MathAddicts at EXIST 2026: Heterogeneous Vision-Language Ensembles with Diverse PEFT Adapters for Multimodal Sexism Detection

MathAddicts at CLEF 2026

Ali Yıldız[1,2,*], Yusuf Hasan Onkun[1,*], Ahmet Nedim Kocatürk[1], Haktan Cebeci[1], Yaşar Cengizhan Açıcı[1], Ahmet Sadık Demirci[1], Alperen Korkmaz[1] and Can Ekkazan[2]

[1]*Istanbul Technical University, Ayazağa, Sarıyer, Istanbul, Turkey*
[2]*Optdcom Technology, Beşiktaş, Istanbul, Turkey*

Abstract

This paper presents our system for the EXIST 2026 shared task (Task 2), targeting the detection and characterization of sexism in memes across English and Spanish. Our approach deploys a heterogeneous ensemble of three Vision-Language Models (VLMs): Qwen 3.5 4B, GLM-4.6V-Flash, and Qwen 3.5 9B. Each of them is fine-tuned from base weights using 4-bit quantized parameter-efficient fine-tuning (PEFT) methods: QLoRA for Qwen 3.5 4B and GLM-4.6V-Flash, and QDoRA for Qwen 3.5 9B. All models process memes through a sequential zero-shot pipeline covering three subtasks: binary sexism identification (Subtask 2.1), source-intention classification (Subtask 2.2), and multi-label sexism categorization (Subtask 2.3). Ensemble predictions are consolidated via majority voting for Subtasks 2.1 and 2.2, and via at-least-one voting for Subtask 2.3. In the official EXIST 2026 evaluation, our system achieved Rank 10 (F1 = 0.7514) on Subtask 2.1, Rank 6 (F1 = 0.5469) on Subtask 2.2, and Rank 7 (F1 = 0.5413) on Subtask 2.3.

Keywords

Sexism Detection, Multimodal Learning, Vision-Language Models, Parameter-Efficient Fine-Tuning, LoRA, DoRA, Ensemble Methods, EXIST 2026

1. Introduction

The widespread adoption of social media platforms has transformed the way harmful content spreads online, with sexism emerging as one of the most prevalent forms of gender-based abuse [1]. Among the formats through which sexist content circulates, memes have become particularly prevalent, since they combine text and imagery that may appear harmless in isolation but together can convey offensive or degrading messages [2].

Automated detection of such content remains a computationally challenging problem. Sexist meaning in memes is frequently implicit, ironic, or culturally embedded, requiring simultaneous understanding of both visual and textual context. Compounding this, the perception of what constitutes sexism is inherently subjective, leading to systematic disagreement among human annotators [3].

The EXIST 2026 [4, 5] shared task (Task 2) targets this challenge directly, requiring systems to classify memes in both English and Spanish across three consecutive subtasks:

- Subtask 2.1 – Binary identification: is the meme sexist or not?
- Subtask 2.2 – Intention classification: is the sexism direct or judgmental (i.e., critical of sexism)?
- Subtask 2.3 – Category assignment: one or more sociological categories of sexism (multi-label).

The dataset contains memes in both English and Spanish, making multilingual robustness a key criterion throughout model selection. Our core strategy is a heterogeneous ensemble of fine-tuned Vision-Language Models (VLMs) [6] resolved via voting-based aggregation, with the strategy adapted per subtask. By pairing models with different architectures and different parameter-efficient fine-tuning (PEFT)

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
yildiza23@itu.edu.tr (A. Yıldız); onkun25@itu.edu.tr (Y. H. Onkun)

methods, we reduce correlated errors across the ensemble. This approach is grounded in ensemble diversity theory, which establishes that combining heterogeneous classifiers consistently outperforms any single strong model when the component models exhibit reasonably low error correlation [7].

2. Literature Review

Research on sexism detection in social media has developed alongside the EXIST benchmark series. Earlier editions established multilingual evaluation frameworks for identifying sexist content on X (previously known as Twitter) and other platforms, progressively introducing annotation disagreement modeling and source-intention classification as additional dimensions of analysis [8]. These shared tasks have driven advances in transformer-based architectures and annotation methodology, providing strong baselines and evaluation protocols that directly inform our approach.

Meme-based sexism detection presents unique challenges beyond text-only classification, since meaning frequently arises from the interplay between overlaid text and visual content [2]. The self-attention mechanism [9] has proven foundational to these advances: cross-modal attention mechanisms and multimodal architectures have been shown to outperform text-only baselines on meme understanding tasks by capturing visual cues, such as facial expressions, scene composition, and stereotype-reinforcing imagery, which are invisible to language models operating on text alone [6]. While LLM-based approaches have proven highly effective for text-based sexism detection, most recently achieving first place across all subtasks of EXIST 2025 Task 1 [10], Task 2 operates on memes rather than plain text, making multimodal architectures a necessity rather than a design choice.

Parameter-efficient fine-tuning (PEFT) has become the dominant paradigm for adapting large pretrained models to downstream tasks under memory constraints. LoRA [11] introduces trainable low-rank decompositions into the weight matrices of attention layers, reducing the number of trainable parameters by orders of magnitude while retaining most of the representational capacity of full fine-tuning. DoRA [12] extends this framework by decomposing weight updates into separate magnitude and direction components, more closely approximating full fine-tuning behavior. The QLoRA methodology [13] further enables fine-tuning of models exceeding 7B parameters on commodity hardware by combining 4-bit NormalFloat quantization with double quantization and LoRA adapters, with negligible performance degradation relative to full-precision baselines [14].

Ensemble methods provide a principled approach to variance reduction in classification tasks. The theoretical justification for heterogeneous ensembles rests on the principle that combining classifiers with low pairwise error correlation reduces overall expected error, even when individual classifiers are imperfect [7]. In the context of LLM-based classifiers, majority voting and related aggregation strategies have been shown to provide consistent gains over individual models, particularly when the component models differ in architecture or optimization objective [15]. Annotator disagreement in sexism labeling datasets further motivates ensemble approaches, as majority-vote label aggregation mirrors the process by which ground-truth labels are produced in the first place [15, 16].

3. System Design

This section describes our data preparation pipeline and system design, covering the dataset split strategy, preprocessing steps applied to training instances, model selection, parameter-efficient fine-tuning configuration, inference pipeline, and the voting mechanisms used to aggregate predictions across the ensemble.

3.1. Data Preparation

Data was split 90:10 into training and validation sets, a ratio chosen to maximize training signal while retaining a sufficiently large held-out set for reliable performance estimation. Because standard single-label stratification is structurally inapplicable to overlapping target variables, we utilized multi-label

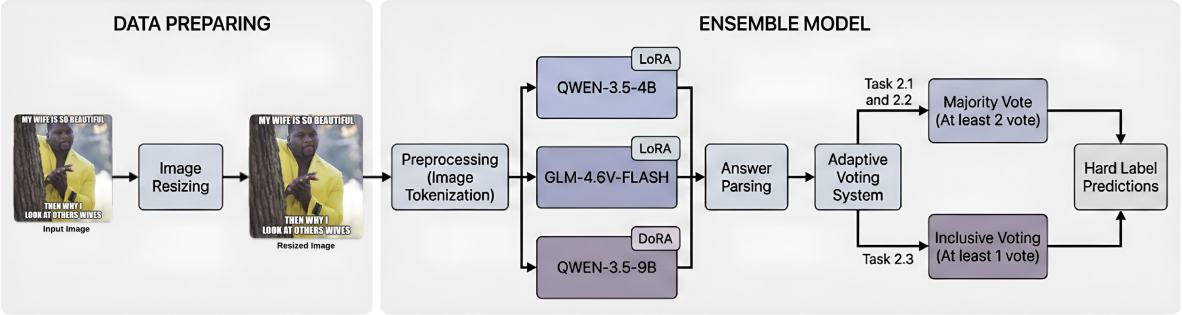

Figure 1: Overview of the proposed pipeline. The system processes input memes through image resizing and tokenization, passes them to an ensemble of three fine-tuned VLMs, and aggregates predictions via an adaptive voting strategy: majority vote for Subtasks 2.1 and 2.2, inclusive vote for Subtask 2.3.

stratified sampling [17] across all subtasks to jointly balance the meme languages (English and Spanish) and task-specific labels, ensuring identical data distributions across splits.

Input images were resized to a maximum of 588 px on the longest side while preserving aspect ratio, then both spatial dimensions were floor-rounded to the nearest multiple of 14 to align with the patch grid of the vision encoders.

3.2. Model Selection

A key constraint throughout this project was the exclusive use of base models. Instruction-tuned or otherwise post-trained variants were excluded because fine-tuning on top of already instruction-tuned models risks interference with the task-specific signal introduced during training, and introduces unknown pre-existing biases toward particular output formats. This constraint also ensures cleaner and more predictable fine-tuning dynamics. Given this restriction, model selection was governed by three criteria: (1) strong benchmark performance on vision-language tasks at small parameter scale; (2) multilingual capability, given the English/Spanish composition of the dataset; and (3) availability as open-weight base models compatible with 4-bit quantized training.

3.2.1. Qwen 3.5 4B

Qwen 3.5 4B [18] is a 4-billion-parameter vision-language model that offers a strong balance between performance and computational efficiency. Although substantially smaller than the 9B variant, it retains strong multimodal capabilities and provides a favorable accuracy-to-parameter ratio. We fine-tuned the model using QLoRA [13] applied to its attention projection layers. Within the ensemble, it serves as a lightweight component whose predictions complement those of the larger models, contributing to the diversity and robustness of the voting scheme.

3.2.2. GLM-4.6V-Flash

GLM-4.6V-Flash [19] is a 9-billion-parameter vision-language model selected for its strong multilingual capabilities and its distinct architectural lineage from the Qwen family of models [18]. Due to differences in architecture, pretraining objectives, and tokenization strategies, we expected GLM to learn complementary representations and capture patterns that might be overlooked by Qwen-based models, thereby increasing ensemble diversity and reducing correlated errors.

3.2.3. Qwen 3.5 9B

Qwen 3.5 9B [18] is the largest and strongest individual model in our system. We apply QDoRA [12] to this model rather than standard QLoRA [13] for two reasons. First, DoRA decomposes weight updates into separate magnitude and direction components, allowing the adapter to more closely approximate

Table 1
Shared adapter hyperparameter configuration across all models.

Hyperparameter	Value
Target Modules	All linear modules
Rank (r)	128
Alpha (α)	256
Learning Rate	8×10^{-6}
Optimizer	AdamW
Batch Size	32
Base Precision	4-bit NF4 + double quantization

full fine-tuning behavior [12], which is particularly valuable for a complex cross-modal task such as meme sexism detection. Second, and critically for ensemble design, we already had a Qwen-family model using QLoRA [13, 18]. Adding a second Qwen model with QLoRA risked having the two Qwen+QLoRA configurations produce highly correlated predictions, effectively reducing the ensemble to two independent voices rather than three. By applying QDoRA to Qwen 3.5 9B, we ensure that each model in the ensemble differs not only in scale or architecture but also in its fine-tuning dynamics, maximizing prediction diversity and reducing the risk of any single configuration dominating the vote.

We additionally note that Ministral 8B [20] was trained to completion using QLoRA with the same hyperparameter configuration as the primary models, but yielded a validation F1-score of 0.68 and consistently degraded ensemble performance when included in the voting pool [15]; it was therefore excluded from the final system.

3.3. Parameter-Efficient Fine-Tuning

All models are loaded in 4-bit quantized precision using the NormalFloat4 (NF4) format with double quantization, following the QLoRA methodology introduced by Dettmers et al. [13]. This precision level has been shown to be near-universally optimal for the trade-off between model size and zero-shot accuracy across LLM families ranging from 19M to 176B parameters [14]. This approach drastically reduces the memory footprint of the base weights while retaining full-precision adapter parameters, enabling fine-tuning of 9B-parameter models within the memory constraints of our hardware. The adapter configuration is shared across all models and is summarized in Table 1. We use AdamW as our optimizer [21] with a cosine annealing learning rate schedule [22], which decouples weight decay from the gradient update step and has been shown to produce more stable training than standard Adam for transformer-based architectures.

To address severe class imbalance during training, losses for both binary subtasks were scaled using inverse-frequency class weights computed strictly from the training split and rescaled to a mean of 1.0, aligning with traditional cost-sensitive learning frameworks [23]. For the multi-label Subtask 2.3, per-class weights were calculated for each category, and instance-level weights were determined via a maximum reduction strategy across positive labels. This design explicitly prevents the loss gradients of rare, co-occurring categories from being mathematically diluted by more frequent labels within the same sample.

3.4. Inference Pipeline and Prompt Design

All three models process each meme through the same three-stage pipeline at inference time. We adopt a zero-shot prompting format, presenting instructions and expected output structure to the model without embedding training examples in the context window [24]. This choice was deliberate: including few-shot examples risks biasing predictions toward the label distribution of those examples, which is problematic given the subjective and culturally variable nature of meme sexism. The system prompts strictly prohibit conversational preambles and markdown formatting fences to prevent output parsing

failures. The system prompts can be seen as below for each subtask:

System Prompt Subtask 2.1 — Sexism Identification
You are a research classifier for sexism detection in memes. You will be given a meme image. The meme may be in English or Spanish.
Classify the meme:
- YES: the meme depicts, promotes, normalizes, or critiques a sexist situation or behavior.
- NO: the meme does not contain sexist content.
Output only the JSON object. No preamble, no markdown fences, no explanations.
{"label": "YES"} or {"label": "NO"}

System Prompt Subtask 2.2 — Intention Detection
You are a research classifier for detecting the author's intention behind sexist memes. You will be given a meme image that has already been identified as sexist. The meme may be in English or Spanish.
Classify the author's intention:
- DIRECT: the meme is itself sexist or incites sexism.
- JUDGEMENTAL: the meme describes a sexist situation or behavior with the aim of condemning it.
Output only the JSON object. No preamble, no markdown fences, no explanations.
{"label": "DIRECT"} or {"label": "JUDGEMENTAL"}

System Prompt Subtask 2.3 — Sexism Categorization
You are a research classifier for categorizing sexism in memes. You will be given a meme image that has already been identified as sexist. The meme may be in English or Spanish.
Assign one or more of the following categories. At least one must be selected.
- "IDEOLOGICAL-INEQUALITY": discredits feminism, denies gender inequality, or frames men as victims.
- "STEREOTYPING-DOMINANCE": reinforces gender roles or claims male superiority.
- "OBJECTIFICATION": treats women as objects, hypersexualizes female attributes, or enforces beauty standards.
- "SEXUAL-VIOLENCE": contains sexual suggestions, harassment, or references to sexual assault.
- "MISOGYNY-NON-SEXUAL-VIOLENCE": expresses hatred or promotes non-sexual violence toward women.
Output only the JSON object. No preamble, no markdown fences, no explanations.
{"labels": ["<category>", ...]}

3.5. Voting Mechanism

3.5.1. Subtasks 2.1 and 2.2: Majority Vote

For the binary sexism identification and intention classification subtasks, a hard majority vote is applied: a label is assigned if at least two of the three models produce that label. This threshold proved effective for both tasks and was retained without modification.

3.5.2. Subtask 2.3: At-Least-One Vote

Subtask 2.3 is a multi-label classification problem. Initial experiments applied the same majority voting rule used in the preceding subtasks. This configuration proved excessively conservative: it systematically suppressed valid but infrequently assigned categories [16], causing the ensemble to underperform individual models on this subtask.

We subsequently adopted an at-least-one voting rule, in which a category is included in the final label set if at least one of the three models assigns it. The empirical comparison between these strategies is presented in Table 2.

Table 2
Comparison of voting strategies for Subtask 2.3 on the validation set.

Voting Strategy (Subtask 2.3)	Ensemble F1
Majority Vote ($\geq$ 2 of 3 models)	0.6697
At-Least-One Vote ($\geq$ 1 of 3 models)	0.7155

To understand the source of this gain, we break down performance by category (Table 3).

Table 3
Per-category F1 under majority-vote and at-least-one voting for Subtask 2.3, validation set. Abbreviations: ID-INEQ = Ideological-Inequality, STEREO-DOM = Stereotyping-Dominance, OBJ = Objectification, SEX-VIOL = Sexual-Violence, MISOGYNY-NSV = Misogyny/Non-Sexual-Violence.

Category	# of Samples	Majority Vote F1	At-least-one-vote F1	ΔF1
ID-INEQ	85	0.750	0.749	-0.001
STEREO-DOM	109	0.683	0.730	$+0.047$
OBJ	90	0.832	0.817	-0.015
SEX-VIOL	46	0.643	0.690	$+0.047$
MISOGYNY-NSV	41	0.441	0.591	$+0.150$

The gain from at-least-one voting is not evenly distributed across categories. The two rarest categories, MISOGYNY-NSV (n = 41) and SEX-VIOL (n = 46), show the largest improvements: at-least-one voting raises recall on MISOGYNY-NSV from 0.317 to 0.634 (F1: 0.441 $\rightarrow$ 0.591) and on SEX-VIOL from 0.587 to 0.848 (F1: 0.643 $\rightarrow$ 0.690). This is because majority voting requires 2/3 models to agree on labels that individual models already under-predict, disproportionately suppressing rare categories. By contrast, for OBJ, where individual-model recall is already high under majority voting (0.878), at-least-one voting slightly reduces F1 (0.832 $\rightarrow$ 0.817) by admitting additional false positives with little remaining recall to recover. This confirms that at-least-one voting is most beneficial precisely where the reviewer's concern about mutually exclusive treatment of categories is sharpest: rare, frequently co-occurring categories that a conservative majority rule tends to suppress.

3.5.3. Ensemble Diversity Analysis

A potential concern in ensemble design is the risk of model dominance, where one component model is sufficiently strong that the remaining models merely introduce noise. To assess whether the Qwen models would overwhelm the ensemble, we analyzed the distribution of 2-1 disagreements (cases where one model dissents from the other two) across the three possible pairs. Cases of unanimous agreement were excluded from this analysis, as they do not differentiate ensemble behavior. Table 4 reports the results for Subtasks 2.1 and 2.2 separately.

The results confirm that no single model pair dominates the disagreement distribution. Notably, the two Qwen models do not consistently ally against GLM-4.6V-Flash. In Subtask 2.1, the two Qwen models form the majority pair in only 27.4% of disagreement cases, a rate identical to when GLM and Qwen 3.5 9B form the majority pair against Qwen 3.5 4B. This balanced distribution indicates that the GLM model captures complementary information rather than acting as a passive third voter. Ultimately, the Qwen models do not dominate the ensemble, validating the inclusion of GLM-4.6V-Flash as a valuable complementary component.

4. Results

We report results on both the internal validation set and the official EXIST 2026 evaluation. Validation results measure each model and ensemble configuration independently; official results reflect the final submitted system.

Table 4
Distribution of 2-1 disagreements across model pairs on the validation set. Counts reflect only cases where the three models did not reach unanimous agreement.

Majority Pair	Subtask 2.1 (n / %)	Subtask 2.2 (n / %)
Qwen 3.5 4B + GLM	79 / 175 (45.1%)	41 / 94 (43.6%)
Qwen 3.5 4B + Qwen 3.5 9B	48 / 175 (27.4%)	31 / 94 (33.0%)
GLM + Qwen 3.5 9B	48 / 175 (27.4%)	22 / 94 (23.4%)

Table 5
Individual model and ensemble F1 performance on the validation set.

Model / Configuration	Subtask 2.1 (F1)	Subtask 2.2 (F1)	Subtask 2.3 (F1)
Qwen 3.5 4B (QLoRA)	0.7657	0.6415	0.6421
GLM-4.6V-Flash (QLoRA)	0.7675	0.6097	0.6772
Qwen 3.5 9B (QDoRA)	0.7923	**0.6552**	0.6469
Consensus Ensemble	**0.7967**	0.6497	**0.7155**

4.1. Validation Performance

Table 5 reports F1-scores for each individual model and the final consensus ensemble on the validation dataset across all three subtasks.

Among the individual models, Qwen 3.5 9B (QDoRA) achieves the highest Subtask 2.1 F1-score of 0.7923, confirming that the additional expressiveness of weight-decomposed adaptation provides a measurable advantage over standard QLoRA at equivalent model scale. GLM-4.6V-Flash achieves an F1-score of 0.7675 on Subtask 2.1 despite its distinct architectural lineage, validating its selection as a source of representation diversity within the ensemble. For Subtask 2.3 (multi-label categorization), GLM-4.6V-Flash achieves the highest individual-model F1-score of 0.6772, suggesting that its pretraining objective and tokenization may yield representations better suited to the fine-grained categorical distinctions required by this subtask.

On the validation set, the ensemble did not provide a clear benefit for Subtasks 2.1 and 2.2. For Subtask 2.1, the ensemble F1-score of 0.7967 sits only marginally above the strongest individual model (0.7923), a difference too small to be meaningful. For Subtask 2.2, the ensemble F1-score of 0.6497 actually falls below the strongest individual model (Qwen 3.5 9B at 0.6552), indicating that majority voting actively hurt performance on this subtask.

This regression can be traced to a class-imbalanced error trade-off. For Subtask 2.1, the ensemble reduces false positives (instances incorrectly predicted as YES) relative to Qwen 3.5 9B alone, from 33 to 31 out of 331 validation instances, explaining its marginal F1 improvement. For Subtask 2.2, however, the ensemble and Qwen 3.5 9B achieve identical overall accuracy (0.753), meaning majority voting does not change the total number of correct predictions, but it only redistributes them. Among the 20 validation instances where the ensemble and Qwen 3.5 9B disagree, the ensemble corrects 7 DIRECT and 3 JUDGEMENTAL errors made by Qwen 3.5 9B alone, while introducing 6 new DIRECT and 4 new JUDGEMENTAL errors. Since JUDGEMENTAL is the minority class (n = 51 vs. 139 for DIRECT) and Macro F1 weights both classes equally, this asymmetric trade produces a net recall loss on JUDGEMENTAL (0.412 $\rightarrow$ 0.392) that outweighs the marginal recall gain on DIRECT (0.878 $\rightarrow$ 0.885), explaining the observed regression despite unchanged overall accuracy.

4.2. Official EXIST 2026 Results

Table 6 reports the official EXIST 2026 evaluation results for our submitted system across all three subtasks. Following the task conventions, Subtask 2.1 uses F1-YES as the primary metric, while Subtasks 2.2 and 2.3 use Macro F1. Task 1 received 213 submissions, Task 2 received 182 submissions, and Task 3 received 183 submissions.

Table 6
Official EXIST 2026 evaluation results.

Task	Rank	ICM-Hard	ICM-HardNorm	F1-YES	Macro F1
Task 2.1 – EN	18	0.3064	0.6556	0.7520	–
Task 2.1 – ES	6	0.3127	0.6593	0.7509	–
Task 2.1 – ALL	10	0.3099	0.6576	0.7514	–
Task 2.2 – EN	14	0.0578	0.5201	–	0.5336
Task 2.2 – ES	5	0.1636	0.5570	–	0.5601
Task 2.2 – ALL	6	0.1110	0.5386	–	0.5469
Task 2.3 – EN	8	−0.2480	0.4473	–	0.5346
Task 2.3 – ES	4	−0.2170	0.4556	–	0.5436
Task 2.3 – ALL	7	−0.2203	0.4543	–	0.5413

The official results indicate a solid and respectable outcome across the entire pipeline, underscoring the strong generalization capacity of our approach. In Subtask 2.1, our model secured the 10th position out of 213 submissions. While the F1-score experienced a slight decrease from 0.7967 on the validation set to 0.7514 on the official test set for this initial stage, this marginal gap reflects a natural performance decay rather than a critical vulnerability. This minor variation could be attributed to inherent statistical fluctuations, subtle edge cases, or potential label ambiguity within the unseen test distribution. Given the cascading nature of the pipeline, such an initial variation would typically be expected to degrade downstream performance. However, our pipeline demonstrated a highly resilient behavior: despite the minor quality loss in the upstream input from Subtask 2.1, the specialized models for the subsequent stages successfully compensated for this variance. This robustness is clearly reflected in the final standings, where our approach advanced further, securing the 6th and 7th positions in Subtasks 2.2 and 2.3, respectively.

4.3. Hardware and Environment

All fine-tuning experiments were conducted on G4 Virtual Machines equipped with NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs providing 96 GB of VRAM. The 96 GB memory capacity was necessary to accommodate 9B-parameter base models loaded in 4-bit quantization alongside adapter parameters and optimizer states.

5. Conclusions

We presented a heterogeneous ensemble of three VLMs fine-tuned via quantized PEFT for the EXIST 2026 Task 2 meme sexism detection challenge. The key findings can be summarized as follows.

First, the exclusive use of base models with high-rank adapters ($r = 128$, $\alpha = 256$) proved essential, while instruction-tuned variants were excluded to preserve clean fine-tuning dynamics. Second, architectural diversity within the ensemble was achieved by pairing Qwen and GLM lineages with QLoRA and QDoRA adapters. This diversity delivered consistent gains over any single model on the validation set, consistent with ensemble diversity theory [7]. Critically, we chose QDoRA for Qwen 3.5 9B specifically to avoid having two Qwen+QLoRA models producing correlated predictions that could dominate the vote. Our analysis of 2-1 disagreements confirmed that the GLM-4.6V-Flash model does not behave as a passive third voter: the two Qwen models do not dominate disagreements, validating the ensemble design. Third, the voting strategy must be adapted to the task structure: majority voting strategy is appropriate for binary and single-label subtasks, while at-least-one voting strategy substantially outperforms majority voting for multi-label categorization by preserving valid but infrequently predicted categories.

Based on the internal validation, the ensemble did not produce a clear improvement over individual models in Subtasks 2.1 and 2.2; however, Subtask 2.3 shows a clear and decisive benefit from the

ensemble approach, driven by the at-least-one voting strategy. The gap between validation and test performance, particularly in the intention and category subtasks, points to the sensitivity of these classifiers to domain shift and motivates further investigation into more robust training regimes.

Several directions remain open for future work. Investigating soft voting, which aggregates model confidence scores rather than hard labels, may further improve ensemble calibration for Subtasks 2.1 and 2.2. Per-category error analysis for Subtask 2.3 would identify which sexism categories are most frequently mispredicted and whether this pattern is consistent across models. Finally, evaluating further rank scaling ($r = 256$) would clarify whether additional adapter capacity yields further gains or encounters diminishing returns on this task.

Declaration on Generative AI

During the preparation of this work, the author(s) used generative AI tools [25, 26, 27] in order to: grammar and spelling check, paraphrase and reword. After using these tools/services, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] C. Fox, C. Cruz, J. Y. Lee, Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media, Computers in Human Behavior 52 (2015) 436–442.

[2] Italiani, et al., Trankiltwice at EXIST 2025, in: Working Notes of CLEF 2025, 2025.

[3] V. Basile, et al., We need to consider disagreement in evaluation, in: Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future (ACL), 2021.

[4] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, M. Aloy-Mayo, P. Rosso, E. García-Arias, D. Spina, Overview of EXIST 2026: Physiological data for multimodal sexism characterization in social media, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), 2026.

[5] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, M. Aloy-Mayo, P. Rosso, E. García-Arias, D. Spina, Overview of EXIST 2026: Physiological data for multimodal sexism characterization in social media (extended overview), in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.

[6] F. Bordes, et al., An introduction to vision-language modeling, 2024. ArXiv:2405.17247.

[7] S. Bian, W. Wang, On diversity and accuracy of homogeneous and heterogeneous ensembles, International Journal of Hybrid Intelligent Systems 4 (2007) 103–128.

[8] F. Rodríguez-Sánchez, J. C. de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of EXIST 2021: sEXism identification in social neTworks, Procesamiento del Lenguaje Natural 67 (2021) 209–221.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems (NeurIPS), volume 30, 2017. ArXiv:1706.03762.

[10] L. Tian, J. R. Trippas, M.-A. Rizoiu, Mario at EXIST 2025: A simple gateway to effective multilingual sexism detection, in: Working Notes of CLEF 2025, 2025.

[11] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations (ICLR), 2021. ArXiv:2106.09685.

[12] S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, M.-H. Chen, DoRA: Weight-decomposed low-rank adaptation, in: International Conference on Machine Learning (ICML), 2024. ArXiv:2402.09353.

[13] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient finetuning of quantized LLMs, in: Advances in Neural Information Processing Systems (NeurIPS), volume 36, 2023. ArXiv:2305.14314.

[14] T. Dettmers, L. Zettlemoyer, The case for 4-bit precision: k-bit inference scaling laws, in: International Conference on Machine Learning (ICML), 2023, pp. 7750–7774.
[15] F. Trad, A. Chehab, To ensemble or not: Assessing majority voting strategies for phishing detection with large language models, 2024. ArXiv:2412.00166.
[16] A. M. Davani, M. Díaz, V. Prabhakaran, Dealing with disagreements: Looking beyond the majority vote in subjective annotations, Transactions of the Association for Computational Linguistics 10 (2022) 92–110.
[17] K. Sechidis, G. Tsoumakas, I. Vlahavas, On the stratification of multi-label data, in: Proceedings of ECML PKDD 2011, Springer, 2011, pp. 145–158.
[18] Qwen Team, Qwen3.5: Towards native multimodal agents, https://qwen.ai/blog?id=qwen3.5, 2026.
[19] V Team, W. Hong, W. Yu, et al., GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. ArXiv:2507.01006.
[20] Mistral AI Team, Ministral 3, 2026. ArXiv:2601.08584.
[21] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations (ICLR), 2019. ArXiv:1711.05101.
[22] I. Loshchilov, F. Hutter, SGDR: Stochastic gradient descent with warm restarts, in: International Conference on Learning Representations (ICLR), 2017. ArXiv:1608.03983.
[23] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9268–9277.
[24] T. Brown, B. Mann, N. Ryder, et al., Language models are few-shot learners, in: Advances in Neural Information Processing Systems (NeurIPS), volume 33, 2020, pp. 1877–1901.
[25] OpenAI, ChatGPT (large language model), https://chatgpt.com, 2026.
[26] Anthropic, Claude Sonnet 4.6 (large language model), https://claude.ai, 2026.
[27] Google, Gemini 3.1 Pro (large language model), https://gemini.google.com, 2026.