaccine Critical* and *Neutral* contribute equally to the overall score as the majority *Pro-vaccine* class, penalizing systems that achieve high accuracy by predominantly predicting the dominant category.

4 Results and Analysis

In this section, we present and analyze the performance of our proposed CSECU-Learners system on the EEUCA 2026 shared task on multimodal identification of vaccine critical content on social media. The full training set is used to train our proposed model, while the validation set is reserved exclusively for hyperparameter tuning. The official evaluation metric is the macro-averaged F1 score, as described in the previous section.

4.1 Performance Comparison with Participating Systems

Table 3 summarizes the performance of our system alongside a selection of participating teams. Our CSECU-Learners system (Codabench username: anchy) achieved a macro-averaged F1 score of 0.8308 and an accuracy of 0.8341, securing 6th place among all participating teams. These results demonstrate that our proposed two-stage early fusion architecture, which consolidates complementary signals from RoBERTa, ViT, and ViLT, yields competitive performance on this multimodal classification task.

Among the top-ranked systems, *lili12* attains the highest macro-F1 of 0.8494, outperforming our system by a margin of 0.0186 points. The 2nd and 3rd ranked systems, *TIU-MI* and *CUET_Synthetica*, achieve macro-F1 scores of 0.8389 and 0.8357, respectively, placing them 0.0081 and 0.0049 points ahead of our submission. Notably, the performance gap between the 1st and 6th ranked systems is relatively narrow at approximately 1.86 percentage points, indicating that our framework operates within a highly competitive performance band at the upper end of the leaderboard. In contrast, the lower-ranked systems exhibit considerably weaker results. The 23rd, 24th, and 25th ranked teams *abs123*, *thatgrass*, and *kannanrrk* record macro-F1 scores of 0.7846, 0.7754, and 0.7436, respectively. Our system surpasses these by margins of 0.0462, 0.0554, and 0.0872 points. The ranking is not unique for each team. On the Codabench test phase leaderboard, we observe multiple entries un-

Team	Macro-F1	Accuracy	Precision	Recall	Rank
lili12	0.8494	0.8517	0.8494	0.8517	1st
TIU-MI	0.8389	0.8420	0.8386	0.8409	2nd
CUET_Synthetica	0.8357	0.8390	0.8383	0.8359	3rd
alexcristea72	0.8340	0.8380	0.8338	0.8351	4th
CUET_Synthetica	0.8332	0.8361	0.8345	0.8340	5th
CSECU-Learners (Ours)	**0.8308**	**0.8341**	**0.8309**	**0.8309**	**6th**
abs123	0.7846	0.7912	0.7868	0.7864	23rd
thatgrass	0.7754	0.7844	0.7858	0.7802	24th
kannanrrk	0.7436	0.7502	0.7435	0.7437	25th

Table 3: Comparative performance of selected systems on the EEUCA 2026 shared task. Our system is highlighted in bold.

der the same team. For example, CUET_Synthetica appears in both 3rd and 5th positions on the leaderboard.

4.2 Analysis of Different Modality Baseline Models

To motivate the design of our proposed multi-encoder fusion framework, we conduct a systematic baseline analysis across three modality categories: textual, visual, and multimodal. For the textual baseline, we adopt the Twitter RoBERTa model, which is particularly well-suited to this task, given that the underlying data originates from social media platforms with informal and hashtag-rich linguistic characteristics. As the visual baseline, we employ ViT, which has demonstrated strong performance across a broad range of image classification benchmarks. For the multimodal baseline, we utilize ViLT, which processes both image and text modalities within a single unified transformer, affording equal priority to visual and linguistic signals through its linear modality interaction mechanism. Each baseline model is evaluated independently on the validation set, and the results are reported in Table 4.

Method	Modality	Prec.	Rec.	Macro-F1
Twitter RoBERTa	Text	0.8084	0.8100	0.8082
ViT	Image	0.6991	0.6993	0.6973
ViLT	Multimodal	0.7718	0.7723	0.7716

Table 4: Performance of individual modality baseline models on the validation set. Here Prec. and Rec. indicate Precision and Recall metrics respectively.

Among the three baselines, Twitter RoBERTa attains the highest macro-F1 score of 0.8082, demonstrating that the textual content carries the most discriminative signal for identifying vaccine-critical content. This is consistent with the nature of the task, where vaccine critical stance is frequently expressed through explicit linguistic cues such as hashtags, emotionally charged phrasing, and misinformation-laden statements. ViT, operating solely on the visual modality, records the weakest performance with a macro-F1 of 0.6973, falling 11.09 percentage points below RoBERTa. This substantial gap suggests that visual features alone are insufficient for reliable vaccine-critical content identification.

ViLT, which jointly encodes image and text within a single transformer through cross-modal attention, achieves a macro-F1 of 0.7716 — surpassing ViT by 7.43 percentage points but falling 3.66 percentage points short of RoBERTa. This intermediate performance highlights both the benefit and the limitation of ViLT in this setting. While its multimodal design allows it to leverage visual-textual interactions, its restricted maximum sequence length of 40 tokens prevents it from fully encoding the often lengthy post captions this dataset, as discussed in Section 2.2.

4.3 Ablation Study

To quantify the individual contribution of each component in our proposed framework, we conduct an ablation study on the validation set by selectively disabling one component at a time while keeping the remaining components. The results are presented in Table 5.

Our proposed CSECU-Learners system achieves the highest macro-F1 of 0.8251, confirming that every component contributes positively to the overall performance. Replacing Focal Loss with standard cross-entropy loss reduces the macro-F1 from 0.8251 to 0.8216, a drop of 0.35 percentage points. It indicates that the class imbalance present in the training data, where *Pro-vaccine* samples constitute approximately 39% of the corpus. Focal Loss effectively mitigates this by suppressing the gradi-

Method	Macro-F1
CSECU-Learners	**0.8251**
−Focal Loss	0.8216
−Fusion 2	0.8180
−Fusion 1	0.7716

Table 5: Ablation study results on the validation set. Each row removes one component from the full system. −Focal Loss denotes training with standard cross-entropy loss; −Fusion 2 removes ViLT and retains only the weighted RoBERTa–ViT fusion; −Fusion 1 removes Stage 1 and relies solely on ViLT.

ent contribution of easily classified majority-class samples.

Disabling Stage 2 fusion, that is, discarding the ViLT pooler output and relying solely on the Stage 1 weighted combination of RoBERTa and ViT, results in a macro-F1 of 0.8180, a decline of 0.71 percentage points relative to the full system. This indicates that the cross-modal attention mechanism of ViLT contributes complementary vision–language interaction signals that the independently encoded RoBERTa and ViT representations alone cannot fully replicate.

The most pronounced degradation occurs when Stage 1 fusion is removed entirely, reducing the system to ViLT alone and yielding a macro-F1 of 0.7716, a drop of 5.35 percentage points relative to the full model. The substantial performance recovery achieved by incorporating Stage 1, which pairs RoBERTa's full-sequence textual encoding with ViT's visual representation. It addresses the ViLT's architectural sequence length limitation and captures the richer linguistic content present in vaccine-critical social media posts.

5 Conclusion and Future Direction

In this study, we tackle the problem of detecting vaccine-critical memes in a multimodal setting as part of the EEUCA 2026 shared task. We introduce a two-stage early fusion framework built on multiple transformer-based encoders. In the first stage, representations from RoBERTa and ViT are merged using a weighted summation guided by their relative performance. In the second stage, this fused representation is further integrated with the joint image–text embedding generated by ViLT. To better handle difficult samples and class imbalance during training, we adopt focal loss as the optimization objective. Experimental findings indicate that the proposed method achieves strong performance, demonstrating its capability in identifying vaccine-critical memes.

For future work, we plan to investigate advanced transformer models that are pre-trained on biomedical corpora, as they may provide more domain-relevant representations. Additionally, we aim to replace the fixed fusion formulation in Stage 1 with a learnable scalar or a lightweight attention-based gating mechanism, allowing the model to adaptively weight feature contributions and improve generalization across diverse datasets.

Limitations

Despite its effectiveness, our approach has several limitations. The use of multiple transformer-based encoders and their fusion increases computational cost, making the model relatively slow and resource-intensive. In this work, we rely on base-sized transformers; however, larger variants are known to achieve better performance in many tasks, which we did not investigate here. Moreover, the performance of the model is influenced by manual hyperparameter tuning, which may vary across different datasets and may not always generalize well to real-world applications.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 1644–1650.

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.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Xiyang Du, Dou Hu, Jin Zhi, Lianxin Jiang, and Xiaofeng Shi. 2022. Pali-nlp at semeval-2022 task 6: isarcasmeval-fine-tuning the pre-trained model for detecting intended sarcasm. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 815–819.

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)*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

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*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Usman Naseem, Matloob Khushi, Jinman Kim, and Adam Dunn. 2021. Classifying vaccine sentiment tweets by modelling domain-specific representation and commonsense knowledge into context-aware attentive gru. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G Dunn. 2023. A multimodal framework for the identification of vaccine critical memes on twitter. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 706–714.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 2783–2796.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. Momenta: A multimodal framework for detecting harmful memes and their targets. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 4439–4455.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Laxmi Thapa, Aryaman Jain, Lakshmojee Koduru, Surabhi Adhikari, Junaid Rashid, Jungeun Kim, Surendrabikram Thapa, and Usman Naseem. 2026a. Concept-grounded detection of vaccine misinformation in multimodal content using interpretable vision-language models. In *Companion Proceedings of the ACM on Web Conference 2026*.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Kritesh Rauniyar, Laxmi Thapa, Surabhi Adhikari, Kristina T. Johnson, Ali Hürriyetoğlu, Hristo Tanev, and Usman Naseem. 2026b. Multimodal identification of vaccine content stance on social media. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Svitlana Volkova, Ellyn Ayton, Dustin L Arendt, Zhuanyi Huang, and Brian Hutchinson. 2019. Explaining multimodal deceptive news prediction models. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 659–662.

Zuhui Wang, Zhaozheng Yin, and Young Anna Argyris. 2020. Detecting medical misinformation on social media using multimodal deep learning. *IEEE journal of biomedical and health informatics*, 25(6):2193–2203.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement 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.

Li Zhang, Haimeng Fan, Chengxia Peng, Guozheng Rao, and Qing Cong. 2020. Sentiment analysis methods for hpv vaccines related tweets based on transfer learning. In *Healthcare*, volume 8, page 307. MDPI.

Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.

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

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.

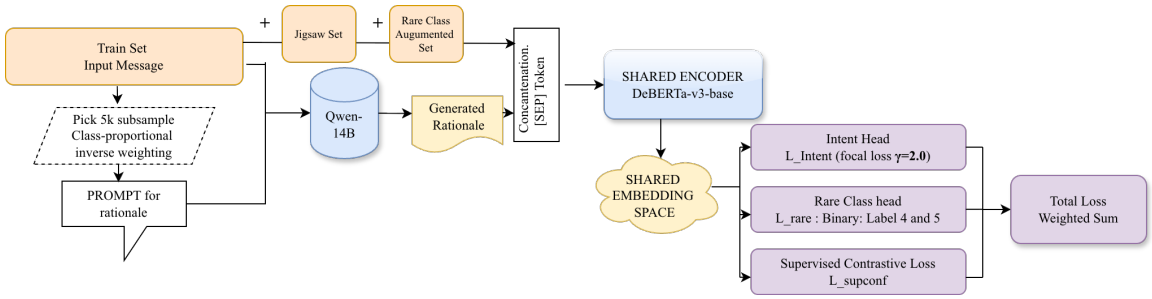


Figure 1: RAKSHAK architecture. The shared DeBERTa-v3-base encoder receives concatenated input messages and teacher-generated rationales. Three loss components operate over the shared embedding space: Focal Loss on the 6-class intent head, binary cross-entropy on dedicated Label 4 and Label 5 heads (weighted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

_alexcristea@EEUCA 2026: A Robust Early-Fusion ERNIE Pipeline for Multimodal COVID-19 Vaccine Meme Classification

Cristea Alexandru-Marian
University of Bucharest
alexandru-marian.cristea@s.unibuc.ro

Ionescu Costin Ioan
University of Bucharest
costin-ioan.ionescu@s.unibuc.ro

Abstract

This paper presents our team's submission for the EEUCA 2026 shared task on Multimodal Vaccine Critical Meme Detection. To tackle the inherent challenges of internet sarcasm and high label noise, we propose a robust, heavily regularized early-fusion text pipeline. Bypassing computationally heavy visual encoders, we extract text directly from meme images via OCR, concatenate it with the user's social media post, and process the unified context through an ERNIE-2.0-Large encoder. To combat severe overfitting, we replace the standard classification head with a Multi-Sample Dropout architecture ($p = 0.3$). Our optimized, lightweight text-only pipeline achieved a peak Macro F1 score of 0.834, securing 4th place on the official leaderboard. Furthermore, an ablation study utilizing Focal Loss demonstrates that our primary solution using standard Cross-Entropy provides superior robustness against the inherent label noise found in internet meme datasets.

1 Introduction

In recent years, social media memes have become a primary vehicle for both public health communication and the spread of medical misinformation (Thapa et al., 2024). While memes can be used to promote awareness, they are also frequently used to spread vaccine skepticism and vaccine-critical narratives, often employing heavy sarcasm, irony, and culturally specific visual puns to subvert their literal text. This ambiguity makes automated stance detection highly susceptible to overfitting, as models tend to memorize surface-level lexical cues rather than learning generalized semantic representations.

The EEUCA 2026 shared task (Thapa et al., 2026b), held in conjunction with the EEUCA Workshop (Hürriyetoğlu et al., 2026), provides a targeted benchmark for this challenge. The overarching goal of the competition is to advance reliable systems for monitoring vaccine-related discourse and supporting myth-debunking efforts on social media platforms.

While previous state-of-the-art approaches to this task have relied on massive, multi-stream architectures combining Graph Neural Networks and deep image encoders, we propose a highly efficient early-fusion text architecture relying on the Enhanced Representation through Knowledge Integration (ERNIE) framework. By explicitly demarcating the original social media post from the embedded image text, we allow the transformer's self-attention mechanism to cross-reference contextual cues effectively without the latency of visual feature extraction. Aggressively regularized via Multi-Sample Dropout and Cross-Entropy loss, our pipeline successfully secured 4th place on the final leaderboard with an F1 score of 0.834.

2 Background

Task Setup and Dataset: The EEUCA 2026 shared task requires systems to process a multimodal input (a social media text post paired with an image) and predict a single output stance representing the post's attitude toward COVID-19 vaccines. For example, an input might consist of a user's post saying "They want us to take it" paired with an image of a grim reaper. The target output is a three-class classification: *Vaccine Critical*, *Neutral*, or *Pro-Vaccine*.

To train and evaluate systems, the shared task utilizes a curated dataset of over 10,000 multimodal social media posts (Thapa et al., 2026a; Naseem et al., 2023). The dataset consists entirely of English-language content originating from social media platforms. The genre is highly informal, characterized by internet slang, grammatical irregularities, and image-macro memes. The annotation schema for determining the stance relies on complex multimodal interactions, mirroring methodologies established in prior multimodal hate

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 185–191

speech and crisis informatics research (Bhandari et al., 2023).

Related Work: Early work in automated stance detection primarily focused on text-only social media posts, consistently struggling with implicit sentiment, sarcasm, and irony. With the emergence of memes as a dominant form of internet communication, the field shifted toward multimodal approaches. Datasets mapping COVID-19 discourse have shown that vaccine-critical memes often rely on deep cultural context and dog-whistles rather than explicit anti-vaccine terminology (Naseem et al., 2023). Furthermore, while Large Language Models (LLMs) have shown immense promise in computational social science for capturing these nuanced social dynamics (Thapa et al., 2025), their application to non-compositional memes (where a benign image paired with a benign text creates a highly sarcastic message) remains challenging. Consequently, researchers frequently utilize massive Large Vision-Language Models (LVLMs) to capture cross-modal interactions. However, these dual-encoder systems often suffer from high computational costs and struggle to align disparate visual and textual feature spaces.

3 System Overview

We propose a highly regularized, early-fusion text pipeline that bypasses the computational overhead of dual-encoder LVLMs. The core algorithm relies on reducing the multimodal task into an advanced text-pair classification problem.

Concrete Algorithmic Flow Example: 1. *Input:* A user posts an image of a grim reaper holding a sign, accompanied by the social media caption: *"They want us all to take it."* 2. *Extraction:* We extract the text embedded within the image via OCR: *"COVID VACCINE WILL KILL YOU!"* 3. *Early Fusion:* The algorithm concatenates these texts into a single string separated by tokens: *"[CLS] They want us all to take it. [SEP] COVID VACCINE WILL KILL YOU! [SEP]"* 4. *Encoding:* ERNIE 2.0 processes this unified string, allowing its self-attention mechanism to instantly recognize the contextual contrast between the vague post and the aggressive image text. 5. *Prediction:* The pooled representation passes through our Multi-Sample Dropout head, predicting *"Vaccine Critical"*.

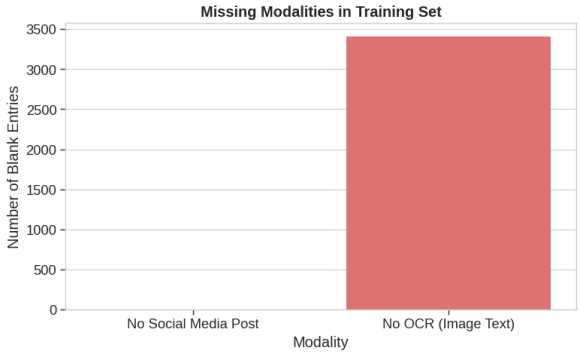


Figure 1: Missing modalities in the training set. A significant portion of the dataset lacks OCR text from the image, necessitating an early-fusion approach to avoid dropping critical contextual data.

3.1 Data Processing and Early Fusion

Because memes frequently rely on the juxtaposition between an image's embedded text and a user's accompanying caption, processing these modalities independently limits the model's ability to capture irony. Furthermore, as illustrated in Figure 1, a massive chunk of the dataset is missing either the social media post or the OCR image text entirely. To address this, we employ our early-fusion strategy.

For a given data instance, we define the social media caption as T_{post} and the OCR-extracted text from the image as T_{image}. We concatenate these into a single sequence, separated by the special '[SEP]' token, allowing the Transformer's self-attention mechanism to cross-reference contextual cues between the two sources:

$$S = [\text{CLS}] \oplus T_{post} \oplus [\text{SEP}] \oplus T_{image} \oplus [\text{SEP}]$$

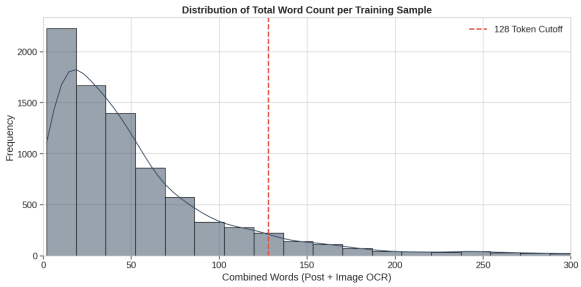


Figure 2: Distribution of the combined word count (post text + OCR image text). The 128-token cutoff captures the vast majority of instances without introducing unnecessary padding latency.

The combined sequence is tokenized using the ERNIE subword tokenizer. As shown in Figure 2, a maximum sequence length of 128 tokens perfectly

encapsulates the vast majority of the combined text sequences. Sequence lengths are strictly padded or truncated to this limit to accommodate the relatively short nature of meme text while maintaining optimal computational efficiency.

3.2 The ERNIE 2.0 Encoder

We utilize the pre-trained ERNIE 2.0-Large (Enhanced Representation through Knowledge Integration) model as our core representation learner (Sun et al., 2020). Unlike standard BERT, ERNIE is pre-trained via continual multi-task learning, explicitly capturing lexical, syntactic, and semantic information through entity-level masking. This makes it particularly adept at recognizing the named entities (e.g., vaccine names, political figures) heavily prevalent in COVID-19 discourse. We extract the hidden state of the '[CLS]' token from the final layer, $h \in \mathbb{R}^d$, where $d = 1024$, to serve as the aggregate representation of the meme.

3.3 Multi-Sample Dropout Classification Head

Standard Transformer classifiers utilize a single dropout layer preceding a linear transformation. However, due to the high variance and ambiguity in subjective meme datasets, a single dropout mask can inadvertently zero-out the specific neurons holding crucial "sarcasm" features for a given batch, leading to unstable gradient updates and poor generalization.

To combat this, we replace the standard head with a Multi-Sample Dropout architecture (Inoue, 2019). Instead of a single pass, the pooled representation h is passed through N parallel, independent dropout layers. The surviving vectors are all processed by the exact same fully connected linear classifier, and the resulting predictions are averaged to produce the final output logits:

$$\text{Logits} = \frac{1}{N} \sum_{i=1}^{N} W(Dropout_i(h)) + b$$

In our optimal configuration, we set $N = 5$ and apply a heavy dropout rate of $p = 0.3$. This architecture acts as an implicit ensemble within a single forward pass, aggressively regularizing the network and smoothing the loss landscape without introducing additional trainable parameters.

3.4 Loss Function and Optimization

As illustrated in Figure 3, the dataset exhibits a class imbalance, which naturally biases standard neural networks toward the majority "Pro-Vaccine" class. To mitigate this, we apply inverse class weighting to our loss functions. The weight for class c is calculated as:

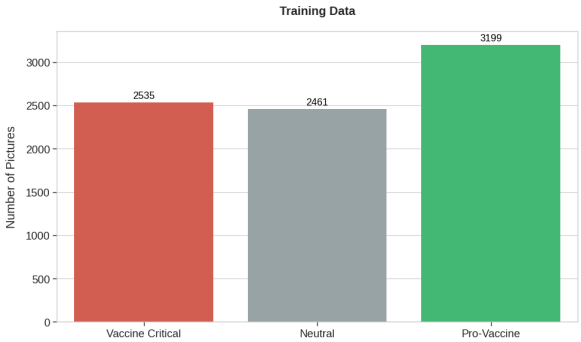


Figure 3: Class distribution of the training dataset, highlighting the natural class imbalance that necessitates our inverse-frequency weighted loss function.

$$w_c = \frac{N_{total}}{C \times N_c}$$

where N_{total} is the total number of training samples, $C = 3$ is the number of classes, and N_c is the number of samples in class c.

For our final solution, the model is optimized using the standard PyTorch Cross-Entropy Loss function, modulated by these class weights:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{C} w_i y_i \log(\hat{y}_i)$$

To validate this choice, we also formulated an alternate configuration utilizing Weighted Focal Loss (Lin et al., 2017) to test if aggressive hard-example mining would yield better results (detailed in Section 4.2). We optimize the network using AdamW with a learning rate of 2×10^{-5} and an increased weight decay of 0.05 to further penalize large weights. We employ a linear learning rate scheduler with a 10% warmup phase over 10 training epochs, utilizing mixed-precision training (FP16) with a gradient accumulation factor of 8 to achieve an effective batch size of 256.

4 Experimental Setup

4.1 Dataset Splits and Preprocessing

We evaluate our proposed pipeline on the official benchmark dataset provided for the EEUCA 2026 shared task (Naseem et al., 2023; Thapa et al., 2026a). The dataset consists of multimodal social media posts annotated for one of three stances:

Vaccine Critical, Neutral, and Pro-Vaccine. We strictly utilized the official shared task data splits: the training set was used for model optimization, the development (dev) set was used for hyperparameter tuning and ablation validation, and the hidden test set was exclusively used for the final leaderboard submission.

During preprocessing, we extracted the OCR text and social media caption, concatenating them using the standard '[SEP]' token. As qualitative analysis of the filtered vocabulary demonstrates, there are distinct lexical signatures for each class, confirming strong textual signals exist for our early-fusion pipeline to exploit. Based on text-length distributions, all sequences were truncated or padded to a maximum length of 128 tokens.

4.2 Implementation Details and Tools

Our pipeline was implemented using the PyTorch framework[1] and the HuggingFace Transformers library[2]. We initialized our text encoder using the pre-trained `ernie-2.0-large-en` weights.

To ensure the statistical validity of our findings and account for the high variance inherent in fine-tuning large language models on noisy data, every architectural variation in our ablation study was trained across three distinct random seeds (42, 22, and 100). The network was optimized using AdamW with a learning rate of 2×10^{-5} and mixed-precision training (FP16). Due to space constraints, the complete list of hyperparameter configurations required to replicate our experiments is provided in Appendix A.

4.3 Evaluation Measures

Following the official EEUCA 2026 Codabench evaluation protocol, system performance is primarily measured using the Macro F1-score. The Macro F1 calculates the F1-score independently for each of the three classes and then computes their unweighted average. This ensures that performance on the minority classes contributes equally to the final score, preventing systems from artificially inflating their metrics by over-indexing on the majority class.

[1] https://pytorch.org, version 2.0.1

[2] https://huggingface.co/docs/transformers, version 4.30.0

5 Results

5.1 Main Quantitative Findings

Our final optimized pipeline (ERNIE 2.0 + Multi-Sample Dropout + Cross-Entropy) achieved a peak Macro F1 score of **0.8340** on the official hidden test set. This performance secured the **4th place** ranking on the final EEUCA 2026 competition leaderboard, demonstrating that a highly regularized, early-fusion text-only pipeline is highly competitive against complex, multi-stream vision-language baseline models.

5.2 Quantitative Analysis: Ablation Studies

To better understand our design decisions, we conducted several ablations evaluated on the development split (summarized in Table 1).

The Impact of Multi-Sample Dropout: Our first experiment evaluated the impact of the classification head. The baseline, utilizing a single 10% dropout layer, achieved a 3-seed average of 0.8136 on the dev split. By replacing this with our Multi-Sample Dropout architecture (five parallel masks at $p = 0.3$), the average Macro F1-score significantly increased. This confirms that creating an implicit ensemble of dropout masks prevents over-indexing on noisy, batch-specific features.

Cross-Entropy vs. Focal Loss: To validate our choice of the primary loss function, we conducted an ablation utilizing Weighted Focal Loss. Contrary to our initial hypothesis, our final solution using standard Cross-Entropy Loss slightly outperformed the Focal Loss ablation. Focal Loss aggressively penalizes misclassifications on "hard" examples. However, in the context of subjective internet sarcasm, these "hard" examples are frequently mislabeled outliers. By forcing the optimizer to over-index on these noisy data points, the Focal Loss model suffered a degradation in generalization. Therefore, standard Cross-Entropy provides the highest robustness against dataset noise.

Cross-Lingual Knowledge Transfer: To test the boundaries of text-only meme classification, we conducted an ablation utilizing ERNIE 3.0 Base, an architecture trained natively on Chinese corpora, by translating the English memes into Chinese. As expected, this cross-lingual pipeline suffered a catastrophic performance drop to 0.7240 Macro F1. This confirms that meme comprehension relies heavily on culturally specific slang and regional context, which machine translation actively strips away.

Model Architecture	Macro F1-Score
Standard Head (Baseline)	0.8136 ± 0.0042
ERNIE 2.0 + MSD + Focal Loss (Ablation)	0.8221 ± 0.0028
ERNIE 2.0 + MSD + Cross-Entropy (Final)	$\mathbf{0.8238 \pm 0.0102}$
ERNIE 3.0 (Translated Data Ablation)	0.7240 ± 0.0013

Table 1: Macro F1-scores for our experimental models and ablation studies. Mean and standard deviation are calculated on the development split across three random seeds.

5.3 Error Analysis

To better understand the limitations of our text-only approach, we generated a confusion matrix based on the development split predictions (Figure 4). While the strong diagonal confirms the efficacy of our inverse class-weighting, the off-diagonal clusters reveal that our pipeline primarily struggles with two highly specific multimodal phenomena, illustrated in Figure 5.

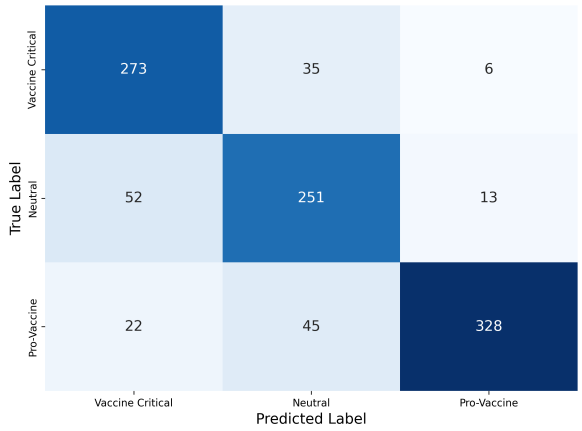


Figure 4: Confusion matrix of our final ERNIE 2.0 early-fusion pipeline on the development split. The diagonal demonstrates balanced learning, while off-diagonal clusters highlight specific vulnerabilities to satirical overlap and visual punchlines.

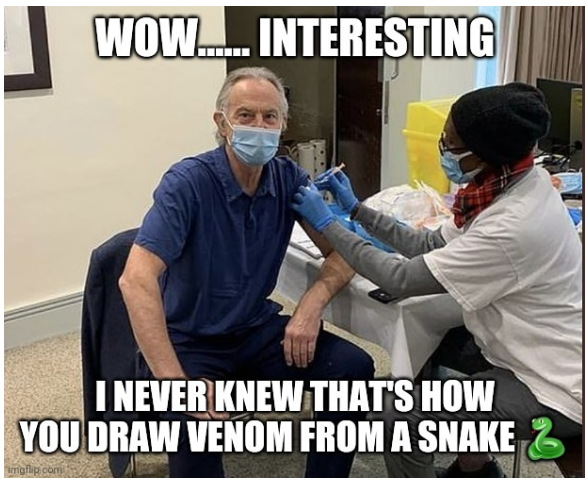
First, the model frequently fails due to the inability to parse complex pragmatic intent in "debunking" or satirical posts. As seen in Figure 5a, the model incorrectly predicts 'Vaccine Critical' due to the presence of highly critical hashtags, such as #vaccineinjury and #tonyblairisawarcriminal. Even though the user included the #vaxxed hashtag, indicating their actual Pro-Vaccine stance, the pooling mechanism fails to recognize that the user is presenting the other information as an object of ridicule. The model essentially lacks the human world knowledge required to distinguish between endorsement and ironic sharing.

Second, the text-only pipeline inherently fails on memes where the stance relies exclusively on visual sarcasm. When a user posts a benign, neutral, or even positive caption paired with a sarcastic "reaction face" (Figure 5b), the true stance is completely subverted by the visual modality. Because our ERNIE encoder is blind to facial expressions and visual tropes, it evaluates the positive text at face value and misclassifies the post. This confirms that while text-only pipelines are highly efficient and robust against general noise, achieving true human-level comprehension on internet sarcasm will ultimately require lightweight visual-feature integration.

6 Conclusion

In this paper, we presented a highly regularized, early-fusion text pipeline for the EEUCA 2026 COVID-19 multimodal meme classification task. Rather than relying on computationally expensive Large Vision-Language Models, we demonstrated that extracting image text via OCR and utilizing a specialized ERNIE 2.0 encoder provides a competitive and efficient alternative.

Crucially, our experiments highlight the dangers of applying hard-example mining techniques, such as Focal Loss, to inherently noisy internet datasets. We found that utilizing standard Cross-Entropy loss combined with a Multi-Sample Dropout architecture yields superior generalization by smoothing the loss landscape and ignoring mislabeled outliers. Without requiring task-specific generative vision fine-tuning, our robust text-only pipeline achieved a peak Macro F1-score of 0.834. Future work will explore applying this lightweight, noise-resistant classification head to open-source multimodal encoders to further capture nuanced visual irony without sacrificing training stability.



(a) Ablating Pragmatic Intent in Debunking Post

(b) Visual Sarcasm (Reaction Face)

Figure 5: Examples of systematic pipeline failures. **(a)** The model predicts 'Vaccine Critical' because it heavily over-indexes on critical hashtags like #vaccineinjury, completely missing the ironic, 'Pro-Vaccine' intent implied by the #vaxxed declaration. **(b)** The model fails to recognize the 'Vaccine Critical' stance because the sarcasm relies entirely on the visual reaction face, which our text-only pipeline cannot process.

7 Limitations

While our early-fusion text pipeline is highly efficient, it inherently fails on memes where the intended stance relies exclusively on visual sarcasm. When a user posts a benign, neutral, or even positive caption paired with a sarcastic "reaction face," the true stance is completely subverted by the visual modality. Because our ERNIE encoder is blind to facial expressions and visual tropes, it evaluates the text at face value and misclassifies the post. Furthermore, it struggles to parse complex pragmatic intents, such as users sharing critical misinformation specifically to debunk it. This confirms that while text-only pipelines are highly robust against general noise, achieving true human-level comprehension on multimodal internet sarcasm will ultimately require lightweight visual-feature integration.

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.

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)*.

Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. *Neural Networks*, 111:66–73.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G Dunn. 2023. A multimodal framework for the identification of vaccine critical memes on twitter. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 706–714.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8963–8970.

Laxmi Thapa, Aryaman Jain, Lakshmojee Koduru, Surabhi Adhikari, Junaid Rashid, Jungeun Kim, Surendrabikram Thapa, and Usman Naseem. 2026a. Concept-grounded detection of vaccine misinformation in multimodal content using interpretable vision-language models. In *Companion Proceedings of the ACM on Web Conference 2026*.

Surendrabikram Thapa, Kritesh Rauniyar, Hariram Veeramani, Aditya Shah, Imran Razzak, and Usman Naseem. 2024. Did you tell a deadly lie? evaluating large language models for health misinformation identification. In *International Conference on Web Information Systems Engineering*, pages 391–405. Springer.

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. 2026b. Multimodal identification of vaccine content stance on social media. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

A Hyperparameters and Reproducibility

To ensure full reproducibility of our ERNIE 2.0-Large fine-tuning experiments, we detail the complete set of hyperparameters in Table 2. All models were trained using PyTorch with Mixed Precision (FP16) on a single NVIDIA T4 GPU.

Hyperparameter	Value
Encoder Architecture	ERNIE 2.0 Large (English)
Max Sequence Length	128 tokens
Batch Size	32
Gradient Accumulation	8 steps
Effective Batch Size	256
Epochs	10
Optimizer	AdamW
Learning Rate	2×10^{-5}
Weight Decay	0.05
Learning Rate Scheduler	Linear with 10% Warmup
Multi-Sample Dropout	5 layers, $p = 0.3$

Table 2: Detailed training hyperparameters for the optimal Cross-Entropy + MSD configuration.

B Dataset Distribution and Class Weights

The EEUCA 2026 dataset exhibits a natural class imbalance reflecting real-world social media distributions. To prevent the model from collapsing into predicting only the majority class, we applied inverse frequency weighting to the loss functions. The class weights were calculated as $w_c = N_{total}/(C \times N_c)$.

C Detailed Per-Class Performance

While the primary evaluation metric is the unweighted Macro F1-score, analyzing the per-class F1-scores provides deeper insight into the model's behavior. As discussed in the main text, the baseline models consistently struggled with the "Neutral" class due to its highly ambiguous and sarcastic nature. The combination of early text fusion, Multi-Sample Dropout, and weighted Cross-Entropy yielded the most balanced performance across all three classes, preventing catastrophic failure on the minority "Vaccine Critical" and "Vaccine Neutral" instances while maintaining high accuracy on the clear-cut "Pro-Vaccine" memes.

To quantitatively evaluate these dynamics, we report the per-class Precision, Recall, and F1-scores on the development split in Table 3.

Class Stance	Precision	Recall	F1-Score
Vaccine Critical	0.7867	0.8694	0.8260
Neutral	0.7583	0.7943	0.7759
Pro-Vaccine	0.9452	0.8304	0.8841
Macro Average	0.8301	0.8314	**0.8238**

Table 3: Detailed performance breakdown per stance class evaluated on the development split for the final ERNIE 2.0 + MSD configuration.

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

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

Abstract

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

1 Introduction

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

The task presents several unique challenges:

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

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

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

2 Background

2.1 Task Description

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

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

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

2.2 Dataset

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

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

Table 1: Training set class distribution showing severe imbalance.

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

2.3 Related Work

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

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

3 System Overview

3.1 Model Architecture

We experimented with multiple architectures:

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

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

3.2 Prompt Engineering

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

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

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

3.3 Synthetic Data Augmentation

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

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

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

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

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

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

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

4 Alternative Approaches

We explored several alternative strategies that ultimately underperformed:

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

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

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

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

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

5 Experimental Setup

5.1 Training Configuration

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

5.2 Evaluation

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

6 Results

6.1 Main Results

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

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

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

6.2 Synthetic Data Ablation

Table 3 shows the critical sensitivity to synthetic ratio.

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

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

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

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

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

6.3 Per-class Performance

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

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

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

7 Analysis

7.1 The Validation Trap

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

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

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

7.2 Why 5% Synthetic Works

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

7.3 Error Analysis

Common error patterns include:

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

8 Conclusion

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

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

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

Limitations

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

Ethics Statement

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

about automated decisions, and appeal mechanisms for users.

References

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

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

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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

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

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

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

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

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

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

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

A Full Test Performance

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

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

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

B Additional Experimental Results

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

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

Table 7: Additional systems and ablations evaluated during development.

TAGA@EEUCA 2026: Token-Attribution Guided Attention for Fine-Grained Toxic Behaviour Classification in Online Gaming Communities

Akshyat Shah and Shashi Sah and Aryan Gupta and Kavinder Singh
Delhi Technological University
Delhi, India
{akshyatshah, sah.shashi2003, aryangupta0419, kavinder85}@gmail.com

Abstract

Online gaming involves large amount of people forming a large community of players who interact in real time. Toxic behavior in online chat is common and can harm players by deterring them. Thus, automated moderation is a necessity but difficult because game chat mixes domain-specific slang, deliberate obfuscation, informal "gamer" language , and limited coverage for categories such as threats and extremism. This paper describes the TAGA (Token-Attribution Guided Attention) system submitted to the EEUCA 2026 Shared Task on Understanding Toxic Behavior in Gaming Communities. We propose TAGA, an architecture that employs a leave-one-out attribution method using the Detoxify toxicity scorer to compute per-token attribution scores across multiple toxicity dimensions, which are then projected into the learned attention biases that steer the model toward toxicity-indicative tokens. By preparing a five phase ablation study, we demonstrate that each component: domain-specific preprocessing, focal loss with label smoothing, attribution-guided attention pooling, and dual-model Detoxify features with strategic oversampling contributes to a cumulative gain in macro-F1 score points over the DeBERTa-v3-base baseline reported. The final system achieves a test macro-F1 score of 0.618 and, importantly, produces non-zero predictions for extreme data imbalance present in the dataset used in the shared task.

1 Introduction

The rapid rise of the culture of online gaming has created several communities where millions of players interact with each other in real time over chat. While most of these interactions are fruitful and often positive, toxic behavior such as harassment, threats, hate speech, and extremism, still remains prevalent and can cause significant mental harm, psychological harm and damage to players (Kwak et al., 2015; da Silva et al., 2020). Hate speech detection using NLP has seen significant improvements with the advancements in language models (Parihar et al., 2021), yet moderating game chat at scale remains challenging due to the unique linguistic characteristics and gamer speak of the community.

Gaming chat exhibits several linguistic properties that distinguish it widely from standard social media text: extensive use of game-specific slang, deliberately obscuring chats through leetspeak substitutions (e.g., “naz1” for “nazi”), and highly compressed messages where a single word may carry the entire toxic intent (Märtens et al., 2015; Blackburn and Kwak, 2014). Furthermore, the distribution of toxicity categories is severely imbalanced: in the GameTox dataset (Naseem et al., 2025), non-toxic messages constitute over 81% of the data, while critical categories like Threats (0.14%) and Extremism (0.06%) contain fewer than 60 samples each. The annotation schema follows the hate speech categorization framework established in Bhandari et al. (2023).

The EEUCA 2026 Shared Task on Understanding Toxic Behavior in Gaming Communities (Thapa et al., 2026), which is organized as a part of the 9th Workshop on Event Extraction and Understanding (Hürriyetoğlu et al., 2026), provides a platform for developing and benchmarking automated toxicity detection systems on this challenging set of data. Naseem et al. (2025) introduced GameTox, which is a dataset consisting of 53K game chat utterances from the World of Tanks game and annotated for both utterance-level intent classification (6 classes) and token-level slot filling (4 slot types). Their baseline experiments showed that using Joint BERT (Chen et al., 2019) achieved the best performance among the 12 baseline models which were evaluated. However, the gap between the performance of slot-filling (0.99 F1) and intent classification (0.89 F1) suggests that understanding the overall intent of a message, particularly for rare

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 198–207

categories, still remains as a challenging task.

In this work, we propose **TAGA**, our submission to the EEUCA 2026 shared task. TAGA combines token-level toxicity signals and utterance-level intent classification through an attribution-guided attention mechanism. Rather than relying on manually annotated slot labels, TAGA automates computation of token-level toxicity attributions using leave-one-out token approach with the Detoxify toxicity scorer (Hanu and Unitary team, 2020). These attributions are then projected into the attention bias terms that guide the model toward determining the tokens which are the most indicative of the utterance's toxicity. Our key contributions are:

1. A **token-attribution guided attention** mechanism that injects externally computed toxicity signals into the attention bias and computes using a pre-trained language model.

2. A **multi-channel attribution** approach using leave-one-out perturbation across four toxicity classes to capture fine-grained token-level signals.

3. A comprehensive **domain-specific preprocessing** pipeline for gaming chat that handles the linguistics of gamer chat such as leetspeak, gaming abbreviations, and censored profanity.

4. A rigorous **five-phase ablation study** demonstrating a cumulative macro-F1 of 0.618 on the shared task.

2 Related Work

Toxicity detection in gaming: Early computational work for classifying game toxicity in chats used crowdsourced moderation signals and rich behavioral features: Blackburn and Kwak (2014) built a large scale corpus of utterances from the game-League of Legends and trained Random Forests classifiers over hundreds of in-game and chat-derived features to predict community level toxicity. Märtens et al. (2015) introduced a lexicon-driven annotation pipeline for the chat for the game-Dota 2 and released a resource called DotAlicious for analysis of utterance-level toxic vs. non-toxic. Stoop et al. (2019) proposed a conversation-aware modeling approach (the HaRe framework), which maintained per-user toxicity estimates and were updated with each new message for detection of real-time harassment in the game-League of Legends. The more recent, Yang et al. (2023) introduced ToxBuster, which conditioned line-level toxicity on the chat history and metadata across several multiplayer titles; in a post-game moderation setting they reported metrics which flagged 82.1% of chat-reported players at a precision of 90.0%, and identifying an additional 6% of toxic players who were not reported by other players.

Multi-class intent and joint token supervision (GameTox): Naseem et al. (2025) released GameTox, 53K utterances from the game-*World of Tanks* with six-way intent labels and token-level slot labels, together with human annotation for classification and LLM verification. On the English-only subset, their baselines indicated that intent level classification is substantially harder than simply using slot filling, the strongest joint model, Joint BERT (Chen et al., 2019), achieved an intent macro-F1 (I-F1) = 0.89, slot macro-F1 (S-F1) = 0.99, and intent accuracy (ICA) = 0.89. The explanation-centric framework called ToXCL proposed in (Hoang et al., 2024) reached an I-F1 = 0.87 and ICA = 0.88. Several of the LLM baselines are below the joint NLU models on this benchmark (e.g., The Gemma-7B model (Gemma Team et al., 2024) I-F1 = 0.74; The Mistral-7B model (Jiang et al., 2023) 0.69; Flan-T5-XL (Chung et al., 2024) 0.68; Llama-2-7B (Touvron et al., 2023) 0.65), this supported the claim that token-level slot supervision helps the models to cope with game-specific obfuscate toxic language.

Offensive language and hate speech (non-gaming benchmarks): The corresponding progress for the toxicity identification on social media included the identification of offensive language (Zampieri et al., 2019), large-scale abusive-behavior characterization (Founta et al., 2018),as well as detection of multi-aspect cyberbullying (Salawu et al., 2021), as well as benchmarks for rationale-grounded hate-speech (Mathew et al., 2021). These resources advanced the toxic language modeling from coarse to fine grain, but the schema for labels as well as domain differ from that of multi-intent game chat, where slang, obfuscation, and community norms dominate (Naseem et al., 2025).

3 Shared Task & Dataset

Task description: The EEUCA 2026 Shared Task on Understanding Toxic Behavior in Gam-

Label	Train	%	Val	%
Non-Toxic	34797	81.00	4349	81.03
Insults & Flaming	5925	13.79	740	13.79
Other Offensive	1874	4.36	234	4.36
Hate & Harassment	279	0.65	34	0.63
Threats	60	0.14	7	0.13
Extremism	24	0.06	3	0.06
Total	**42959**	100	**5367**	100

Table 1: Intent label distribution in our train/validation splits (test size: 5375).

ing Communities (Thapa et al., 2026; Hürriyetoğlu et al., 2026) challenges participants to classify game chat utterances into six toxicity intent categories. The task is framed as a single-label multi-class classification problem, evaluated on the macro-F1 score to equally weight rare and frequent classes to properly account for the class imbalances.

Dataset: The task uses the GameTox dataset (Naseem et al., 2025), comprising about 53,000 utterances from the World of Tanks game chat. The annotation schema follows the toxicity categorization framework established in Bhandari et al. (2023). Each utterance is classified into one of six categories: Non-Toxic, Insults and Flaming, Other Offensive Texts, Hate and Harassment, Threats, and Extremism. The annotations were produced through a human-LLM collaborative process with a three-phase schema achieving a Fleiss' Kappa of 0.91 (Falotico and Quatto, 2015).

Data split: The English-only subset of utterances is split into training (42959), validation (5367), and test (5,375) sets. As shown in Table 1, the dataset exhibits severe class imbalance, with Non-Toxic comprising 81.00% and Extremism only 0.06% of the data.

4 Methodology

Figure 1 illustrates the overall architecture of our proposed TAGA approach. It consists of four components:

4.1 Pre-processing

Gaming chat requires specialized normalization to recover the semantic meaning which is obscured in the content by informal writing conventions. We implement three domain specific methods utterance identifiers for this:

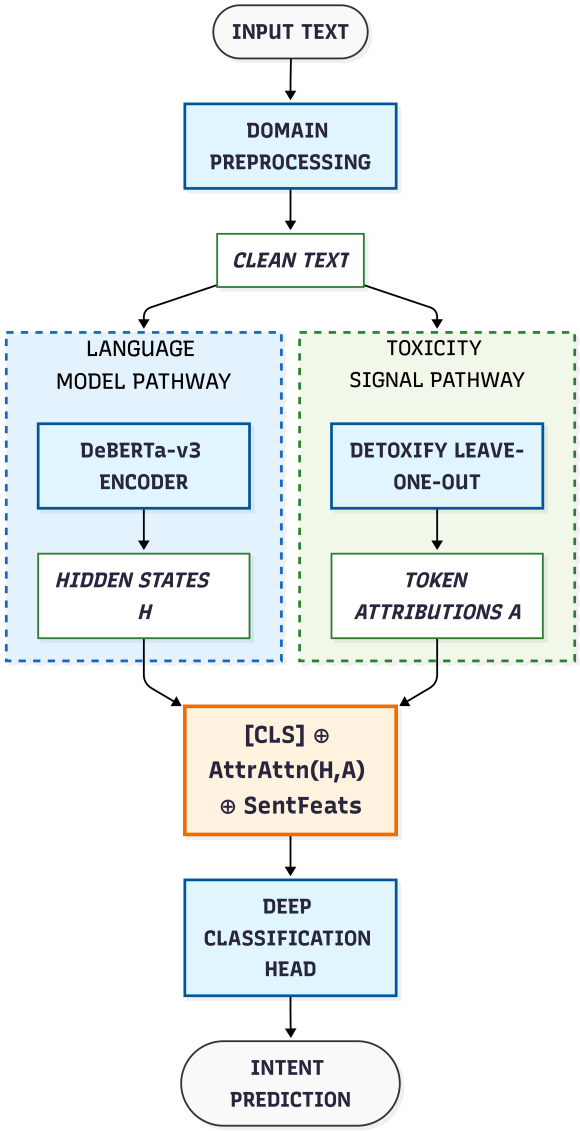


Figure 1: Overview of the TAGA architecture.

Leetspeak Decoding: We decode common character substitutes using numbers found in the dataset like (e.g., "naz1" → "nazi", "b1tch" → "bitch", "h1tler" → "hitler"). These substitutes are generally prevalent in Extremism and Hate categories where users attempt to evade the hate speech filters in various games.

Gaming Abbreviation Expansion: We manually handle 22 gaming-specific abbreviations which are commonly used in World of Tanks and other similar games (e.g., "kys" → "kill yourself", "stfu" → "shut the fuck up", "inting" → "intentional feeding"). This normalization is crucial as many abbreviations carry strong sentiments representing toxicity that would be opaque to the models and not classified as toxic if trained on standard English.

Uncensoring: We use regex-based patterns checks to recover censored words and phrases where characters are replaced with special symbols like (e.g., "f**k" → "fuck", "sh#t" → "shit"). This causes the model to understand and thus leverage the full semantic content of the utterance.

4.2 Token-Level Attribution

A key insight which motivates the use of TAGA is that the contribution of individual tokens to an utterance's toxicity can be estimated by measuring the effect on toxicity on their removal. We compute token-level attributions using the Detoxify toxicity scorer (Hanu and Unitary team, 2020), which is a suite of models trained on the Jigsaw toxicity datasets (Jigsaw/Conversation AI, 2018).

Leave-One-Out Attribution: For each utterance $\mathbf{x} = (w_1, w_2, \ldots, w_n)$, we first obtain baseline toxicity scores considering all tokens $\mathbf{b} \in R^C$ across $C = 4$ channels (toxicity, threat, insult, identity attack) using the Detoxify unbiased model. For utterances where $b_{\text{toxicity}} > \tau$ (we use $\tau = 0.15$), we compute the toxicity attribution of each token w_i by measuring the score drop when that token is removed:

$$a_{i,c} = \max\left(0,\ b_c - s_c(\mathbf{x}_{\setminus i})\right) \tag{1}$$

where $\mathbf{x}_{\setminus i}$ represents the utterance with token w_i removed and $s_c(\cdot)$ returns the score for the channel c. The $\max(0, \cdot)$ clipping makes sure that only tokens that contribute to decreasing toxicity contribute positively to toxicity attributions. This produces an attribution matrix $\mathbf{A} \in R^{n \times C}$ for each utterance.

Efficiency: We minimize attribution computation to the first 30 tokens per sentence and skip non-toxic utterances ($b_{\text{toxicity}} \leq \tau$), setting their attributions to zero which reduces the total number of Detoxify inference calls substantially and retains attributions for all of the remaining toxic content as required.

4.3 Sentence-Level Toxicity Features

In addition to the token-level attributions, we also extract sentence-level features from two Detoxify model variants:

$$\mathbf{f}_{\text{sent}} = [\mathbf{d}_{\text{unbiased}}; \mathbf{d}_{\text{original}}] \in R^{14} \tag{2}$$

where each $\mathbf{d} \in R^7$ contains score for toxicity, severe toxicity, obscenity, threat, insult, identity attack, and sexual explicitness. The dual-models used capture complementary perspectives on toxicity, as the two Detoxify variants were trained on completely different subsets of the Jigsaw data and different debiasing strategies.

4.4 TAGA Model Architecture

Encoder: We implement an encoder, pre-trained DeBERTa-v3-base model (He et al., 2023) as our backbone encoder, which produces contextualized token representations $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_L) \in R^{L \times d}$ where L is the sequence length and $d = 768$ is the hidden dimensions. DeBERTa's disentangled attention mechanism (He et al., 2021) separately encodes the content and position information and provides stronger representations than standard BERT.

Attribution-Guided Attention Pooling: Instead of relying solely on the [CLS] token, we compute a weighted sum of all the token representations and use attention scores that are biased using the token-level attributions. The attention logits are given as:

$$e_i = \underbrace{\mathbf{w}^\top \tanh(\mathbf{W}_a \mathbf{h}_i)}_{\text{content attention}} + \underbrace{g(\mathbf{a}_i)}_{\text{attribution bias}} \tag{3}$$

where $\mathbf{W}_a \in R^{256 \times d}$ and $\mathbf{w} \in R^{256}$ are learnable parameters for content-based attention, $\mathbf{a}_i \in R^C$ is the attribution vector for token i (aligned to subword tokens), and $g : R^C \to R$ is a learned projection:

$$g(\mathbf{a}) = \mathbf{v}^\top \text{GELU}(\mathbf{W}_g \mathbf{a} + \mathbf{b}_g) + b_v \tag{4}$$

with $\mathbf{W}_g \in R^{16 \times C}$. The attribution projection is initialized with small weights (std = 0.02) to ensure that the model relies initially on content-based attention and then gradually learns to incorporate token level attribution signals during training.

The attention-pooled representation is computed as:

$$\mathbf{h}_{\text{attn}} = \sum_{i=1}^{L} \alpha_i \mathbf{h}_i, \quad \alpha_i = \frac{\exp(e_i)}{\sum_j \exp(e_j)} \tag{5}$$

where padding positions are masked with $-\infty$ before applying softmax.

Classification Head: The final representation combines the [CLS] token, the attribution-guided attention pooling, and sentence-level features together into one:

$$\mathbf{z} = [\mathbf{h}_{\text{CLS}}; \mathbf{h}_{\text{attn}}; \mathbf{f}_{\text{sent}}] \in R^{2d+14} \quad (6)$$

This is then sent through a three-layer classification head:

$$\hat{\mathbf{y}} = \mathbf{W}_3 \cdot \text{GELU}(\mathbf{W}_2 \cdot \text{GELU}(\mathbf{W}_1 \mathbf{z})) \quad (7)$$

having hidden dimensions of 512 and 256, and dropout (rate 0.15) being applied after each activation.

4.5 Training

Focal Loss: In order to address the severe class imbalance, we use focal loss (Lin et al., 2017) alongside class-dependent weights:

$$\mathcal{L}_{\text{focal}} = -\sum_{k=1}^{K} w_k (1 - p_k)^{\gamma} y_k \log p_k \quad (8)$$

where $\gamma = 2.0$ is the focusing parameter, y_k is the one-hot target label with label smoothing (Szegedy et al., 2016) ($\epsilon = 0.05$), and $w_k = \sqrt{N/(K \cdot n_k)}$ are class weights derived from the *original* (pre-oversampling) class frequencies which are then clipped to $[1, 5]$.

Auxiliary Token-Level Loss: We implement an auxiliary token-level loss that allows the model to predict token-level toxicity from the hidden representations:

$$\mathcal{L}_{\text{aux}} = \lambda \cdot \beta(t) \cdot \text{MSE}\big(\sigma(\hat{\mathbf{t}}), \text{clamp}(\textstyle\sum_c \mathbf{a}_c, 0, 1)\big) \quad (9)$$

where $\hat{\mathbf{t}}$ are individual token predictions from a lightweight head, $\lambda = 0.02$ controls the auxiliary loss weight, and $\beta(t) = \max(0, 1 - t/T)$ is a linear anneal that reduces the auxiliary supervision over the epochs. This loss causes the encoder to develop token-level toxicity awareness early in training and is gradually relaxed as the main classification objective begins to take precedence.

Total Loss:

$$\mathcal{L} = \mathcal{L}_{\text{focal}} + \mathcal{L}_{\text{aux}} \quad (10)$$

Strategic Oversampling: We apply specific class level oversampling with augmentation to address the extreme class imbalance in the dataset. Target sample counts are set per class (e.g., 2,500 for Other Offensive, 600 for Threats, 300 for Extremism), with duplication up to 15× and text augmentation (word swaps, deletions, duplications, shuffling) for classes with fewer than 100 original samples. Oversampled copies however retain the original Detoxify features while augmented texts receive zero attributions, preventing the model from overfitting to noisy augmented attribution patterns.

Optimization: We use AdamW optimizer with differential learning rates: 1.5×10^{-5} for the encoder and 7.5×10^{-5} (5×) for the classification head and attribution projections. We also use a cosine scheduler with 6% warmup (Loshchilov and Hutter, 2017) that controls the learning rate over 5 epochs of training and has a batch size of 32. We employ mixed-precision training (Micikevicius et al., 2018) and gradient checkpointing (Chen et al., 2016) for memory efficiency on a single NVIDIA T4 GPU.

5 Experiments and Ablation Study

We conducted a rigorous ablation study to isolate the contribution of each component in the system. Starting from a DeBERTa-v3-base baseline, we incrementally added gaming-domain preprocessing, a redesigned loss function, architectural enhancements, and our proposed TAGA feature set incrementally. A final competition configuration (E5) is used to train on the combined train+validation split. With the release of true test labels, all experiments E0–E4 are now evaluated directly on the held-out test split. Performance is reported using Test Macro-F1.

5.1 Results

Table 2 summarizes the cumulative effect of each component on the true test Macro-F1.

5.1.1 Baseline (E0)

Five transformer models are fine-tuned with a single linear classification head on raw, uncleaned game-chat data using standard cross-entropy loss, without class balancing or preprocessing (Table 3). All five models remain below 0.51 macro-F1 which is a disconnect driven by the near-complete failure

Phase	Description	Macro-F1	Accuracy	Precision	Recall
E0	Vanilla DeBERTa-v3, CLS pooling, CE loss, raw text	0.4983	0.8789	0.4731	0.5348
E1	+ L33t decode + slang normalisation + uncensoring	0.5086	0.8798	0.4817	0.5425
E2	+ Focal loss (γ=2) + class weights + label smoothing	0.5114	0.8945	0.5013	0.5245
E3	+ CLS+attention pooling + deeper head + differential LR	0.4959	0.9003	0.4900	0.5047
E4	+ Dual Detoxify (14-d) + attribution + oversample + aux loss	0.5905	0.8915	0.5841	0.6012
E5	Full data (train+val), all E4 components	**0.6186**	**0.8902**	**0.6047**	**0.6497**

Table 2: Incremental test performance across phases.

Model	Macro-F1	Accuracy	Precision	Recall
BERT-base	0.4747	0.8969	0.5079	0.4588
HateBERT	0.4961	0.8966	0.5104	0.4876
ToxicBERT	**0.5044**	0.8997	0.5177	0.4935
RoBERTa-base	0.4976	**0.9018**	**0.5351**	0.4720
DeBERTa-v3-base	0.4983	0.8789	0.4731	**0.5348**

Table 3: Baseline Model Performance on Test Set (Phase E0) where models are arranged in ascending order of parameter size.

on minority classes, where Threats and Extremism yield near-zero F1 across all models. ToxicBERT achieves the best macro-F1 (0.5044), while RoBERTa-base leads in precision (0.5351) and accuracy (0.9018). DeBERTa-v3-base, despite the lowest precision (0.4731), achieves the highest recall (0.5348), reflecting greater sensitivity to minority class instances. Given its disentangled attention mechanism and ELECTRA-style pre-training, DeBERTa-v3-base is selected as the backbone for all subsequent phases.

5.1.2 Gaming-Domain Preprocessing (E1)

Gaming communities employ three main strategies to obscure toxicity in game chat that defeat the baseline tokenizer: Leetspeak ("n4z1" $\rightarrow$ "nazi"), community slang ("noob", "rekt"), and censored profanity ("f***" $\rightarrow$ "fuck"). We evaluate each normalisation step and then run it in combination reporting a test macro-F1 of 0.5086.

5.1.3 Loss Function Redesign (E2)

The six-class distribution is severely imbalanced towards non toxic utterances: Non-toxic (81.0%) versus Extremism (0.06%). The standard cross-entropy loss ignores this imbalance. Thus, we evaluate two specific alternatives:

Focal loss with weights and smoothing (E2b) achieves the highest test F1 of 0.5114, outperforming class-weighted CE (E2a, 0.4961). The combined configuration is nonetheless adopted for its stable training foundation across subsequent phases. Despite improvements, loss reweighting alone cannot overcome extreme data scarcity.

Variant	Configuration	Val F1	Test F1
E2a	Class-weighted CE	0.5061	0.4961
E2b	Focal + weights + smoothing	0.5102	0.5114

Table 4: Loss function ablation across class weighted CE and Focal loss with weights and smoothing.

5.1.4 Architecture Enhancements (E3)

We replace the single CLS token with a learned attention pooling mechanism that combines CLS with all token representations. Thus we hypothesized that toxic signals are concentrated in specific words or threat tokens rather than the global CLS embedding. We also deepened the classification head (Linear$\rightarrow$GELU$\rightarrow$Dropout$\rightarrow$Linear) and applied a $5\times$ differential learning rate multiplier to the head versus the backbone of the model. On the true test set, these changes yield a macro-F1 of 0.4959, a regression from the F1-score in E2. This drop suggests that the deeper head and attention pooling mechanism slightly overfit the training distribution on the unseen test data, confirming that architectural capacity alone is not the primary bottleneck. The performance gap is instead attributed to the absence of explicit toxicity signals, which motivates the feature-level augmentation introduced in E4.

5.1.5 Dual TAGA Feature Set (E4)

Phase 4 introduces our primary contribution to the task: combines the TAGA feature set which comprises (i) dual Detoxify sentence vectors, (ii) token-level attribution scores, (iii) minority-class oversampling, and (iv) an auxiliary toxicity regression loss. We ablate for the Detoxify components in Table 5.

E4a, which introduces sentence-level Detoxify features without oversampling or auxiliary supervision, yields a test macro-F1 of 0.4480, below the E0 baseline (0.4983). This regression indicates that Detoxify features alone are insufficient and may introduce noise without a training objective that directs the model to exploit them; the minority classes remain suppressed under standard cross-entropy. The combination of oversampling and auxiliary loss (E4b) is the decisive factor: even with a single-model 7-d Detoxify vector, test macro-F1 jumps to 0.5887 and Extremism F1 becomes non-zero (0.480) for the first time. The full dual-model configuration (E4) further improves Hate F1 to 0.500 and consolidates Extremism at 0.500. E4 achieves 0.5905, a significant improvement over the E0 baseline.

5.1.6 Final System Configuration (E5)

E5 maximises the available training signal by combining the train and validation splits before fine-tuning, ensuring that every annotated example - including the rarest Extremism (27 samples) and Threats (67 samples) - contributes to the final model. Training on the full labelled data yields a test macro-F1 of **0.618**, improving over E4 and our best submitted result.

5.2 Per-Class Analysis

Table 6 traces per-class F1 across all phases. Non-toxic and Insults remain stable across all phases, confirming that majority classes are well-captured from the baseline. The Hate class improves consistently from 0.476 (E0) to 0.542 (E2), driven by slang normalisation in E1 and loss redesign in E2, before settling at 0.500 in E4. Threats shows rising sharply from 0.338 (E0) to 0.410 (E1) due to L33t-speak decoding, dipping through E2-E3, and recovering to 0.420 in E5 with oversampling and auxiliary loss. Extremism remains zero across E0–E3 and first becomes non-zero in E4 (0.500), rising further to **0.667** in E5 the single largest per-class gain across all phases, directly attributable to the TAGA oversampling and token attribution strategy providing sufficient training signal for this 24-sample minority class.

5.3 Error Analysis

We believe that documenting negative results and error analyses is as valuable as reporting performance gains. Therefore, we detail both the successes of the TAGA system and the surprising failure modes that emerged across our five-phase ablation.

Architecture regression in E3. A counterintuitive finding is that the introduction of attention pooling and a deeper classification head in E3 *decreased* test macro-F1 by 1.55 pp relative to E2, despite these components being theoretically well-motivated. We hypothesize that the increased capacity of the head, combined with the absence of explicit toxicity signals, caused the model to overfit to the majority-class distribution of the training data. This result suggests that architectural complexity without complementary feature-level inductive bias is insufficient, and potentially harmful, for severely imbalanced datasets.

Detoxify features without oversampling are counterproductive. E4a, which adds sentence-level Detoxify features without oversampling or auxiliary supervision, produces a test macro-F1 of 0.4480 below the E0 baseline of 0.4983. This negative result reveals that injecting external toxicity signals into a model trained under standard cross-entropy on an imbalanced dataset can actively degrade minority class performance. The Detoxify features introduce a richer signal space that the model cannot exploit without a complementary training objective that forces recovery of minority classes. This finding motivates the joint introduction of oversampling and auxiliary loss in E4b, which together produce the decisive performance jump.

Confusion between Extremism and Hate. Even in our best configuration (E5), the model achieves only 0.469 on the Hate class despite achieving 0.667 on Extremism. Manual inspection of misclassified utterances reveals that group-targeted language involving political or ethnic references is frequently confused between the two categories, as the distinction requires contextual world knowledge that extends beyond the surface form of the utterance. This is an inherent limitation of the classification of the level of utterance without the

Variant	Components	Test F1	Threats	Extremism	Hate
E4a	7-d unbiased (Sent-level only) (no oversample, no aux)	0.4480	0.2500	0.000	0.3950
E4b	7-d unbiased + oversample + aux	0.5887	0.4200	0.4800	0.4750
E4	Dual TAGA (14-d + attr + oversample + aux)	**0.5905**	0.3880	**0.5000**	**0.5000**

Table 5: Detoxify component ablation showing the impact of sentence-level features, oversampling, and token attribution on Test F1 across the rarest toxicity classes.

Phase	Non-toxic	Insults	Other	Hate	Threats	Extremism	Macro-F1
E0 Baseline	0.952	0.768	0.456	0.476	0.338	0.000	0.498
E1 Preprocessing	0.951	0.764	0.418	0.509	0.410	0.000	0.508
E2 Loss redesign	0.946	0.762	0.460	0.542	0.350	0.000	0.511
E3 Architecture	0.948	0.765	0.442	0.495	0.325	0.000	0.495
E4 Dual TAGA	0.948	0.762	0.445	0.500	0.388	0.500	0.590
E5 Full TAGA	0.947	0.763	0.436	0.469	0.425	**0.667**	**0.618**

Table 6: Per-class Test F1 across ablation phases.

context of the discourse.

6 Conclusion

We presented our TAGA architecture, which is a submission to the EEUCA 2026 Shared Task on Understanding Toxic Behavior in Gaming Communities (Thapa et al., 2026). Through a systematic and rigorous five-phase ablation, we demonstrated that each component contributes meaningfully to the final result.

Our work demonstrates that extracting explicit toxicity attribution from pre-trained scorers can serve as an effective and meaningful inductive bias for attention-based models, providing a principled way to incorporate domain knowledge without manual human token-level annotation. Future work could explore extending TAGA to joint intent-slot classification, applying the attribution mechanism to other toxicity detection domains, and investigating gradient-based attribution methods to reduce the $O(n)$ preprocessing cost of leave-one-out perturbation.

Limitations

Our work has several limitations. First, the leave-one-out attribution computation requires $O(n)$ time forward passes through the Detoxify model per toxic utterance, making it expensive to run at scale. Gradient-based attribution methods could provide a more efficient alternative. Second, our approach is evaluated only on a single dataset (GameTox) from one game (World of Tanks), and generalization to other gaming communities with different linguistic norms and game talk remains to be validated. Third, the extreme rarity of certain categories (27 Extremism, 67 Threats samples total) makes robust evaluation statistically challenging, and results on these classes should be interpreted with appropriate uncertainty. Finally, our preprocessing pipeline relies on manually curated lexicons for leetspeak and gaming abbreviations, which may not generalize to evolving gaming slang and may require manual updating.

Ethical Considerations

This work involves the processing and classification of toxic, hateful, and extremist language from real-world gaming chat. While our goal is to advance automated moderation to protect players from harm, several ethical considerations warrant attention. First, the GameTox dataset contains genuine instances of hate speech, threats, and extremist content; access to and use of such data should be restricted to research purposes and handled responsibly to avoid amplifying harmful content. Second, automated toxicity classifiers are inherently imperfect, our system achieves a test macro-F1 of 0.618 and deploying such a system in a production moderation pipeline without human intervention risks both false positives, which may unfairly penalize legitimate players using game-specific language, and false negatives, which may allow genuinely harmful content to go unmoderated. Third, the preprocessing lexicons and Detoxify scorer used in TAGA reflect toxicity norms primarily from English-language Western gaming communities; application to other languages, cultures, or game genres may introduce cultural bias and should be validated independently before deployment. Finally, we acknowledge that systems trained to de-

tect extremist and hateful content could, if misused, be repurposed to identify and target individuals who express such views rather than to protect potential victims, and we strongly discourage any such application of this work.

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.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Bruno Mendes da Silva, Mirian Tavares, Filipa Cerol, Susana Mendes da Silva, Paulo Falcão, and Beatriz Isca Alves. 2020. Playing against hate speech – how teens see hate speech in video games and online gaming communities. volume 3, pages 34–52.

Rosa Falotico and Piero Quatto. 2015. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49:463–470.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Laura Hanu and Unitary team. 2020. Detoxify. https://github.com/unitaryai/detoxify.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Nhat M Hoang, Xuan Long Do, Duc Anh Do, Duc Anh Vu, and Luu Anh Tuan. 2024. Toxcl: A unified framework for toxic speech detection and explanation. *arXiv preprint arXiv:2403.16685*.

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)*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jigsaw/Conversation AI. 2018. Toxic comment classification challenge. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge.

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

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

Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*.

Marcus Märtens, Siqi Shen, Alexandru Iosup, and Fernando Kuipers. 2015. Toxicity detection in multiplayer online games. In *2015 International Workshop on Network and Systems Support for Games (NetGames)*, pages 1–6.

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.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh

Venkatesh, and Hao Wu. 2018. Mixed precision training. In *International Conference on Learning Representations*.

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.

Semiu Salawu, Jo Lumsden, and Yulan He. 2021. A large-scale english multi-label twitter dataset for cyberbullying and online abuse detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 146–156.

Wessel Stoop, Florian Kunneman, Antal van den Bosch, and Ben Miller. 2019. Detecting harassment in real-time as conversations develop. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 19–24.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826.

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)*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Zachary Yang, Nicolas Grenon-Godbout, and Reihaneh Rabbany. 2023. Towards detecting contextual real-time toxicity for in-game chat. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9894–9906, Singapore. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

A Hyperparameter Configuration

Table 7 lists the full hyperparameter configuration for the TAGA model.

Hyperparameter	Value
Encoder	DeBERTa-v3-base
Max sequence length	128
Batch size	32
Encoder learning rate	1.5×10^{-5}
Head learning rate multiplier	$5\times$
Weight decay	0.01
Dropout rate	0.15
Epochs	5
Warmup ratio	6%
Max gradient norm	1.0
Focal loss γ	2.0
Label smoothing ϵ	0.05
Auxiliary loss weight λ	0.02
Attribution toxicity threshold τ	0.15
Max attribution tokens	30
Attribution channels	4
Detoxify models	unbiased, original
Sentence features dim	14
Oversampling targets	
Other Offensive	2,500
Hate and Harassment	1,500
Threats	600
Extremism	300

Table 7: Full hyperparameter configuration for TAGA.

B Model Architecture Details

The TAGA model consists of the following components:

- **Encoder:** DeBERTa-v3-base (184M parameters) with gradient checkpointing enabled.
- **Attention projection:** Linear(d, 256) $\rightarrow$ Tanh $\rightarrow$ Linear(256, 1).
- **Attribution projection:** Linear(C, 16) $\rightarrow$ GELU $\rightarrow$ Linear(16, 1), initialized with $\mathcal{N}(0, 0.02)$.
- **Token head (auxiliary):** Linear(d, 64) $\rightarrow$ GELU $\rightarrow$ Dropout(0.15) $\rightarrow$ Linear(64, 1).
- **Classification head:** Linear($2d + 14$, 512) $\rightarrow$ GELU $\rightarrow$ Dropout(0.15) $\rightarrow$ Linear(512, 256) $\rightarrow$ GELU $\rightarrow$ Dropout(0.15) $\rightarrow$ Linear(256, 6).

LilyMeme@EEUCA 2026: Multimodal Vaccine Meme Stance Detection with Task-Adapted MemeCLIP and Complementary Ensembling

Yixuan Li[1], Xiaolong Yin[2], Yang Yang[1]*
[1]Nanjing University of Science and Technology
[2]Nanjing University
liyixuan25@njust.edu.cn, yinxl@lamda.nju.edu.cn, yyang@njust.edu.cn

Abstract

Memes have emerged as a prominent medium for conveying public sentiment on sensitive health topics such as vaccination. Unlike conventional multimodal tasks, memes feature implicit stances, sarcastic nuances, and complex cross-modal interactions, posing significant challenges for accurate stance detection. This paper presents our approach for the VaxMeme Shared Task @EEUCA 2026, which aims to classify vaccine-related memes into three distinct classes: Vaccine-critical, Neutral, and Pro-vaccine. Building upon MemeCLIP, we systematically enhance our framework via task-specific adaptation, lightweight cross-modal fusion, noise-aware training, LLM-assisted semantic augmentation, and inference-stage optimization, ultimately ensembling multiple complementary variants for final predictions. Our ensemble method achieves a Macro-F1 score of 0.8494 on the official test set, securing first place and demonstrating the critical efficacy of noise-aware training and late-stage ensembling for robust stance identification.

1 Introduction

With the development of social media, memes have become an important medium for expressing viewpoints on public issues, especially on sensitive topics such as vaccination, public health, and misinformation. Unlike ordinary text or images, memes often rely jointly on images, short text, embedded image text, sarcastic rhetoric, and symbolic visual elements to convey opinions. Therefore, automatically identifying their stance is more difficult than general classification tasks. The Shared Task on Multimodal Vaccine Critical Meme Detection (VaxMeme) @EEUCA 2026 was proposed precisely around this challenge (Thapa et al., 2026b; Hürriyetoğlu et al., 2026). This task focuses on fine-grained stance understanding in public health scenarios and provides a valuable benchmark for studying multimodal reasoning under noise, indirect expression, and social context.

Figure 1: An example from the shared task dataset labeled as Vaccine critical.

This shared task is challenging in multiple aspects. First, memes usually contain heterogeneous information sources, which may be complementary, partially redundant, or even semantically conflicting. Second, vaccine-related memes often rely on sarcasm, exaggeration, and implicit rhetoric, making literal interpretation of the text often unreliable. Third, some samples contain only weak textual evidence, noisy OCR, or highly compressed symbolic visual content, so the model must rely on subtle multimodal cues rather than explicit sentiment words to determine stance (Naseem et al., 2023; Thapa et al., 2026b).

To address these difficulties, we take MemeCLIP (Shah et al., 2024) as the core framework and design multiple enhancement methods for the VaxMeme shared task. We first reorganize the input for VaxMeme, introduce explicit missing-text markers, replace the original simple fusion strategy with lightweight token-level cross-modal interaction, and strengthen training with stratified

*Corresponding author.

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 208–215

cross-validation, class weighting, label smoothing, and cosine learning rate scheduling. On this basis, we further explore multiple complementary enhancement methods, including noise-aware weighted training, an LLM-assisted semantic description branch, a three-branch architecture that explicitly distinguishes post text from OCR text, and inference-stage enhancement that combines test-time augmentation with retrieval priors. Our final submission does not rely on any single model, but is instead obtained by ensembling multiple complementary variants.

2 Related Work

In recent years, multimodal meme understanding has evolved from coarse-grained harmful content detection to fine-grained pragmatic and stance analysis (Guan et al., 2025). Early research predominantly focused on identifying hate speech, offensive content, and humor; however, the proliferation of novel datasets and advanced vision-language models has shifted attention toward more complex semantic properties embedded within memes, including targets, stances, and implicit contexts.

Recent studies (Liang et al., 2024; Yang et al., 2024) have systematically explored text-image stance detection by curating diverse social media datasets and introducing target-aware cross-modal prompting strategies. Furthermore, the introduction of the PrideMM dataset and the associated MemeCLIP framework (Shah et al., 2024) transitioned meme analysis from isolated harmful content detection to a comprehensive multi-task paradigm encompassing hate, target, stance, and humor recognition. Subsequently, the CASE 2025 Shared Task formalized multimodal stance recognition in meme scenarios by establishing it as an independent evaluation track (Thapa et al., 2025). Collectively, these advancements highlight that the core challenge of meme stance recognition extends beyond naive image-text fusion; it necessitates the intricate modeling of multimodal synergies, contradictions, and the underlying socio-cultural contexts (Yu et al., 2026; Jiang et al., 2025).

In comparison, research on stance classification for vaccine-related memes remains relatively limited. MMCoVaR provides a multimodal dataset for COVID-19 vaccines, covering both news and tweets, for misinformation- and credibility-related classification tasks (Chen et al., 2021). (Naseem et al., 2023) were the first to systematically introduce the VaxMeme task and dataset, collecting a large-scale manually annotated set of vaccine-related memes and designing a multimodal framework that combines global and local representations for identifying vaccine-critical memes. More recent work has also started to extend multimodal vaccine-content analysis from coarse-grained classification toward more interpretability-oriented directions, and the EEUCA 2026 Shared Task can be seen as a natural extension of earlier multimodal meme stance research into the public-health domain (Thapa et al., 2026a,b).

3 Task and Dataset

3.1 Task Definition

The Shared Task on Multimodal Vaccine Critical Meme Detection (VaxMeme) @EEUCA 2026 is a three-class multimodal stance identification task (Thapa et al., 2026b; Hürriyetoğlu et al., 2026). For each meme with a unique identifier `index`, the model is required to determine its stance toward vaccination, namely Vaccine critical, Neutral, or Pro-vaccine. The shared task adopts Macro-F1 as the primary evaluation metric, which means that the model must not only achieve strong overall classification performance, but also maintain as balanced recognition performance as possible across the three classes. The challenge of this task mainly comes from the complexity of meme expression. On the one hand, the stance of a meme is often not directly expressed through a single modality, but is jointly conveyed through the interaction between image and text. On the other hand, sarcasm, exaggeration, metaphor, and image–text incongruity are very common in this type of data, making it difficult to fully model the true semantics by relying only on visual features or only on textual features.

3.2 Dataset

This study utilizes the official dataset curated for the EEUCA 2026 Shared Task on Multimodal Vaccine Critical Meme Detection. This corpus is primarily derived from the VaxMeme dataset alongside associated data collection initiatives (Thapa et al., 2026b; Naseem et al., 2023; Bhandari et al., 2023; Thapa et al., 2026a). Originally introduced by (Naseem et al., 2023) for multimodal vaccine-critical meme identification, VaxMeme comprises over 10,000 English meme samples sourced from Twitter. These samples encapsulate both visual and textual modalities, where the embedded text within

Table 1: Class distribution of image samples in the shared task dataset.

Dataset	Vaccine critical	Neutral	Pro-vaccine	Total
Train	2535	2461	3199	8195
Val	308	327	389	1024
Test	314	316	395	1025

the images is automatically extracted via optical character recognition (OCR) (Naseem et al., 2023). Following the official shared task configuration, the dataset is partitioned into training, validation, and test sets, comprising 8,195, 1,024, and 1,025 instances, respectively. Each instance consists of a meme image, the associated optical character recognition output (`image_text`), and the corresponding social media post (`post_text`). The classification schema encompasses three stance categories: Pro-vaccine, Vaccine-critical, and Neutral, with the detailed class distribution summarized in Table 1.

4 Method

4.1 MemeCLIP Baseline

Our work is built upon the MemeCLIP baseline (Shah et al., 2024). MemeCLIP is a CLIP-based framework for text-embedded meme classification, whose core idea is to preserve the pre-trained knowledge of CLIP while only performing lightweight adaptation on the downstream classification modules. Specifically, the original implementation adopts CLIP ViT-L/14 as the vision–language backbone and freezes its parameters (Radford et al., 2021); for each meme, the model first extracts image features and text features separately, and then maps them into a shared feature space through independent linear projection layers. Subsequently, both the image branch and the text branch pass through lightweight Adapters and are residually mixed with the original projected features. The original cross-modal fusion is implemented as element-wise multiplication between the image and text representations, and the final classification prediction is produced by a lightweight feed-forward layer together with a cosine classifier (Shah et al., 2024).

4.2 MemeCLIP for VaxMeme

MemeCLIP provides a starting point for text-embedded meme understanding, but it was not directly designed for VaxMeme. Therefore, we adapt MemeCLIP based on the requirements of the shared task.

First, we reorganize the original Pride-based meme classification setting into a three-class task for VaxMeme, and rebuild the data processing pipeline. We construct a unified metadata file and split the training/validation data using five-fold StratifiedKFold. Second, in terms of input construction, we no longer follow a single text field, but explicitly integrate two text sources: the post-level text `post_text` and the embedded text in the image, `image_text`, and organize them as a structured template

[POST] `post_text` [IMG] `image_text`.

We further introduce explicit missing markers `[NO_POST]` and `[NO_OCR]` to distinguish "missing text" from ordinary empty input. To improve robustness, we also apply lightweight text dropout during training, so as to reduce the model's over-reliance on any single textual signal.

To enhance the multimodal fusion mechanism, we upgrade the naive element-wise operation originally employed in MemeCLIP to a lightweight cross-modal Transformer. The input sequence to this module comprises three distinct entities: a parameterized `[CLS]` token, an image token, and a text token. Following modality-specific embedding and self-attention-driven interaction, the ultimate fused representation aggregates the updated `[CLS]` state with the cross-modal interaction terms derived from the visual and textual tokens. This architectural refinement substantially augments the capacity of the network to capture fine-grained, token-level cross-modal alignments.

4.3 Enhancement Methods

Noise-Aware Weighted Training. We observe that the vaccine meme data may contain a certain proportion of highly ambiguous samples or suspiciously mislabeled samples, and therefore further explore a noise-aware weighted training method. In the offline stage, this method constructs nearest-neighbor consistency analysis based on image features, text features, and fused features, respectively, and computes the agreement between each sample and the labels of its neighbors, thereby deriving a noise score and conflict count for each sample. Subsequently, instead of directly removing suspected noisy samples, we convert them into different training weights `sample_weight`, so as to reduce the contribution of suspicious samples during training.

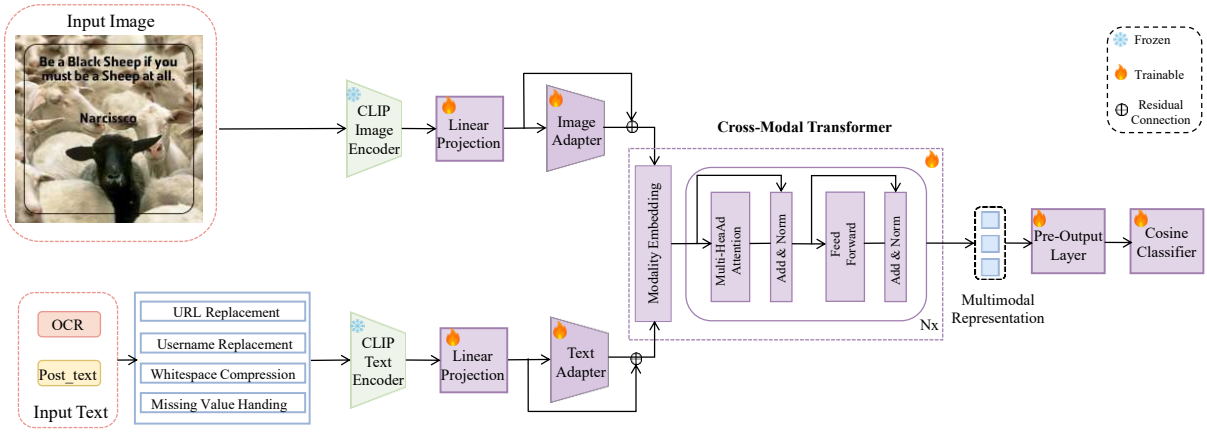


Figure 2: Simplified Architecture of the Adapted MemeCLIP Framework.

LLM-Assisted Multimodal Fusion. We attempt to use large language models to provide additional semantic supplementation. Specifically, we introduce an LLM-assisted multimodal fusion method by adding an auxiliary description branch, `llm_desc`. These descriptions are generated by Qwen2.5-VL-7B-Instruct for memes with poor OCR quality (Bai et al., 2025). Through this selective generation strategy, we aim to examine whether neutral visual supplementary descriptions generated by an LLM can provide additional benefit for vaccine meme stance classification when OCR evidence is insufficient.

During generation, we use prompts that require the model to output neutral and factual supplementary descriptions, focusing on visible persons, objects, actions, symbols, and layout information, while explicitly prohibiting the inference of hidden intent or direct prediction of stance categories. Structurally, we extend the fusion tokens from three to four, namely `[CLS]`, image, text, and description. The final fused representation combines the updated `[CLS]` representation together with three types of interaction terms: image–text, image–description, and text–description. We additionally apply a small-probability dropout to the description branch during training to avoid over-reliance on this auxiliary input.

Three-Branch Multimodal Fusion. We aim to distinguish text information from different sources in a more fine-grained manner. In the standard adapted version, `post_text` and `image_text` are concatenated as a single text input; in this variant, however, we explicitly model them separately, forming a three-branch multimodal fusion architecture: an image branch, a post-text branch, and an OCR-text branch. The motivation for this design is that these two types of text naturally play different semantic roles: `post_text` often serves as contextual supplementation, whereas `image_text` is more likely to provide the most direct stance evidence inside the image.

In this architecture, the image, post text, and OCR text each extract frozen CLIP features, and are then adapted into a shared hidden space through their own projection layers and Adapters(Radford et al., 2021). After residual mixing and normalization, the three branches, together with a learnable `[CLS]` token, are fed into a lightweight cross-modal Transformer. The final fused representation is composed of the updated `[CLS]` representation and the three pairwise interaction terms. Compared with the standard adapted version, this design preserves the branch-specific characteristics of different text sources while modeling their relations through a unified token-level interaction mechanism.

Inference-Stage Enhancement. Beyond architectural optimizations and training-level enhancements, we introduce inference-stage refinement by integrating lightweight test-time augmentation (TTA) with retrieval-augmented prediction. Specifically for TTA, we generate a set of conservative augmented views for each meme image, encompassing the standard CLIP preprocessing pipeline and resizing along the shorter edge followed by center cropping (Radford et al., 2021). To implement retrieval augmentation, we construct a global feature bank comprising all training samples, where the representation of each instance is derived by averaging its fused features extracted across multiple cross-validation folds. During inference, we compute the averaged fused feature for an incoming

test sample and subsequently retrieve its nearest neighbors from the feature bank utilizing cosine similarity. Utilizing the ground-truth labels of the top-k neighbors alongside their temperature-scaled similarity weights, we formulate a k-nearest neighbor (knn) probability distribution, denoted as p_{knn}. Ultimately, the parametric model prediction p_{model} is interpolated with the non-parametric retrieval prior p_{knn} to yield the final probability distribution:

$$p_{\text{final}} = \alpha p_{\text{model}} + (1 - \alpha) p_{\text{knn}}$$

This formulation effectively harnesses both the generalization capabilities inherent in the parametric model and the robust, instance-level evidence provided by non-parametric nearest neighbors.

Final Ensemble System. Our final submission relies on an ensemble framework comprising multiple complementary model variants rather than a single monolithic architecture. Specifically, we integrate models derived from different cross-validation folds alongside variants diversified across visual backbones, robustness optimization, text-source modeling, auxiliary semantics, and inference-stage enhancements. Within this integrated framework, the task-adapted MemeCLIP (Shah et al., 2024) serves as a stable, lightweight baseline, while more advanced backbones extract superior foundational representations. Furthermore, noise-aware training mitigates the interference from ambiguous instances, the three-branch architecture explicitly disentangles heterogeneous text sources, and inference-stage augmentations introduce nearest-neighbor priors to bolster generalization. Consequently, this integrated system exemplifies a highly pragmatic methodology for tasks, maximizing both predictive accuracy and systemic stability by fully exploiting the synergistic complementarity among distinct configurations.

5 Experimental Setup

Data Preprocessing. We implement a standardized preprocessing pipeline across all textual modalities to ensure semantic consistency. Specifically, within the `post_text` and `image_text` fields, we systematically resolve null values and normalize noisy artifacts, including URLs, user mentions, and redundant whitespace. To explicitly encode the absence of specific modalities without losing structural information, missing text entries are substituted with predefined special tokens, namely `[NO_POST]` and `[NO_OCR]`.

Table 2: Results of different model variants on the official shared-task test set, where T stands for the Three-branch Multimodal Fusion method, L stands for the LLM-assisted Multimodal Fusion method, W stands for the Noise-aware Weighted Training method, and I stands for the Inference-stage Enhancement method.

Model	Macro-F1	Accuracy	Precision	Recall
CLIP ViT-L/14	0.8145	0.8166	0.8191	0.8154
EVA-CLIP	0.8170	0.8195	0.8179	0.8182
EVA-CLIP+T	0.8159	0.8185	0.8173	0.8185
EVA-CLIP+L	0.7980	0.8000	0.8035	0.7984
EVA-CLIP+W	0.8239	0.8263	0.8243	0.8256
EVA-CLIP+W+I	0.8355	0.8380	0.8361	0.8377
ENSEMBLE MODEL	**0.8494**	**0.8517**	**0.8494**	**0.8517**

Model Settings and Hyperparameters. Our primary architecture is constructed upon a frozen vision-language backbone. The standard configuration employs CLIP ViT-L/14 (Radford et al., 2021), whereas the more advanced variant utilizes EVA02-L-14, implemented via OpenCLIP (Sun et al., 2023). Across the primary adapted model and most subsequent variants, the core hyperparameters are uniformly set as follows: an input image resolution of 224, a batch size of 16, a residual mixing ratio of 0.2, and a cosine classifier scaling factor of 30. The lightweight cross-modal Transformer comprises 2 layers and 8 attention heads, with dropout rates configured as [0.10, 0.30, 0.20] across its internal components. Optimization is performed using AdamW with a learning rate of 3×10^{-5}, a weight decay of 5×10^{-4}, and a maximum of 10 training epochs. All experiments are executed on a NVIDIA GeForce RTX 3090 GPU.

For the LLM-enhanced variant, we incorporate Qwen2.5-VL-7B-Instruct (Bai et al., 2025) to generate auxiliary textual descriptions. To mitigate the introduction of superfluous noise, this generation process is strictly triggered only for samples exhibiting poor OCR quality. The generation prompt is carefully crafted to elicit neutral and factual visual supplements, explicitly prohibiting the model from inferring hidden intents or directly predicting stance categories. Consequently, this descriptive branch functions exclusively as a supplementary semantic signal rather than a primary decision-making pathway.

Evaluation Metrics. Following the official shared-task setting, we use Macro-F1 as the primary evaluation metric. Macro-F1 computes the F1 score for each class separately and then averages them across classes, thereby assigning equal

Table 3: Hyperparameters used across the main backbone variants.

Parameter	Value
Backbone Variants	CLIP ViT-L/14; EVA02-L-14
Max Token Length	77
Batch Size	16
Max Epochs	10
Optimizer	AdamW
Learning Rate	3×10^{-5}
Weight Decay	5×10^{-4}
Label Smoothing	0.05
Residual Mixing Ratio	0.2
Cosine Classifier Scale	30

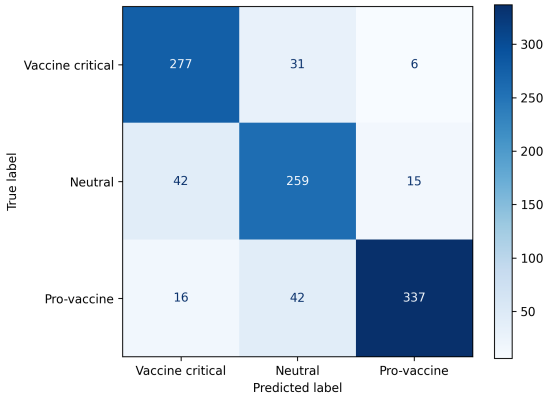


Figure 3: Confusion matrix of the final ensemble system on the official test set.

importance to *Vaccine critical*, *Neutral*, and *Pro-vaccine*. This metric is particularly suitable for the current task because it better reflects balanced classification performance across different stance categories and is less likely to be dominated by relatively easier or more frequent classes.In addition to Macro-F1, we also report Accuracy, Precision, and Recall to provide a more comprehensive view of system behavior. Accuracy reflects the overall proportion of correctly classified samples, while Precision and Recall help us analyze whether the model tends to be overly conservative or overly aggressive for certain classes. Reporting these auxiliary metrics allows us to better understand the trade-offs among overall correctness, class-wise reliability, and class-wise coverage, and also provides additional evidence for interpreting error patterns and comparing model variants.

6 Results and Discussion

6.1 Experimental Results

Our final submission achieved a Macro-F1 score of 0.8494 on the official test set, securing the first place in the shared task (Thapa et al., 2026b). Table 2 summarizes the empirical results of our primary models and their respective variants. The experimental findings indicate that the task-adapted MemeCLIP establishes a robust baseline on the VaxMeme dataset. Building upon this foundation, the integration of a more advanced backbone, noise-aware training, retrieval augmentation, and multi-model ensembling yields substantial performance gains. Notably, noise-aware weighted training contributes significantly to these improvements, suggesting that inherently noisy or highly ambiguous instances profoundly affect training stability within this specific context.

Conversely, the variant incorporating EVA-CLIP+Qwen auxiliary descriptions registers a noticeable performance degradation compared to alternative strategies. This observation implies that while LLM-generated descriptions theoretically enrich the semantic context for samples with poor OCR quality, these supplementary features do not consistently translate into classification improvements under the current experimental setup. This discrepancy may be attributed to residual noise within the generated text, semantic misalignment between the auxiliary descriptions and the original image-text pairs, or the model's suboptimal utilization of the supplementary semantic branch. Ultimately, the superior performance of the ensemble framework over all standalone configurations underscores a strong complementarity among the diverse variants. Techniques such as noise-aware training and text-source disentanglement prove to be non-redundant; rather, they synergistically provide comprehensive and robust evidence for decision-making during the ensembling phase.

6.2 Further Analysis

To further elucidate the limitations of our system, we conduct a detailed error analysis based on the confusion matrix. The most prominent failure mode involves the misclassification between the Neutral category and stance-bearing classes. This phenomenon suggests that the model is prone to over-inferring stances when processing memes characterized by purely factual content, weak opinion signals, or insufficient background context. Furthermore, instances degraded by low-quality OCR extraction, severe textual noise, or intricate visual layouts exhibit a significantly higher susceptibility to misclassification. Another intrinsic challenge arises from samples with inherently ambiguous

class boundaries, such as memes that ostensibly share objective information but implicitly convey supportive or critical nuances within specific socio-cultural contexts. Collectively, these error patterns reveal that while the proposed system effectively integrates multimodal signals, it retains pronounced limitations in handling weak-evidence scenarios, modeling implicit pragmatics, and resolving fine-grained semantic boundaries.

7 Conclusion

In this paper, we describe our proposed approach for the Shared Task on Multimodal Vaccine Critical Meme Detection (VaxMeme) @EEUCA 2026. Using MemeCLIP as the foundational framework, we systematically enhance the model through task-specific adaptation, noise-aware training, auxiliary semantic injection, text-source disentanglement, and inference-stage optimization. Our final system, constructed by ensembling multiple complementary model variants, achieves highly competitive performance on the official shared-task test set. Experimental results demonstrate the effectiveness of noise-aware training and late-stage ensembling strategies for robust vaccine meme stance identification. Future work will focus on developing more robust architectures and exploring interpretable multimodal reasoning mechanisms for complex meme analysis.

8 Limitations

Despite integrating multiple complementary variants, the system inherently operates within a supervised classification paradigm and lacks explicit modeling of nuanced socio-cultural contexts. Furthermore, while the LLM-generated auxiliary descriptions are designed for neutral and selective utilization, they may occasionally introduce semantic noise or over-interpretation. The retrieval-augmented module is also constrained by its reliance on semantically analogous instances in the training corpus, rendering it less robust when processing rare or out-of-distribution memes. Moreover, the current findings are primarily validated on the VaxMeme benchmark, and their generalizability to alternative public health domains, diverse social media platforms, or cross-cultural scenarios requires further empirical verification. Achieving optimal performance necessitates model ensembling, a strategy that inevitably introduces higher inference complexity and deployment costs.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL technical report. *Preprint*, arXiv:2502.13923.

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.

Mingxuan Chen, Xinqiao Chu, and K. P. Subbalakshmi. 2021. MMCoVaR: Multimodal COVID-19 vaccine focused data repository for fake news detection and a baseline architecture for classification. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 31–38.

Zhi-Hao Guan, Qing-Yuan Jiang, and Yang Yang. 2025. Balance-aware sequence sampling makes multi-modal learning better. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 2842–2850.

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)*.

Qing-Yuan Jiang, Zhouyang Chi, and Yang Yang. 2025. Interactive multimodal learning via flat gradient modification. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 5489–5497. ijcai.org.

Bin Liang, Ang Li, Jingqian Zhao, Lin Gui, Min Yang, Yue Yu, Kam-Fai Wong, and Ruifeng Xu. 2024. Multi-modal stance detection: New datasets and model. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12373–12387, Bangkok, Thailand. Association for Computational Linguistics.

Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G. Dunn. 2023. A multimodal framework for the identification of vaccine critical memes on twitter. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 706–714.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. MemeCLIP: Leveraging CLIP representations for multimodal meme classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17320–17332, Miami, Florida, USA. Association for Computational Linguistics.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. EVA-CLIP: Improved training techniques for CLIP at scale. *arXiv preprint arXiv:2303.15389*.

Laxmi Thapa, Aryaman Jain, Lakshmojee Koduru, Surabhi Adhikari, Junaid Rashid, Jungeun Kim, Surendrabikram Thapa, and Usman Naseem. 2026a. Concept-grounded detection of vaccine misinformation in multimodal content using interpretable vision-language models. In *Companion Proceedings of the ACM Web Conference 2026*.

Surendrabikram Thapa, Siddhant Bikram Shah, Kritesh Rauniyar, Shuvam Shiwakoti, Surabhi Adhikari, Hariram Veeramani, Kristina T. Johnson, Ali Hurriyetoglu, Hristo Tanev, and Usman Naseem. 2025. Multimodal hate, humor, and stance event detection in marginalized sociopolitical movements. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts*, pages 20–31, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Kritesh Rauniyar, Laxmi Thapa, Surabhi Adhikari, Kristina T. Johnson, Ali Hürriyetoğlu, Hristo Tanev, and Usman Naseem. 2026b. Multimodal identification of vaccine content stance on social media. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. 2024. Facilitating multimodal classification via dynamically learning modality gap. In *Advances in Neural Information Processing Systems*, volume 37, pages 62108–62122. Curran Associates, Inc.

Feng Yu, Xiangyu Wu, Yang Yang, and Jianfeng Lu. 2026. Multimodal classification via total correlation maximization. In *The Fourteenth International Conference on Learning Representations*.

LINUS@EEUCA 2026: Fine-grained Toxicity Detection in Gaming Chat using Multilingual Transformers

Prajwal Ghimire
Kathmandu University
prajwalghimire22@gmail.com

Aashish Mahato
Kathmandu University
aashishmtho24@gmail.com

Sunil Regmi
Kathmandu University
sunil.regmi@ku.edu.np

Abstract

The detection of toxic behavior in online gaming communities is crucial for maintaining safe digital spaces, yet remains challenging due to subtle context-dependent and intent-driven language. The GameTox dataset consists of around 53K World of Tanks chat utterances annotated across six categories: Non-toxic, Insults and Flaming, Other Offensive Texts, Hate and Harassment, Threats, and Extremism (Naseem et al., 2025). Our best performing approach, across multiple transformer-based architecture experimentations, is based on the multilingual BERT variant mmBERT-base fine-tuned with class-weighted cross-entropy loss. The best mmBERT-base model achieved a Macro F1 of 0.5882 during validation and an official test Macro F1 of 0.5104 on the shared task leaderboard. An internal held-out evaluation on a development split yielded 0.4282, which we analyze to understand distributional sensitivity to gaming slang and class imbalance. The code is available at: https://github.com/sunilRegmi-ai/eeuca-toxicity-detection.

1 Introduction

The global gaming industry has reached an unprecedented scale. Prior research has explored contextual bandit algorithms to balance exploration and exploitation, optimizing costly real-time toxicity monitoring resources in multiplayer environments where no pre-existing predictive models are available (Morrier et al., 2025). Although chat-enabled titles foster social engagement, they also facilitate toxic behavior and cyberbullying, particularly in team-based competitive environments (Kwak et al., 2015; Naseem et al., 2025).

The detection of hate speech remains a complex societal challenge, with various machine learning and natural language processing architectures being developed to classify diverse forms of online toxicity ranging from religious hate to sexism (Parihar et al., 2021). Early rule-based approaches relied on lexicon and pattern-based annotation systems, which covered only about 16% of distinct vocabulary but over 60% of actual word usage, reflecting the repetitive nature of in-game chat(Märtens et al., 2015). However, such approaches remain limited handling the adversarial nature of in-game slang, where players frequently use creative obfuscation, special character substitutions, and evolving abbreviations to circumvent moderation, prompting the development of robust character-level CNN and hybrid transformer models (Lee et al., 2025). Content moderation systems of this kind require responsible usage, as LLMs can inherit and amplify biases present in training data, necessitating community engagement and bias mitigation to ensure fairness in public discourse and policy-making (Thapa et al., 2025b). Thus, modern dataset creation is actively recognizing the critical role of the author's behavioral intent rather than just surface level lexical features to align automated moderation with actual community policies (Wang et al., 2024). These methodologies are highly relevant to the goals of ToxIntent@EEUCA 2026, which aims to understand toxic behavior in gaming communities to promote healthier digital spaces (Thapa et al., 2026; Hürriyetoğlu et al., 2026).

The GameTox dataset (Naseem et al., 2025) provides around 53K chat utterances from *World of Tanks* annotated across six fine-grained categories: Non-toxic (0), Insults and Flaming (1), Other Offensive Texts (2), Hate and Harassment (3), Threats (4), and Extremism (5). The annotation schema used for this task shares its foundational principles with the multimodal CrisisHateMM dataset (Bhandari et al., 2023), which highlights the importance of distinguishing between directed and undirected toxic intent. Our approach is consistent with findings regarding the reliability of models such as FastText and BERT while BERT's contextual embeddings are better suited to identifying and distinguishing complex targets, such as indi-

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 216–222

viduals, organizations, and communities whereas FastText was found to be exceptional for language identification and was less reliable for the more nuanced task of target identification (Acharya et al., 2025). Prior work on fine-grained target classification in hate speech detection demonstrates that contextual embeddings support models to distinguish between nuanced category types, including individual, community, and organizational targets across Devanagari-script languages (Thapa et al., 2025a).

The monolingual pre-training on massive domain-specific corpora has shown superior results in capturing rich semantics and grammatical structures for specific low-resource scripts like Nepali (Timilsina et al., 2022; Maskey et al., 2022), yet we opt to leverage multilingual mmBERT variants to ensure robust feature representation across diverse gaming communities, building on the proven workings of domain adapted transformers retrained on abusive corpora (Caselli et al., 2021).

However, mmBERT variants remain vulnerable to the linguistic volatility of gaming slang. Furthermore, reliance on machine-translated non-English samples introduces semantic misalignment and translation noise, which can degrade classification performance, particularly in informal contexts (Lamin and Aziz, 2025). Therefore, existing approaches struggle to generalize across rare toxicity categories under severe class imbalance and distributional shift, where intent is subtle and context-dependent.

Our participation in ToxIntent@EEUCA 2026 secured a rank of 18th out of 35 competing teams. As a shared-task system description, our primary contributions are a systematic benchmark of five multilingual transformer encoders on a severely class-imbalanced gaming-domain dataset, and an error analysis demonstrating that class-weighted loss alone is insufficient for extremely rare toxicity categories. Our findings reveal that mmBERT-base, optimized with balanced class-weighted cross-entropy loss, substantially outperforms other multilingual encoder baselines on validation and achieves an official test Macro F1 of 0.5104. An internal dev-split evaluation of 0.4282 further reveals sensitivity to distributional shift in rare toxicity categories such as *Extremism* and *Threats*, where players actively employ irony, sarcasm, and localized slang to disguise toxic intent. The code is available at :`https://github.com/sunilRegmi-ai/eeuca-toxicity-detection`.

2 Background

Modern Natural Language Understanding (NLU) is grounded in deep bidirectional pre-training, allowing models to jointly condition on both left and right context in all layers (Devlin et al., 2019). Slot-gated architectures have proven reliable for mapping specific relationships between intent categories and semantic slots by utilizing a slot gate that optimizes the global relationship between intent and slot attention vectors (Goo et al., 2018). This paradigm has been successfully extended via joint fine-tuning strategies to unify classification and slot filling using a shared representation (Chen et al., 2019). Domain-adapted models like HateBERT utilize retraining on curated datasets of banned communities to improve performance on out-of-distribution toxic text to improve the detection of rare and highly offensive categories (Caselli et al., 2021). Frameworks such as ToXCL unify detection with explanation generation using knowledge distillation from a teacher classifier to mitigate error propagation and provide transparency in the automated moderation of implicit hate (Hoang et al., 2024).

The impact of Large Language Models (LLMs) has marked a new era in Computational Social Science (CSS), offering the capacity to interpret human communication nuances and patterns in ideological shifts (Thapa et al., 2025b). However, substantial generalization gaps often remain, as existing top-tier models continue to struggle with the added complexity of diverse NLU benchmarks spanning inference, similarity, and masked evaluation (Nyachhyon et al., 2025). Research on low-resource scripts reveals that dedicated, fine-tuned masked language models (like NepBERTa) frequently outperform generalized LLMs on sequence tagging tasks such as Named-Entity Recognition (NER) and POS tagging (Subedi et al., 2024).

A strategic research on informal digital discourse further shows that multi-aspect annotation schemes uncover nuanced layers of intent such as profanity, violence, feedback and sarcasm overlooked by basic binary systems (Singh et al., 2020). This is particularly evident in anti-establishment discourse and election-related text, where multi-target classification (e.g., individual vs. community) is crucial for understanding the propagation of hate (Rauniyar et al., 2023; Thapa et al., 2023). Bidirectional LSTM modeling paired with appropriate word embeddings remains highly effective for separating

offensiveness and profanity into distinct classification tasks, even amid the noise of informal social media text for specific behaviors like profanity detection (Adhikari et al., 2024).

The necessity of specialized encoder strategies to manage domain-specific linguistic complexity is underscored in applications such as legal machine translation, where custom-built parallel corpora are required to ground encoder-decoder architectures against data sparsity (Poudel et al., 2024). This principle informs our architectural adaptation for the slang-heavy gaming domain. Cross-lingual challenges including semantic misalignment and translation noise remain pertinent when modeling the multilingual roots of global subcultures (Lamin and Aziz, 2025). Finally, empirical evaluations confirm that fine-tuned lightweight transformers (like DistilBERT) continue to provide optimal accuracy-cost trade-offs in continuous gaming moderation when compared to the computational expense of large generative LLMs or Retrieval-Augmented Generation (RAG) pipelines (Tereshchenko and Hämäläinen, 2025).

3 Dataset and Task

The Shared Task on Understanding Toxic Behavioral Intent in Gaming Chat Logs utilizes the GameTox dataset (Naseem et al., 2025). The dataset comprises around 53K chat utterances sourced from the multiplayer online game *World of Tanks*. The primary objective of this task is to capture the complex relationship between user intent and linguistic features across a fine-grained classification task across six categories.

Class ID	Category	Instances
0	Non-toxic	43,497
1	Insults and Flaming	7,407
2	Other Offensive Texts	2,343
3	Hate and Harassment	349
4	Threats	75
5	Extremism	30

Table 1: Distribution of the dataset

The vast majority of utterances fall into the Non-toxic class, while severe toxicity categories, such as Threats and Extremism, are extremely rare that shows severe class imbalance. The gaming specific slang, obfuscation techniques, and informal syntax, also creates a highly noisy and adversarial text environment. The official evaluation metric for the shared task is the Macro F1-score, which weighs the performance across all classes equally, heavily penalizing models that overfit to the majority class.

4 Methodology

Our system approaches the toxicity classification task with multiple experimental tracks explained below.

4.1 Model Architectures and Training Setup

We systematically benchmarked several modeling paradigms. In the initial phase, we evaluated a broad set of multilingual encoder models — ToxicXLM-RoBERTa, XLM-RoBERTa, m-DistilBERT, and m-BERT — using the Hugging Face Trainer API (Wolf et al., 2020). The raw chat utterances are loaded from CSV files and merged on a common index column, with the text field standardized and labels cast to integers across all splits. Each input is tokenized using the model's native tokenizer with truncation and padding to a maximum sequence length, producing fixed-length token sequences fed directly into the classification head. Class weights are computed from the training label distribution using scikit-learn's `compute_class_weight` and passed to a customized `WeightedTrainer` subclass that overrides the default cross-entropy loss to address class imbalance. All non-DeBERTa models are trained with FP16 mixed precision when CUDA is available. The implementation relies on the Hugging Face `transformers` and `datasets` libraries, with `pytorch` as the framework, `scikit-learn` for class weight computation, and `accelerate` for distributed training support. The full implementation details and reproducibility instructions are available at: `https://github.com/sunilRegmi-ai/eeuca-toxicity-detection`.

The experimental logs in the subsequent phase identified mmBERT variants as the superior architecture due to its pre-training on massively multilingual social media and informal web corpora (Marone et al., 2025). The mmBERT-base variant was fine-tuned using the Hugging Face Trainer API with a customized WeightedTrainer applies class-weighted cross-entropy loss to address class imbalance. An initial run with a learning rate of 3e-06, batch size of 32, and sequence length of 64 yielded a validation Macro F1 of 0.4933. A subsequent run with an increased learning rate of 1e-05, batch size of 64, and reduced sequence length of 32 produced a substantial improvement and achieved a valida-

tion Macro F1 of 0.5882 — a gain of ∼0.09 that underscores the sensitivity of mmBERT to learning rate and batch size scaling. Both runs used weight decay of 0.01 and early stopping with a patience of 3 epochs, with the best checkpoint selected based on validation Macro F1.

5 Results and Discussion

The evaluation metrics for our primary transformer benchmarking experiments are listed in Table 2.

Model	Val F1	Test F1	Test Acc
Toxic-XLM-RoBERTa	0.3558	0.3520	0.8281
XLM-RoBERTa	0.3830	0.3839	0.8130
m-DistilBERT	0.3907	0.3578	0.7942
m-BERT	0.4146	0.4243	0.8249
mmBERT-base	**0.5882**	**0.5104**	**0.8716**

Table 2: Experimental Results on Validation and Test Sets

These results, reported on the official validation and test splits, identify mmBERT as the most reliable base architecture for this specific dataset. The standard generalized models and domain-specific toxicity variants like Toxic-XLM-RoBERTa struggled to surpass the ∼0.42 Macro F1 barrier. Our best mmBERT-base model achieved a validation Macro F1 of 0.5882 and an official leaderboard Test Macro F1 of **0.5104** securing a rank of 18th out of 35 participating teams. An additional internal evaluation on a held-out portion of the development set yielded a Test Macro F1 of 0.4282; this lower figure reflects sensitivity to the specific distributional characteristics of that split rather than the true held-out test performance. The ablation between the two mmBERT-base runs further confirms that scaling the learning rate from 3e-06 to 1e-05 and the batch size from 32 to 64 were the primary determinants of the ∼0.09 validation F1 improvement. The two runs also differ in sequence length (64 vs. 32), so this hyperparameter was co-varied and cannot be isolated from the ablation alone. However, both the token length distribution of the training split and the nature of the dataset provide strong empirical justification: the 75th, 90th, and 95th percentiles of whitespace-tokenized utterance lengths are 4, 6, and 8 tokens respectively, with a mean of 3.02 and a maximum of 30 tokens, confirming that a `max_length` of 32 provides complete coverage for all training utterances. This is consistent with the source domain: Naseem et al. (2025) collected utterances from World of Tanks real-time in-game chat, where annotation guideline examples reach a maximum of 6 whitespace tokens, and the most discriminative toxic and game-slang tokens identified in their dataset are predominantly single-word items (e.g., *die*, *cancer*, *kill*), confirming that toxicity signal in this domain is lexically concentrated rather than requiring long contextual spans.

5.1 Error Analysis

Our official test Macro F1 of 0.5104 reflects a moderate ∼0.08 gap from the validation score of 0.5882. A wider gap of ∼0.16 was observed on an internal development split evaluation. This generalization gap underscores the core difficulty of this shared task: extreme out-of-distribution linguistic volatility.

Error analysis indicates that the minority classes (*Extremism* and *Threats*), comprising only 30 and 75 instances respectively out of 53,701 total, suffer heavily from false negatives. The extreme scarcity of these categories means the model has insufficient exposure to their linguistic patterns during training despite balanced class weighting. Players frequently disguise severe intent using irony, sarcasm, inside jokes, and highly localized slang that models struggle to interpret without broader context. The distributional shift between the validation and unseen test interactions further compounds this, as gaming communities continuously evolve new obfuscation strategies that fall outside the training distribution. Future iterations should explore few-shot augmentation strategies for rare categories and more context-aware architectural solutions to address active linguistic obfuscation.

6 Conclusion

This paper presents our system submission to ToxIntent@EEUCA 2026, a shared task on fine-grained toxicity detection in gaming chat logs using the GameTox dataset (Naseem et al., 2025). As a shared-task system description, our primary contributions are a systematic benchmark of five multilingual transformer encoders on a severely class-imbalanced gaming-domain dataset, and an error analysis demonstrating that class-weighted loss alone is insufficient for extremely rare toxicity categories. We identified mmBERT-base (Marone et al., 2025) as the most effective architecture for this task, owing to its pre-training on massively multilingual social media and informal web corpora. Our

best mmBERT-base model fine-tuned with class-weighted cross-entropy loss and optimized hyperparameters resulted in validation Macro F1 of 0.5882 and an official test Macro F1 of 0.5104. Our ablation analysis further demonstrates that learning rate and batch size scaling are critical determinants of performance gain for mmBERT on this task. The substantial ~ 0.08 F1 generalization gap between validation and test environments underscores the inherent difficulty of detecting fine-grained toxic intent in adversarial, slang-heavy gaming discourse. Future work should explore few-shot augmentation for rare toxicity categories, context-aware architectures, and ensemble strategies to bridge this generalization gap.

Limitations

The severe class imbalance in the GameTox dataset (Naseem et al., 2025), particularly for *Extremism* (30 instances) and *Threats* (75 instances), limits the model's exposure to rare toxicity patterns, and balanced class weighting alone is insufficient to fully compensate for this scarcity. Our evaluation is limited to the GameTox dataset sourced from a single game (*World of Tanks*), and generalization to other gaming communities or platforms remains unverified. The absence of explicit data augmentation or few-shot strategies for minority classes is a known weakness of our current pipeline. Finally, the distributional shift between validation and test interactions suggests that our models are sensitive to evolving gaming slang and obfuscation strategies that fall outside the training distribution, a challenge that requires more robust cross-domain generalization techniques in future iterations.

Acknowledgments

We thank the organizers of ToxIntent@EEUCA 2026 (Thapa et al., 2026; Hürriyetoğlu et al., 2026) for providing the datasets and their support throughout this research.

References

Darwin Acharya, Sundeep Dawadi, Shivram Saud, and Sunil Regmi. 2025. Paramananda@NLU of Devanagari script languages 2025: Detection of language, hate speech and targets using FastText and BERT. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 334–338.

Abiral Adhikari, Prashant Manandhar, Reewaj Khanal, Samir Wagle, Praveen Acharya, and Bal Krishna Bal. 2024. Profanity and offensiveness detection in Nepali language using bi-directional LSTM models. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 515–521.

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 Mitrovi'c, 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.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *Preprint*, arXiv:1902.10909.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, and 1 others. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the NAACL: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757.

Nhat M. Hoang, Xuan Long Do, Duc Anh Do, Duc Anh Vu, and Luu Anh Tuan. 2024. ToXCL: A unified framework for toxic speech detection and explanation. In *Proceedings of the 2024 Conference of the ACL: Human Language Technologies (Volume 1: Long Papers)*, pages 6460–6472.

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

Haewoon Kwak, Jeremy Blackburn, and Seungyeop Han. 2015. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*, page 3739–3748. Association for Computing Machinery.

Nor Zakiah Lamin and Azwa Abdul Aziz. 2025. Cross-lingual sentiment analysis in low-resource languages: A recent review on tasks, methods and challenges. *International Journal of Advanced Computer Science and Applications*.

Jaehong Lee, Pavinee Rerkjirattikal, and Sanggyu Nam. 2025. Toxic chat detection in online games using

hybrid bert and character-level cnn. In *MakeLearn, TIIM & PIConf 2025: Accelerated Innovation (AI); Sustainability for Better Humanity*. ToKnowPress.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning. *Preprint*, arXiv:2509.06888.

Utsav Maskey, Manish Bhatta, Shiva Bhatt, Sanket Dhungel, and Bal Krishna Bal. 2022. Nepali encoder transformers: An analysis of auto encoding transformer language models for Nepali text classification. In *Proceedings of the 1st Annual Meeting of the SIGUL*, pages 106–111.

Jacob Morrier, Rafal Kocielnik, and R. Michael Alvarez. 2025. Bandit algorithms for efficient toxicity detection in competitive online video games. *IEEE Access*, 13:103109–103117.

Marcus Märtens, Siqi Shen, Alexandru Iosup, and Fernando Kuipers. 2015. Toxicity detection in multiplayer online games. In *2015 International Workshop on Network and Systems Support for Games (NetGames)*, pages 1–6.

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 NAACL: Human Language Technologies (Volume 2: Short Papers)*, pages 440–447.

Jinu Nyachhyon, Mridul Sharma, Prajwal Thapa, and Bal Krishna Bal. 2025. Consolidating and developing benchmarking datasets for the Nepali natural language understanding tasks. In *Proceedings of the 14th IJCNLP and the 4th AACL*, pages 1906–1925.

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.

Shabdapurush Poudel, Bal Krishna Bal, and Praveen Acharya. 2024. Bidirectional English-Nepali machine translation(MT) system for legal domain. In *Proceedings of the 3rd Annual Meeting of the SIGUL @ LREC-COLING 2024*, pages 53–58.

Kritesh Rauniyar, Sweta Poudel, Shuvam Shiwakoti, Surendrabikram Thapa, and 1 others. 2023. Multi-aspect annotation and analysis of nepali tweets on anti-establishment election discourse. *IEEE Access*, 11:143092–143115.

Oyesh Mann Singh, Sandesh Timilsina, Bal Krishna Bal, and Anupam Joshi. 2020. Aspect based abusive sentiment detection in nepali social media texts. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 301–308.

Bipesh Subedi, Sunil Regmi, Bal Krishna Bal, and Praveen Acharya. 2024. Exploring the potential of large language models (LLMs) for low-resource languages: A study on NER and POS tagging for Nepali language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics (LREC-COLING 2024)*, pages 6974–6979.

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*.

Surendrabikram Thapa, Rauniyar Kritesh, Shiwakoti Shuvam, and 1 others. 2023. Nehate: Large-scale annotated data shedding light on hate speech in nepali local election discourse. In *ECAI 2023*, pages 2346–2353. IOS Press.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Surabhi Adhikari, Kengatharaiyer Sarveswaran, Bal Krishna Bal, Hariram Veeramani, and Usman Naseem. 2025a. Natural language understanding of Devanagari script languages: Language identification, hate speech and its target detection. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 71–82.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025b. 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)*.

Sulav Timilsina, Milan Gautam, and Binod Bhattarai. 2022. NepBERTa: Nepali language model trained in a large corpus. In *Proceedings of the 2nd AACL and 12th IJCNLP (Volume 2: Short Papers)*, pages 273–284.

Xinyu Wang, Sai Koneru, Pranav Narayanan Venkit, and 1 others. 2024. The unappreciated role of intent in algorithmic moderation of social media content. *Preprint*, arXiv:2405.11030.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 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, Online. Association for Computational Linguistics.

Linus@EEUCA 2026: Multimodal and Text-Only Approaches to Vaccine-Critical Meme Detection.

Darwin Acharya
Kathmandu University
acharyadarwin5@gmail.com

Shiv Ram Saud
Kathmandu University
saudshivram373@gmail.com

Sunil Regmi
Kathmandu University
sunil.regmi@ku.edu.np

Abstract

In this paper, we describe our participation in the Shared Task on Multimodal Identification of Vaccine Critical Content on Social Media (VaxMeme) of EEUCA 2026, a satellite of ACL 2026. We tackle the classification of Twitter-based vaccine memes into anti-vaccine, neutral, and pro-vaccine categories using the VaxMeme dataset with 8,195 train, 1,024 val, and 1,025 test samples. We experiment with two different architecture families: (i) Multimodal hybrids: CLIP ViT-B/32 for images + BERT-based models for texts (BERT-base-uncased, ModernBERT) with late fusion strategy based on concatenation of L2-normalized feature vectors and (ii) Text-only: pre-trained models for texts (BERT-base-uncased, RoBERTa-base, ModernBERT-base, DistilBERT-base, Deberta-v3-base) for post_text. In both cases, we use a three-layer feed-forward network with GELU activation for classification. We use class-weighted cross-entropy loss, differential learning rates, AdamW optimizer, gradient accumulation, OneCycleLR scheduler, and early stopping on the val set for optimization. Data augmentation is applied for the multimodal CLIP-based approach only. The winning approach among those tested is the text-only BERT-base-uncased with a macro-F1 of 0.8102 which is ahead of the performance of the CLIP + BERT-base hybrid model, which achieves a test macro-F1 of 0.7603.

1 Introduction

The rapid spread of health misinformation online poses significant challenges to public health, potentially leading to confusion, undermining trust in health authorities, and hindering effective health interventions (Thapa et al., 2024). The internet meme, defined by its concise and visually salient nature and its reliance on a combination of image and text for information and meaning, has risen to become a powerful and spreading vehicle for vaccine-critical information. Unlike plain text-based misinformation, which can often be addressed using conventional natural language processing techniques, the use of humor and irony in such memes makes them highly engaging and significantly more challenging to detect and mitigate automatically. The EEUCA 2026 shared task (Thapa et al., 2026b; Hürriyetoğlu et al., 2026), which is a satellite event of ACL 2026, involves the classification of vaccination-related memes into three types: anti-vaccine, neutral, and pro-vaccine. The evaluation is done based on macro-averaged F1 score. The task is based on the VaxMeme benchmark dataset (Naseem et al., 2023). Our task is part of the main track.

Our main strategy is based on the development of a classification framework that is modular and enables a comparison of unimodal and multimodal settings in a controlled and identical manner. Two different approaches are implemented: the first approach is a multimodal hybrid pipeline which is based on a late fusion multimodal approach, where visual and textual information are first encoded with separate transformer-based models, which are then concatenated for a classification task. We are using the ViT-B/32 encoder from clip (Radford et al., 2021) for the visual modality, which was trained with contrastive learning on image-text pairs and has shown excellent performance in multimodal reasoning tasks. For the textual modality, we experimented with two different encoders: BERT-base-uncased (Devlin et al., 2019), and ModernBERT (Warner et al., 2025) fusing the L2-normalized representations through late concatenation. The second part of the pipeline is a text-only pipeline, which consists of the fine-tuning of five pre-trained language models, namely BERT-base-uncased (Devlin et al., 2019), RoBERTa-base (Liu et al., 2019), ModernBERT-base (Warner et al., 2025), DistilBERT-base (Sanh et al., 2020), and DeBERTa-v3-base (He et al., 2023)—using only the post_text metadata field. Both pipelines utilize

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 223–232

the identical feedforward classification head, along with the identical training approach, thus making it possible to compare them directly.

Carrying out this task has led to some interesting findings. Firstly, the best-performing model, BERT-base-uncased with text-only input, achieves a test macro-F1 of 0.8102, outperforming the CLIP + BERT-base hybrid model, achieving 0.7603. This is unexpected, suggesting that the post_text tweet feature is a highly discriminative feature for vaccine stance classification, and CLIP image features, although individually useful (CLIP-only: 0.7189), do not seem to bring consistent improvements over powerful text encoders when used with late concatenation. Secondly, the RoBERTa-base model, when used with text-only input, achieves a competitive test macro-F1 of 0.8091, showing how effective the post_text feature is across models. Finally, from error analysis, it is clear that, across all models, sarcasm and implicit culture remain the main failure cases, especially when the intended stance depends on the nuanced interaction between the images and the text. Our top-performing model ranked 14th on the official CodaBench leaderboard with a macro-F1 score of 0.81. The code is available at: `https://github.com/sunilRegmi-ai/VCC-Social-Media`.

2 Background

The VaxMeme Shared Task (Thapa et al., 2026b; Hürriyetoğlu et al., 2026) is a three-class classification task that attempts to predict the stance expressed in a meme. Each data point is composed of a meme image in PNG format and post_text metadata corresponding to the tweet. The task is to predict one of three classes: 0 for anti-vaccine, 1 for neutral, and 2 for pro-vaccine. For example, a meme showing a syringe with the caption "CCP Virus Variant Affects Vaccinated People More Than Unvaccinated People: Study" is classified as anti-vaccine (class 0), while a meme showing a healthcare worker holding a vaccination record is classified as pro-vaccine (class 2). Memes that depict vaccination without expressing any opinion are classified as neutral. Evaluation is done using macro F1 score which assigns equal importance to each class.

The dataset used from VaxMeme consists of 10,244 English-language memes which are labeled as 0, 1, 2. The dataset has been split into official training, evaluation, and test sets. The training set comprises 8,195 samples, whereas the evaluation set has been used to provide a held-out validation partition.

This work was inspired by the extensive amount of prior work in the area of multimodal harmful content detection. For instance, the Hateful Memes Challenge (Kiela et al., 2021) highlighted the importance of multimodal reasoning in meme classification tasks and showed that it was possible to achieve state-of-the-art results by utilizing a model that excelled in each modality individually, albeit at a much lower level of performance than humans. Most recently, (Bhandari et al., 2023) proposed a multimodal dataset, CrisisHateMM, for detecting hate speech from text-embedded conflict images. This demonstrates the generalization of vision-linguistic approaches for detecting harmful content. In the context of vaccine misinformation, (Pramanick et al., 2021) presented MOMENTA, a framework that makes use of both global image and text features, as well as local entity-level representations and object attributes, in order to effectively capture the fine-grained semantic information that is necessary in meme analysis tasks. (Hayawi et al., 2022) also presented ANTi-Vax, a text-based dataset that targets vaccine misinformation related to COVID-19. In a further extension of the above studies, (Naseem et al., 2023) presented the VaxMeme dataset, which targets the multimodal setting of vaccine misinformation analysis. At the same time, (Thapa et al., 2026a) also proposed concept-grounded vision-language models for interpretable vaccine misinformation detection, offering an alternative approach to the classification-centric models used in this shared task. Unlike the aforementioned works, our work differs in the following aspects: (i) comparative evaluation of five encoders in a unified text-only setting, (ii) direct comparison of the proposed approaches with the multimodal CLIP fusion variants in the same setting, and (iii) the evaluation of the proposed approaches on recently proposed state-of-the-art models such as ModernBERT (Warner et al., 2025) and DeBERTa-v3 (He et al., 2023), which have not been previously evaluated on this benchmark.

3 System Overview

3.1 Fusion and Classification Head

All systems both unimodal and multimodal, use a classification head that is the same and acts on the

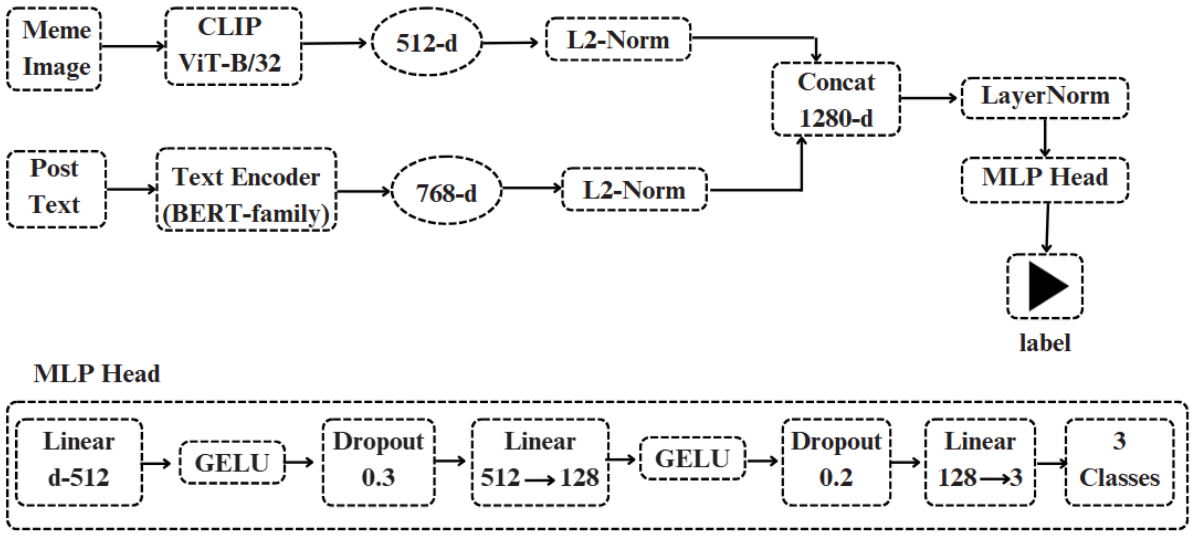


Figure 1: Architecture diagram showing hybrid system with MLP Head

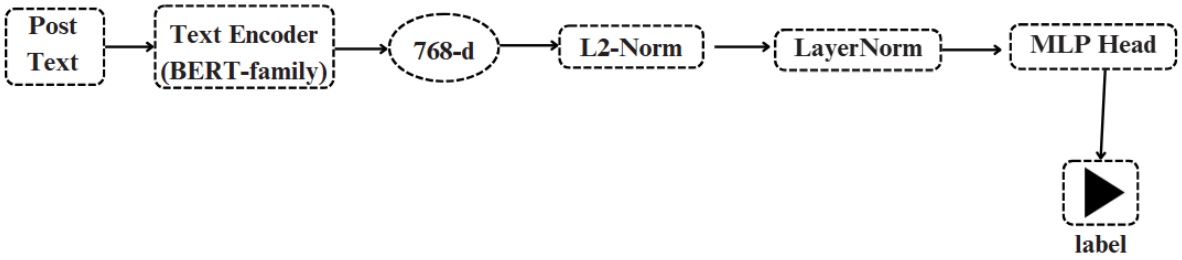


Figure 2: Architecture diagram showing text-only systems.

last feature representation. Let $z \in \mathbb{R}^d$ denote the input to the classification head.

In the multimodal (hybrid) setting, the L2-normalized visual feature vector $v \in \mathbb{R}^{512}$ and textual feature vector $t \in \mathbb{R}^{768}$ are concatenated to form a joint representation:

$$z = [v; t] \in \mathbb{R}^{1280}. \quad (1)$$

In the unimodal settings, z corresponds directly to the modality-specific representation:

$$z = \begin{cases} v \in \mathbb{R}^{512}, & \text{(image-only)} \\ t \in \mathbb{R}^{768}, & \text{(text-only)} \end{cases} \quad (2)$$

Thus, the input dimension d varies depending on the modality: $d = 512$ (image-only), $d = 768$ (text-only), and $d = 1280$ (hybrid).

Before classification, Layer Normalization is applied to z. The classification head is implemented as a three-layer feedforward network with GELU activations and dropout regularization:

$$\text{Linear}(d \to 512) \to \text{GELU} \to \text{Dropout}(0.3)$$
$$\text{Linear}(512 \to 128) \to \text{GELU} \to \text{Dropout}(0.2)$$
$$\text{Linear}(128 \to 3)$$

The last layer of the network generates logits for the three classes of stances. The classification head is randomly initialized and has a learning rate higher than the one used for training the pre-trained backbone encoders.

3.2 Multimodal Hybrid Systems (CLIP + Text Encoder)

The same late fusion multimodal framework is used by our hybrid systems. A visual encoder and a BERT-family textual encoder (discussed in Section 3.3) are used to get fixed-length feature vectors for the memes. The modality flag is used to specify the active encoders (image only, text only, or hybrid) and to specify if the normalized feature vectors need to be concatenated.

3.2.1 Visual Encoder: CLIP ViT-B/32

For all hybrid experiments, we use the visual subnetwork of CLIP ViT-B/32 (Radford et al., 2021) as the image encoder. CLIP is pre-trained via contrastive learning on 400 million web-scraped image-text pairs. This model is particularly beneficial for meme classification because the semantic meaning

of visual elements is often defined in relation to text.

In this work, we used the model via the OpenAI CLIP library, loading the pre-trained weights for ViT-B/32 with clip.load("ViT-B/32", jit=False) .

Each image is preprocessed through the default CLIP pipeline (resizing to 224x224 and normalizing through CLIP channel statistics). The visual encoder then produces a 512-dimensional vector. During training time, the images also go through a stochastic augmentation pipeline (as described in Section 3.5) before CLIP preprocessing. This model is cast to float32. During training, the whole visual encoder is fine-tuned end-to-end with a backbone learning rate of 2×10^{-6}. In the case of ModernBERT systems, a learning rate of 3×10^{-6} is used.

3.3 Text Encoders

We experiment with five text encoders from BERT-family to assess the effect of text model choice on multimodal meme and text-only systems classification performance.

3.3.1 BERT-base-uncased (Devlin et al., 2019)

The most commonly used transformer-based English language pre-trained model is trained using BooksCorpus and English Wikipedia datasets using a masked language and next sentence prediction task. The input is passed through a tokenizer and a maximum length of 128 is considered. The [CLS] token representation from the last hidden state is used for text representation, with a dimensionality of 768.

3.3.2 ModernBERT (Warner et al., 2025)

ModernBERT is a recent advancement in the BERT family of transformer-based language encoders that leverage FlashAttention, an alternating attention mechanism, and extensive pre-training on 2 trillion tokens using a context window of up to 8,192 tokens. In our experiment, we are using a variant of ModernBERT developed by answerdotai/ModernBERT-base variant. The hidden state of the last layer of the [CLS] symbol is used for text representation and has a dimensionality of 768. The ModernBERT model, along with the BERT-base model, has been utilized as a performance benchmark in all the experiments for both families of systems.

3.3.3 RoBERTa-base (Liu et al., 2019)

The RoBERTa-base is a 125 million parameter encoder that is based on the BERT-base architecture but employs a significantly enhanced pre-training procedure. This includes dynamic masking, where the masking is not static as in the original BERT, the removal of the next sentence prediction task, the use of ten times more data in the pre-training procedure (approximately 160 GB of data, as opposed to the 16 GB of the original BERT), and the employment of bigger batches and longer training steps. In fact, the RoBERTa-base outperforms the original BERT-base in all the standard natural language processing tasks. In our work, the RoBERTa-base is our primary strong text-only baseline, as our hypothesis is that the more and varied pre-training data, especially the web crawled material which is closer in style to the Twitter dataset due to its informal nature, might transfer particularly well to our meme post task.

3.3.4 DistilBERT-base-uncased (Sanh et al., 2020)

DistilBERT-base-uncased is a knowledge-distilled model with 66 million parameters that is a variant of BERT-base. It has been designed to reproduce the output distributions of BERT-base while reducing the number of parameters by 40%. It also reduces inference time by 60%. The knowledge distillation process uses a compound loss function that includes a loss from a masked language model (MLM), cosine embedding alignment from a BERT teacher model, and a soft cross-entropy loss over teacher logits. Nevertheless, DistilBERT has around 97% of the performance of BERT-base on the GLUE test set. It has been designed to be an efficiency baseline to calculate the performance cost of model compression for meme stance classification tasks and to check whether the reduced model performance has a significant effect on task performance.

3.3.5 Deberta-v3-base (He et al., 2023)

DeBERTa-v3-base is an encoder model with 86M parameters. It has two major improvements over BERT: "disentangled attention" unlike traditional BERT, disentangled attention calculates weight using distinct content and position vectors, which enables the model to more effectively understand token relationships at different positions, and "Enhanced Mask Decoding" (EMD), which is similar to ELECTRA and includes "replaced token detec-

tion." It is used in text-only systems as a text encoder.

For all encoders, the `post_text` field is tokenized with padding and truncation to a maximum of 128 tokens. The representation at the `[CLS]` position from the last hidden state (768-d) is extracted and ℓ_2-normalized as

$$\hat{\mathbf{f}} = \frac{\mathbf{f}}{\|\mathbf{f}\|_2 + \epsilon}, \quad \epsilon = 10^{-8}, \tag{3}$$

before being passed to the classification head.

3.4 Loss Function

We employ a weighted cross-entropy loss to address class imbalance in the training data. Let $y \in \{0, 1, 2\}$ denote the ground-truth label and $\hat{p} \in \mathbb{R}^3$ the predicted class probabilities. The loss is defined as:

$$\mathcal{L} = -\sum_{c=1}^{3} w_c \, y_c \log(\hat{p}_c), \tag{4}$$

where w_c denotes the class-specific weight, and $y_c \in \{0, 1\}$ is the one-hot indicator for class c.

Accordingly, the class weights $w = [1.2, 1.2, 1.0]$ associated with [anti-vaccine, neutral, pro-vaccine] respectively were defined based on the distribution of training set (pro-vaccine: 3,199; anti-vaccine: 2,535; neutral: 2,461). This weighting strategy boosts the influence of relatively infrequent classes in both training and evaluation, which helps address bias towards larger classes.

When training with gradient accumulation, the per-batch loss will be divided by the number of accumulation steps before backpropagation is executed to ensure correct gradient scaling.

3.5 Data Augmentation

To improve visual generalization, we apply a stochastic data augmentation pipeline to all training images prior to CLIP preprocessing. It is not applied to the text-only systems, for which no image data is processed. The augmentation pipeline consists of: (i) random resized cropping to 224×224 with a scale range of $(0.8, 1.0)$, ensuring that at least 80% of the original image content is retained; (ii) random horizontal flipping with a probability of 0.5; (iii) random rotation within $\pm 15°$; and (iv) color jittering with brightness, contrast, and saturation factors of 0.2.

No data augmentation is applied to validation or test images. The same augmentation pipeline is used across all hybrid model and image only model (CLIP Vit-B/32) configurations.

4 Experimental Setup

4.1 Data Splits

For model training, we use the official VaxMeme training partition, and for hyperparameter tuning and early stopping, we use the official evaluation partition as the development (validation) set. The test labels are not available, and evaluation is conducted via the official CodaBench submission system. No external data is used for training.

4.2 Hyperparameters

All models are trained using a batch size of 32 with gradient accumulation over 4 steps, yielding an effective weight update batch size of 128. The CLIP visual encoder and text backbone use AdamW for optimization with a learning rate of 2×10^{-6} (for BERT-base systems) or 3×10^{-6} (for ModernBERT systems) and a weight decay of 0.01.

The classification head is trained with a learning rate of 1×10^{-4} and weight decay of 0.05 for BERT-base systems, and a learning rate of 1.25×10^{-3} for ModernBERT systems.

We use a OneCycleLR scheduler with pct_start $= 0.2$, which performs linear warm-up over the first 20% of training steps and then cosine anneals to zero. Prior to each optimizer step we perform gradient clipping with maximum norm of 0.5.

We train for up to 50 epochs with early stopping based on validation macro-F1 and a patience of 5 epochs. The checkpoint with the best validation macro-F1 is selected for evaluation on the final test set.

4.3 Preprocessing

We load all images with the Pillow library, in RGB format. In training, each image is processed through data augmentation as described in Section 3.5, followed by CLIP-specific preprocessing, consisting of resizing images to 224x224 and normalization using CLIP channel statistics, whereas validation and test use CLIP preprocessing only. In case an image is either missing from the filesystem or corrupted, a zero tensor of shape $(3, 224, 224)$ is used as a filler for those images.

We tokenize the textual input (*post_text*) separately for all models using its respective tokenizer with padding and truncation to up to a maximum sequence length of 128 tokens. In particular, we use `padding="max_length"`, `truncation=True`, `max_length=128` and `return_tensors="pt"`. All inputs are explicitly cast to string type for robustness against missing or null data.

4.4 Evaluation Measures

The evaluation measure is based on the macro-averaged F1 score. This is calculated using the unweighted mean of F1 scores for individual classes. This ensures that equal weight is given to every class irrespective of their frequency. In addition to this, we also calculate macro-averaged precision and recall.

The evaluation is done using the `scikit-learn` library. Specifically, we have used `f1_score`, `precision_score`, and `recall_score` with `average="macro"` arguments. In addition, we have used `zero_division=0` to handle undefined values.

4.5 Tools and Libraries

The experiments use Python 3.10, PyTorch 2.x, and Hugging Face Transformers v4.40+. Other libraries include OpenAI CLIP for visual encoding, scikit-learn for evaluation, Pillow for images, and pandas for data handling.

For training, experiments use free Google Colab GPUs with CUDA. If multiple GPUs are present, data parallel training is performed via PyTorch `DataParallel`.

5 Results

Table 1 displays the entire results of all systems over the official validation and test sets. The top-performing systems over all metrics are the BERT-base-uncased text-only model with a test macro-F1 of 0.8102, which slightly outperforms the second-best-performing model, RoBERTa-base text-only model with 0.8091. Other top-performing systems are the Deberta-v3-base text-only model with 0.7950 and the DistilBERT-base text-only model with 0.7938. The top-performing hybrid systems are the CLIP + ModernBERT model with 0.7783, followed by the CLIP + BERT-base model with 0.7603. The performance of the CLIP-only baseline model is 0.7189. This shows that the textual post_text features are much more discriminative than image features alone for this dataset.

5.1 Text-Only vs. Multimodal Hybrid

One of the interesting observations from the results in Table 1 is the consistent outperformance of text-only systems over their multimodal hybrid counterparts, even when the same text encoder is used. For example, the text-only system using the BERT-base-uncased text encoder achieves 0.8102, which is 4.99 percentage points ahead of the CLIP + BERT-base hybrid system, which achieves 0.7603. In another case, the ModernBERT text-only system achieves 0.7797, which is almost comparable to the performance of the CLIP + ModernBERT hybrid system, which achieves 0.7783. The metadata field "post_text" describing the text of the tweet in which the meme was originally shared is seen to perform exceptionally well as a feature for vaccine stance classification. The addition of the CLIP features does not seem to provide any advantage over the text-only system, which is seen to be the best-performing configuration in our experiments. The reason for this might be the fact that the text of the tweet originally shared by the author of the meme actually contains the stance, which might not require the use of the image. Future work might involve the use of cross-attention fusion or the use of the text read directly from the embedded text in the images, i.e., OCR.

5.2 Comparison Across Text Encoders

Among the text-only models, BERT-base-uncased has the highest test macro-F1 of 0.8102. This is followed closely by RoBERTa-base with a test macro-F1 of 0.8091. These two models have high performance compared to ModernBERT-base with a test macro-F1 of 0.7797, DistilBERT-base with a test macro-F1 of 0.7938 and Deberta-v3-base with a test macro-F1 of 0.7950. The underperformance of ModernBERT-base compared to BERT-base is also noteworthy, especially considering that ModernBERT-base was designed with a more recent architecture and a pretraining corpus several orders of magnitude larger. This may be due to the nature of the post_text field, which consists of short-form social media text with fewer than 128 tokens. In this case, the large context window of up to 8,192 tokens and the complex attention mechanisms of ModernBERT-base do not provide any additional value and may pose optimization difficulties. The test macro-F1 score for DeBERTa-v3-base is 0.7950, ranking it in the third position among text-only models despite its superior bench-

Table 1: Results on the official validation and test sets. P = macro precision; R = macro recall. Val = validation set

System	Pipeline	Val F1	Val P	Val R	Test F1	Test P	Test R
BERT-base-uncased	Text-only	0.8166	0.8182	0.8162	0.8102	0.8126	0.8113
RoBERTa-base	Text-only	0.8140	0.8141	0.8158	0.8091	0.8121	0.8117
DistilBERT-base	Text-only	0.8140	0.8135	0.8149	0.7938	0.7941	0.7949
ModernBERT-base	Text-only	0.7923	0.7945	0.7955	0.7797	0.7819	0.7819
Deberta-v3-base	Text-only	0.8024	0.8020	0.8030	0.7950	0.7957	0.7964
CLIP + BERT-base	Hybrid	0.7604	0.7676	0.7587	0.7603	0.7682	0.7603
CLIP + ModernBERT	Hybrid	0.7791	0.7874	0.7785	0.7783	0.7882	0.7789
CLIP ViT-B/32	Image-only	0.7217	0.7223	0.7220	0.7189	0.7192	0.7190

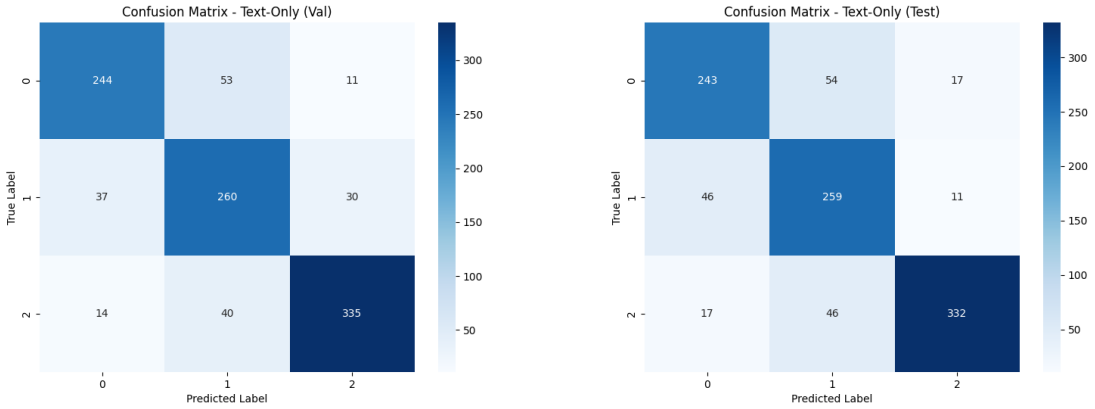


Figure 3: Confusion Matrices for text-only model BERT-base-uncased on validation and test sets.

mark performance on traditional NLP tasks. This shows that its disentangled attention mechanism provides little advantage in handling short informal text in tweets. The high performance of the RoBERTa-base model is also expected due to the pretraining of this model over a larger and more diverse corpus compared to BERT-base. The corpus used by RoBERTa-base also includes web-crawled text, which may be closer to Twitter text. The high validation F1 of 0.8140 and relatively lower test F1 (0.7938) by DistilBERT-base suggests some overfitting or train-test distribution mismatch, consistent with its reduced model capacity.

5.3 Error Analysis

The confusion matrices of the best-performing text-only system are presented in Figure 3 and its corresponding hybrid model's confusion matrices are presented in Figure 4, showing their performance on the validation and test sets. For both models, the neutral class, which corresponds to label 1, is the key challenge. For the test set, the hybrid system misclassifies 84 vaccine-critical and 57 pro-vaccine memes as neutral, whereas the text-only system misclassifies 54 and 46, respectively. Therefore, the text-only system is better for the neutral class. The pro-vaccine class, which corresponds to label 2, is the best for both models. For the hybrid system, the F1 score is 0.8548, whereas for the text-only system, the score is 0.8795, likely because pro-vaccine memes tend to have strong sentiment.

We manually analyzed the 100 cases of misclassification in the validation set of our best-performing model (BERT base uncased, text only). Three types of misclassifications were found.

i. **Sarcasm and ironic framing (42%):** Firstly, there are 42% of misclassified memes that express the vaccine-critical message with the use of sarcasm or irony. In these memes, words such as "Trust the science!" and "Safe and effective!" are used in an ironic way. However, the text encoder does not recognize the irony.

ii. **Implicit cultural references (31%):** Secondly, 31% of misclassifications are due to the meme implicitly referring to something, such as knowledge of a political figure, an in-

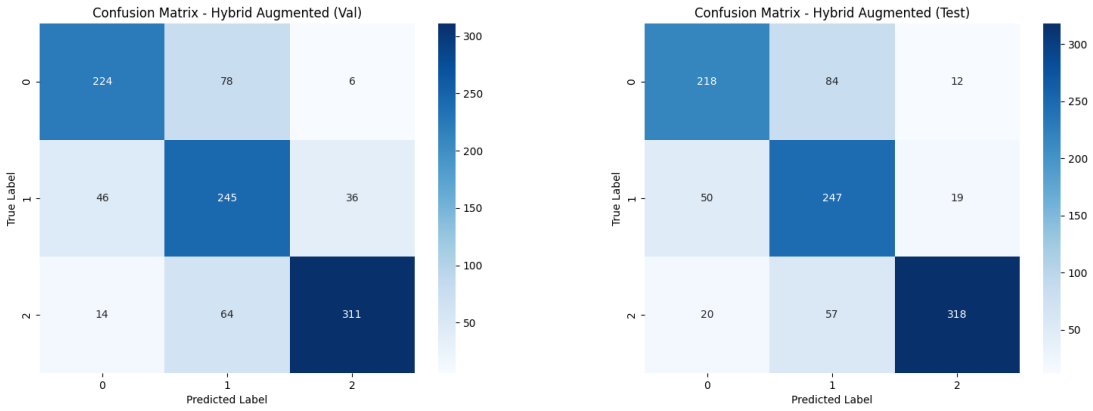


Figure 4: Confusion Matrices for hybrid model CLIP + BERT-base on validation and test sets.

ternet meme template, or a cultural event that cannot be inferred from the text alone. The model does not have enough world knowledge to comprehend the reference.

iii. **Neutral/vaccine-critical ambiguity (27%):** Lastly, 27% of misclassifications are due to the ambiguity of the meme, which represents a nuanced view that acknowledges both the benefits of vaccines as well as the concerns, thus placing it near the boundary of neutral and vaccine-critical in the original dataset.

6 Conclusion

We present our comprehensive system for the VaxMeme Shared Task at EEUCA 2026, comparing the performance of multimodal hybrid CLIP-fusion models with text-only fine-tuned language models for vaccine-critical meme stance classification. Our main discovery is that text-only models, especially the BERT-base-uncased model (0.8102 test macro F1), even outperform their multimodal counterparts when using the post_text tweet metadata signal. This suggests that the tweet text signal is highly discriminative for vaccine stance classification in this dataset. The CLIP model, using only image data, performs worse (0.7189) than text-based models, validating that vaccine stance cannot be reliably inferred from visual content alone. Common problems among configurations include sarcasm, culture, and ambiguity at the neutral/vaccine critical boundary.

Several avenues for enhancement have been identified. Firstly, the text extracted from the meme image using OCR can be leveraged as an additional visual-textual feature, different from post_text. Secondly, using cross-attention for fusion can improve the modeling of inter-modal relationships, thereby improving the understanding of sarcasm and implicit cultural cues. Thirdly, larger vision-language models, namely LLaVA and InstructBLIP, can improve the results for sarcasm detection and implicit cultural understanding. Lastly, using a retrieval model with an external knowledge base can improve the results by addressing the principal failure modes identified.

7 Limitations

- **Data dependency and overfitting:** The model heavily relies on *post_text*, which caused overfitting and weak generalization to visual inputs, as seen in the poor performance of image-only CLIP.
- **Simplistic multimodal fusion:** The fusion method is relatively simple and fails to capture complex relationships between text and images.
- **No image-text extraction:** The system does not extract text from images, missing important semantic information in memes.
- **Class ambiguity:** The model struggles with sarcasm, cultural context, and the distinction between neutral and vaccine-critical classes.
- **No external knowledge:** The approach does not use external knowledge or retrieval methods to resolve ambiguity.

8 Acknowledgments

We thank the organizers of the VaxMeme Shared Task at EEUCA 2026 (Thapa et al., 2026b; Hür-

riyetoğlu et al., 2026) for their efforts in creating the VaxMeme dataset and for running the shared task, along with maintaining the CodaBench platform. We also thank the anonymous reviewers for their helpful comments. This work was done without any external funding.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Kadhim Hayawi, Sakib Shahriar, Mohamed A. Serhani, Issam Taleb, and Sujith S. Mathew. 2022. Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. *Public Health*, 203:23–30.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

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)*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The hateful memes challenge: Detecting hate speech in multimodal memes. *Preprint*, arXiv:2005.04790.

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. *Preprint*, arXiv:1907.11692.

Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G Dunn. 2023. A multimodal framework for the identification of vaccine critical memes on twitter. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 706–714.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Laxmi Thapa, Aryaman Jain, Lakshmojee Koduru, Surabhi Adhikari, Junaid Rashid, Jungeun Kim, Surendrabikram Thapa, and Usman Naseem. 2026a. Concept-grounded detection of vaccine misinformation in multimodal content using interpretable vision-language models. In *Companion Proceedings of the ACM on Web Conference 2026*.

Surendrabikram Thapa, Kritesh Rauniyar, Hariram Veeramani, Aditya Shah, Imran Razzak, and Usman Naseem. 2024. Did you tell a deadly lie? evaluating large language models for health misinformation identification. In *International Conference on Web Information Systems Engineering*, pages 391–405. Springer.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Kritesh Rauniyar, Laxmi Thapa, Surabhi Adhikari, Kristina T. Johnson, Ali Hürriyetoğlu, Hristo Tanev, and Usman Naseem. 2026b. Multimodal identification of vaccine content stance on social media. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

A Full Hyperparameter Table

Table 2 provides a complete listing of all hyperparameter settings used in each experimental pipeline to allow full reproduction of reported results.

Table 2: Hyperparameter settings for text-only and hybrid pipelines.

Hyperparameter	**Text-Only**	**Hybrid**
Batch size	32	32
Gradient accumulation	4	4
Effective batch size	128	128
Max epochs	50	50
Early stopping (patience)	5	5
LR — backbone	3×10^{-6}	2×10^{-6} (BERT), 3×10^{-6} (ModernBERT)
LR — classifier head	1×10^{-4}	1×10^{-4}
Weight decay (backbone / head)	0.01 / 0.05	0.01 / 0.05
Optimizer	AdamW	AdamW
Scheduler	OneCycleLR (pct_start = 0.2)	OneCycleLR (pct_start = 0.2)
Gradient clipping	0.5	0.5
Dropout (512 / 128 layer)	0.3 / 0.2	0.3 / 0.2
Max text tokens	128	128
Image size	N/A	224×224
Image augmentation	N/A	RandomResizedCrop, HFlip, Rotation ($\pm 15^{\circ}$), ColorJitter
Class weights [0,1,2]	[1.2, 1.2, 1.0]	[1.2, 1.2, 1.0]

Author Index