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.

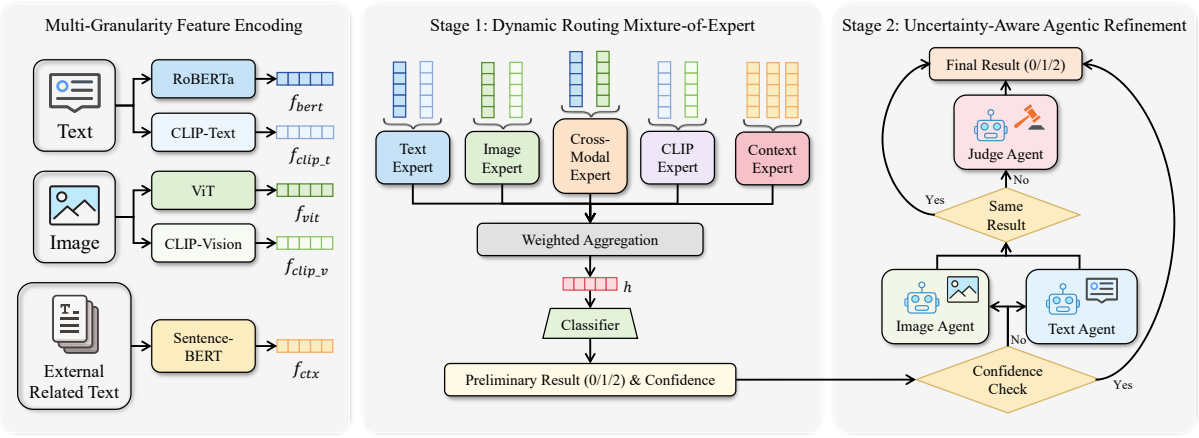

Figure 1: The overall architecture of the MoEs-VaxAgent. The framework consists of three parts: (1) **Multi-Granularity Feature Encoding**: utilizing RoBERTa, CLIP, ViT, and Sentence-BERT to extract comprehensive textual, visual, and external context features; (2) **Stage 1: Dynamic Routing Mixture-of-Experts**: 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.

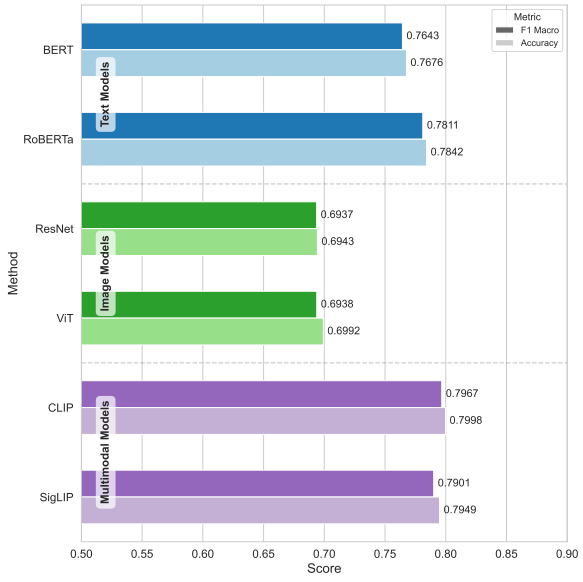

Figure 2: Performance comparison of various baseline models on the validation set.

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:

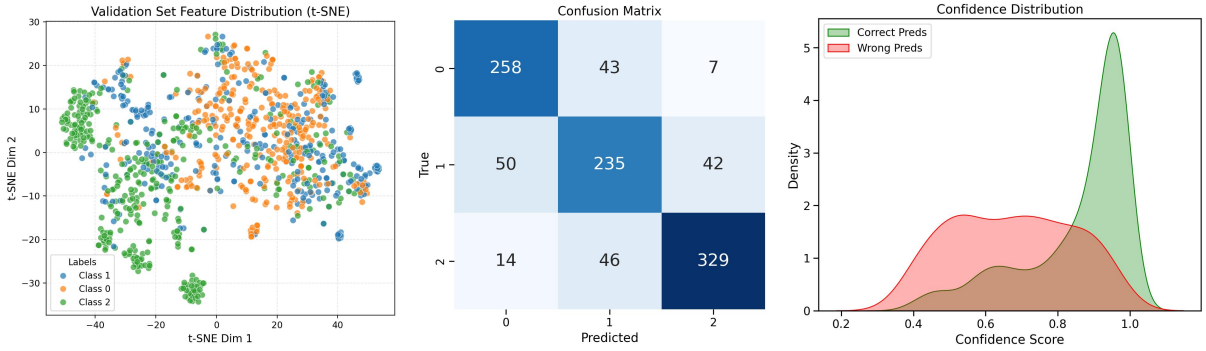

Figure 3: Visualization of model performance on the validation set. The figure displays the t-SNE feature distribution (Left), the Confusion Matrix (Middle), and the Confidence Distribution (Right) distinguishing between correct and incorrect predictions.

- *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.