Axiom@MMLow 2026: Cascaded Prompting with QLoRA-Tuned Qwen2-VL for Hierarchical Tamil Political Meme Classification

Md. Ajwad Hossain
Chittagong University of Engineering and Technology
md.ajwadhossain@gmail.com

Abstract

We present our system for the MMLow@AACL-IJCNLP 2026 shared task on multi-level classification of Tamil political memes. Our approach fine-tunes Qwen2-VL-2B-Instruct with QLoRA under a cascaded prompting strategy, in which a Level 1 stance prediction (Support/Praise vs. Troll/Oppose) conditions the prompt used for Level 2 target classification (party vs. person, per stance). On the official test set, our system achieves a Level 1 macro-F1 of 0.7173 and a Level 2 macro-F1 of 0.4429, for a final score of 0.5801, placing 5th of 7 teams. Our error analysis shows that all Level 1 errors propagate downstream, accounting for the majority of Level 2 mistakes. Additionally, a preliminary qualitative review suggests that sarcasm interpreted literally drives a large share of Level 1 misclassifications, pointing to sarcasm detection and cross-modal reasoning as key bottlenecks for this task.

1 Introduction

Political memes compress complex messages into shareable images, shaping discourse, mobilizing supporters, and occasionally spreading misinformation (Mihailescu, 2024; AlAfnan, 2025; Munawar, 2025). Understanding them requires genuine cross-modal reasoning, as a meme's meaning typically relies on the cross-modal interaction between individually benign image and text rather than anything literally stated (Kiela et al., 2020; Pramanick et al., 2021; Dimitrov et al., 2021).

Tamil political memes compound this difficulty. They draw on code-mixed Tamil-English, a linguistic setting that remains under-resourced (Chakravarthi et al., 2021) and heavily utilize culturally specific humor and sarcasm (Suryawanshi et al., 2020). While meme classification has matured considerably for English content, code-mixed settings "remain comparatively underexplored; historical Tamil multimodal benchmarks have consistently peaked well below comparable English tasks (Suryawanshi et al., 2020; Suryawanshi and Chakravarthi, 2021; B. et al., 2022).

This paper describes our system[1] for the MMLow@AACL-IJCNLP 2026 shared task, targeting hierarchical stance and target classification for Tamil-English political memes.Our contribution is threefold: we fine-tune Qwen2-VL-2B under tight hardware constraints using QLoRA, structure inference as a two-step cascaded prompting process, and provide a detailed error analysis identifying literal interpretations of sarcasm as a primary failure mode, corroborating prior warnings in this domain. (Suryawanshi et al., 2020; Singh et al., 2024).

2 Related Work

Meme Classification and VLMs. The Hateful Memes Challenge established the standard for subtle, multimodal reasoning (Kiela et al., 2020), which later expanded into political and propagandistic domains (Pramanick et al., 2021; Dimitrov et al., 2021). Modeling approaches for English memes have rapidly evolved from early fusion baselines (Hasan et al., 2022; Abdullakutty and Naseem, 2024) to CLIP-aligned encoders (Kumar and Nandakumar, 2022; Shah et al., 2024), and most recently to pipelines that prompt large vision-language models (VLMs) directly (Hee and Lee, 2025; Tzelepi and Mezaris, 2025). Our cascaded approach adapts this latest VLM paradigm to a hierarchical task.

Code-Mixing and Sarcasm. Our task extends the DravidianCodeMix trajectory, spanning from offensive language identification to multimodal troll meme classification (Chakravarthi et al., 2021; Suryawanshi et al., 2020; B. et al., 2022). Across

[1] Code available on GitHub.

Proceedings of the Fifth Workshop on Multimodal Machine Learning in Low-resource Languages, pages 64–69
November 9, 2026 ©2026 Association for Computational Linguistics

these tasks, sarcasm remains a notorious obstacle due to cross-modal incongruity that surface-level sentiment cues miss (Yue et al., 2023; Liang et al., 2022). While prior efforts confirm the difficulty of Tamil meme analysis, none have addressed the fine-grained hierarchical labels targeted in the current shared task.

Multilingual Grounding and PEFT. Separately, predominantly English-trained VLMs frequently degrade on low-resource tasks due to weak multilingual instruction-following (Schmidt et al., 2025; Atuhurra et al., 2025; Qu et al., 2026; Li et al., 2025). To adapt these massive backbones to code-mixed Tamil under severe compute constraints, we employ QLoRA. This parameter-efficient fine-tuning (PEFT) technique injects trainable low-rank matrices into frozen quantized weights, drastically reducing the memory footprint while maintaining competitive downstream performance (Hu et al., 2021; Martín-Fernández et al., 2025; Elgendy et al., 2025).

3 Methodology

3.1 Data

The shared task dataset consists of Tamil and Tamil-English code-mixed political memes, each labeled at two hierarchical levels. Table 1 shows the class distribution in the working training split, before and after balancing.

The raw distribution is lopsided, memes opposing an individual outnumber every other category by a wide margin, while support-for-party memes are scarce. Training on this directly risks pushing the model toward the majority class. The balanced subset fixes this by matching counts first across the two Level 1 classes, then across the Level 2 subclasses within each, leaving 324 training examples split almost evenly at Level 1 (163 Support/Praise, 161 Troll/Oppose).

Level 2 Class	Raw	Balanced
Troll/Oppose Against Person	465	83
Support for Person	150	81
Troll/Oppose Against Party	124	80
Support for Party	84	80
Total	823	324

Table 1: Training data class distribution before and after hierarchical balancing.

The full training set contains 969 labeled examples, 823 of which form the working split; the rest are held out for validation. The official test set comprises 244 unlabeled examples, with gold labels released by the organizers only after the evaluation period closed.

3.2 Model Architecture

Qwen2-VL-2B-Instruct serves as the backbone, chosen mainly for practical reasons: native support in the transformers library, decent OCR on embedded meme text out of the box, and stable performance under SDPA attention on hardware without FlashAttention2.

3.3 Cascaded Prompting

Rather than predicting both labels in one pass, the task splits into two sequential generation steps. The model first judges Level 1 stance from the image and its embedded text; that prediction then feeds into a second prompt asking for the specific Level 2 category (exact prompts in Appendix A). Training conditions the second step on the ground-truth Level 1 label, but at inference the model has only its own prediction to work with. That gap between training and inference matters a great deal, as Section 6 shows. Because the VLM generates free-text, we apply a "label snapping" function to raw outputs: we first check for an exact label match, then fall back to a substring match, and finally default to a fallback class if the generation is completely invalid.

3.4 Parameter-Efficient Fine-Tuning

Fine-tuning uses QLoRA with 4-bit NF4 quantization and bfloat16 compute precision. LoRA adapters (rank 16, alpha 32, dropout 0.05) attach to all attention and MLP projection layers, giving roughly 18.5M trainable parameters out of 2.23B total (0.83%). Full hyperparameters are in Appendix B.

4 Experimental Setup

Training runs on a single Tesla T4 GPU (15.6GB VRAM) on a free-tier cloud notebook, and that hardware shapes several design decisions. Qwen2-VL's dynamic-resolution tokenizer produces a different image-token count per sample, ruling out large batch sizes; instead, examples were processed with a micro-batch size of 2 and gradient accumulation steps of 8. Gradient checkpointing is likewise not optional: turning it off causes immediate out-of-memory failures even at the smallest batch size tested. Three epochs of training run in total, with

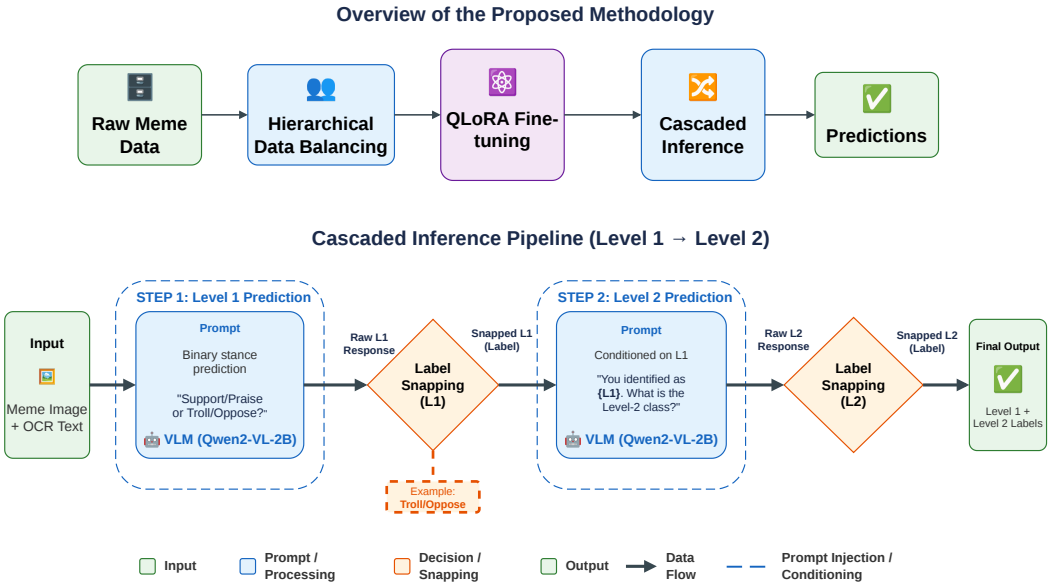


Figure 1: Cascaded inference pipeline (Level 1 → Level 2). The Level 1 binary stance prediction is inserted into the Level 2 prompt template, conditioning the detailed target classification on the model's own upstream prediction.

the final checkpoint chosen by validation accuracy on a held-out training split. These constraints also rule out a hyperparameter sweep, a LoRA rank ablation, and any comparison against larger or more explicitly multilingual backbones — open questions for future work rather than conclusions drawn without testing.

5 Results

Split	L1 Acc	L2 Acc
Zero-shot baseline	70.0%	50.0%
Epoch 1 (val, best)	92.5%	62.5%
Epoch 2 (val)	80.0%	52.5%
Epoch 3 (val)	87.5%	60.0%
Test (official)	75.0%	57.0%

Table 2: Validation accuracy across training epochs vs. final official test accuracy, using the epoch-1 checkpoint.

The epoch-1 checkpoint goes into the final submission, and the choice is fairly clear-cut: training loss sharply drops after the first epoch before plateauing slightly ($0.3275 \rightarrow 0.1138 \rightarrow 0.1181$), but validation accuracy peaks after epoch 1 and never comes back indication of classic overfitting on a small balanced subset. On the official test set, this checkpoint scores 0.7173 macro-F1 at Level 1 and 0.4429 at Level 2, for a final combined score of 0.5801, good for 5th place out of 7 teams.

The gap between the validation numbers (92.5%/62.5%) and the official test result (75.0%/57.0%) is larger than ideal, and three factors likely drive most of it. The validation split comes from the same source distribution as training, so it probably understates how different the real test set turns out to be. Validation scoring also relies on lenient substring matching rather than exact match, which almost certainly inflates the numbers relative to the official strict scorer. And epoch checks cover only a 40-sample subset of validation data to save time, adding noise to an already tight selection call.

6 Error Analysis

Since the Level 2 prompt conditions on the model's own Level 1 prediction at inference, not the ground truth, any Level 1 mistake gets a direct shot at derailing Level 2. That propagation turns out to be complete rather than partial: Level 2 accuracy sits at 75.96% when Level 1 is correct, but drops to exactly **0.00%** when Level 1 is wrong ($n = 61$). While all 61 Level 1 errors successfully derailed Level 2, approximately 44 additional Level 2 errors occurred even when Level 1 was correct. Thus, roughly 58% of Level 2 mistakes originate directly from upstream Level 1 failures; nothing in the current pipeline lets the model second-guess an earlier call.

The Level 1 confusion matrix shows a clear lean: genuine Troll/Oppose content gets mistaken for

Support/Praise about twice as often as the reverse (41 vs. 20 cases). Target identification alone (person vs. party, ignoring stance) fares better at 75.8% accuracy, though a bias shows up here too, "Support for Person" gets over-predicted relative to gold (81 vs. 45) while "Support for Party" gets under-predicted (10 vs. 25), pointing to a default toward naming a recognizable individual even when a meme's real target is a party or organization.

A manual review of 15 misclassified examples (categorized in Appendix C) digs into what actually drives the Level 1 errors. While this sample is too small for definitive statistical conclusions, this preliminary review suggests sarcasm read literally is the largest single pattern, at roughly 40% of the sample: memes quoting a political figure's self-congratulatory claim, undercut moments later by an ironic follow-up, get scored as genuine praise. A smaller, opposite pattern shows up too (about 27%): intensely worded but genuinely supportive fan rhetoric gets mistaken for hostility, likely because surface tone outweighs underlying intent in the model's judgment. The rest split between cases needing specific cultural or political background knowledge and a handful that are genuinely ambiguous even under close manual review.

7 Discussion

Set against the top-performing system's 0.9027/0.6694 L1/L2 macro-F1, this result points to a real multilinguality gap. Qwen2-VL-2B is pretrained mainly on English and Mandarin text, and while OCR on embedded Tamil script works well enough, its grip on Tamil-English code-mixed political discourse seems noticeably looser, especially around sarcasm, which leans on culturally specific cues a model without strong grounding in the language and its politics would likely miss. That lines up with the top-ranked teams' apparent choice of more broadly multilingual backbones, though this hypothesis hasn't been tested directly against an alternative backbone given the compute budget here, so it stands as a plausible explanation rather than a demonstrated one.

8 Conclusion

This paper presents a cascaded system for hierarchical Tamil political meme classification, fine-tuning Qwen2-VL-2B-Instruct with QLoRA under fairly tight hardware constraints. The final system reaches a combined score of 0.5801 (L1 macro-F1 0.7173, L2 macro-F1 0.4429), placing 5th among 7 teams. Two failure modes stand out from the error analysis: the strict downstream propagation of Level 1 errors under the cascade, and preliminary qualitative evidence pointing to literal interpretations of sarcasm as a primary contributor to Level 1 mistakes. Both point fairly directly to where the work should go next: architectures that let the model revisit an early misstep instead of committing to it, backbones with stronger multilingual and code-mixed grounding, and training data curated with sarcasm and irony specifically in mind.

Limitations

The validation setup relies on lenient substring matching and checks only a 40-sample subset of validation data per epoch, likely overstating real performance relative to the official strict-match scorer and contributing to potential instability in model selection; exact-match scoring from the start would serve future work better. Additionally, discarding over half of the available training data (823 down to 324 examples) for hierarchical balancing likely harmed generalization in an already low-resource setting. Future iterations should explore class-weighted losses or oversampling instead of strict undersampling.

Crucially, because the Level 2 labels inherently combine stance and target, a direct four-way classification model might implicitly recover both levels without requiring an upstream prediction. A limitation of this work is that cascaded prompting was not explicitly ablated against a joint or direct four-way prediction baseline, meaning the isolated benefit of the cascade remains an open question.

Finally, a non-native Tamil speaker carried out the qualitative error review. While the broad categories assigned seem reasonably solid, subtler sarcasm markers may have gone unnoticed; native-speaker annotation would strengthen this analysis. Compute constraints also rule out hyperparameter ablations, LoRA rank sweeps, and head-to-head comparisons against larger or more explicitly multilingual backbones, which remain the most promising next steps rather than settled questions.

References

F. Abdullakutty and Usman Naseem. 2024. Decoding memes: A comprehensive analysis of late and early fusion models for explainable meme analysis. *Companion Proceedings of the ACM Web Conference 2024*.

M. AlAfnan. 2025. The role of memes in shaping political discourse on social media. *Studies in Media and Communication*.

Jesse Atuhurra, Iqra Ali, Tomoya Iwakura, Hidetaka Kamigaito, and Tatsuya Hiraoka. 2025. Vlures: Benchmarking vlm visual and linguistic understanding in low-resource languages. *Preprint*, arXiv:2510.12845.

P. B., Bharathi Raja Chakravarthi, Malliga Subramanian, Bharathi B., Soman Kp, Kishore D., Sreelakshmi K., Arunaggiri Pandian, and Prasanna Kumar Kumaresan. 2022. Findings of the shared task on multimodal sentiment analysis and troll meme classification in dravidian languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*.

Bharathi Raja Chakravarthi, R. Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, E. Sherly, and John P. McCrae. 2021. Dravidiancodemix: sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*, 56:765–806.

Dimitar I. Dimitrov, Bishr Bin Ali, S. Shaar, Firoj Alam, F. Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting propaganda techniques in memes. pages 6603–6617.

Hosam Elgendy, Ahmed Sharshar, Ahmed Aboeitta, Yasser Ashraf, and Mohsen Guizani. 2025. Geollava: Efficient fine-tuned vision-language models for temporal change detection in remote sensing. *Preprint*, arXiv:2410.19552.

M. Hasan, Nusratul Jannat, Eftekhar Hossain, Omar Sharif, and M. M. Hoque. 2022. Cuet-nlp@dravidianlangtech-acl2022: Investigating deep learning techniques to detect multimodal troll memes. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*.

Ming Shan Hee and Roy Ka-Wei Lee. 2025. Demystifying hateful content: Leveraging large multimodal models for hateful meme detection with explainable decisions. *Preprint*, arXiv:2502.11073.

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.

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 *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 171–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Juncheng Li, Qing Yang, Bojian Jiang, Shaolin Zhu, and Qingxuan Sun. 2025. Lrm-llava: Overcoming the modality gap of multilingual large language-vision model for low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 24449–24457.

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1767–1777, Dublin, Ireland. Association for Computational Linguistics.

Iván Martín-Fernández, Sergio Esteban-Romero, Fernando Fernández-Martínez, and Manuel Gil-Martín. 2025. Parameter-efficient adaptation of large vision—language models for video memorability prediction. *Sensors*, 25(6).

Mihaela-Georgiana Mihailescu. 2024. Never mess with the "memers": How meme creators are redefining contemporary politics. *Social Media + Society*, 10.

Zunaira Munawar. 2025. The role of political social media memes on shaping public opinion: A survey analysis. *Lead Sci Journal of Management, Innovation and Social Sciences*.

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.

Xiaoye Qu, Mingyang Song, Wei Wei, Daizong Liu, Jianfeng Dong, and Yu Cheng. 2026. Mitigating multilingual hallucination in large vision-language models. *ACM Transactions on Multimedia Computing, Communications and Applications*, 22(5).

Fabian David Schmidt, Florian Schneider, Chris Biemann, and Goran Glavaš. 2025. MVL-SIB: A massively multilingual vision-language benchmark for cross-modal topical matching. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16285–16312, Vienna, Austria. Association for Computational Linguistics.

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.

Aakash Singh, Deepawali Sharma, and V. K. Singh. 2024. Mimic: Misogyny identification in multimodal internet content in hindi-english code-mixed language. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on troll meme classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 126–132, Kyiv. Association for Computational Linguistics.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John P. McCrae, and Paul Buitelaar. 2020. A dataset for troll classification of TamilMemes. In *Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation*, pages 7–13, Marseille, France. European Language Resources Association (ELRA).

Maria Tzelepi and Vasileios Mezaris. 2025. Improving multimodal hateful meme detection exploiting lmm-generated knowledge. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 202–211.

Tan Yue, Rui Mao, Heng Wang, Zonghai Hu, and Erik Cambria. 2023. Knowlenet: Knowledge fusion network for multimodal sarcasm detection. *Information Fusion*, 100:101921.

A Cascade Prompts

Level 1 prompt:

```
Read the text in this meme
and look at the image.  Is
this 'Support/Praise' or
'Troll/Oppose'?  Answer
with just the label.
```

Level 2 prompt (cascaded on predicted/gold L1 label):

```
You identified this meme
as {level1_label}.  Based
on the image and text,
what is the specific Level
2 category?  Answer with
just the label.
```

B Hyperparameters

Parameter	Value
Backbone	Qwen2-VL-2B-Instruct
Quantization	4-bit NF4
Compute dtype	bfloat16
LoRA rank / alpha / dropout	16 / 32 / 0.05
Target modules	q,k,v,o,gate,up,down_proj
Trainable params	18.5M / 2.23B (0.83%)
Batch size / grad accum	2 / 8 (eff. 16)
Max new tokens	8
Epochs	3
Training samples	324 (balanced)
Attention implementation	SDPA
Hardware	1× T4, 15.6GB VRAM
Training time	∼3.7h/epoch (∼11h total)

Table 3: Training hyperparameters.

C Qualitative Error Categorization

We manually reviewed a sample of 15 Level 1 misclassifications and identified four recurring failure categories:

- **Sarcasm/irony read literally (6/15, 40%):** Memes quoting a political figure's self-congratulatory claim, undercut by an ironic follow-up line, scored as literal endorsement.
- **Reverse bias intense phrasing mistaken for hostility (4/15, 27%):** Exclamatory or combative language in genuinely supportive fan rhetoric misread as opposition.
- **Cultural/political knowledge gaps (2–3/15, ∼15–20%):** Cases requiring specific historical or intra-party knowledge (e.g. references to deceased political figures, party succession politics) underrepresented in general-purpose VLM pretraining.
- **Genuinely ambiguous (∼2/15, ∼13%):** Borderline cases even under manual, non-native-speaker review.