Detecting Sexism in Memes with Vision-Language Models for EXIST 2026*

UNAM_GIL Lab at CLEF 2026

Luis A. H. Miranda[1,2,†], Estela Monserrat Arriaga Santana[1,3,†], Aline Flores Reyes[1,5,†], Luis Eduardo Quintero Cuevas[1,6,†], Francisco F. López-Ponce[1,4,†] and Gerardo Sierra[1,†]

[1]*Universidad Nacional Autónoma de México, Instituto de Ingeniería, Grupo de Ingeniería Lingüística, Circuito Escolar, 04510 Mexico City, Mexico*

[2]*Universidad Nacional Autónoma de México, Programa de Posgrado en Psicología, Unidad de Posgrado, Circuito de Posgrados, 04510 Mexico City, Mexico*

[3]*Universidad Nacional Autónoma de México, Facultad de Ciencias, Circuito Exterior, 04510 Mexico City, Mexico*

[4]*Universidad Nacional Autónoma de México, Posgrado en Ciencias e Ingeniería de la Computación, Circuito Escolar 3000, 04510 Mexico City, Mexico*

[5]*Universidad Nacional Autónoma de México, Facultad de Psicología, Avenida Universidad 3004, 04510 Mexico City, Mexico*

[6]*Universidad Nacional Autónoma de México, Facultad de Filosofía y Letras, Circuito Interior, 04510 Mexico City, Mexico*

Abstract

This work describes the participation of UNAM_GIL Lab in the EXIST 2026 challenge, which focuses on automatic sexism detection in memes. The study evaluates three Qwen2.5-VL vision-language systems: Qwen 2.5 72B VL FT, Qwen 2.5 72B VL BASE, and Qwen 2.5 32B VL. The Qwen 2.5 72B VL FT system was fine-tuned with QLoRA, while Qwen 2.5 72B VL BASE and Qwen 2.5 32B VL were used only for direct inference and were not fine-tuned. The systems were evaluated across three subtasks: binary sexism detection, source intention classification, and multi-label sexism categorization. Each model was prompted using task-specific instruction strategies: a panel-based prompt, which simulates multiple evaluative perspectives, and a mirror-test prompt, which asks the model to infer whether the meme reinforces sexism or criticizes it. Results show that the fine-tuned model generally achieved better scores under the official ICM-Hard metric, while Qwen 2.5 72B VL BASE tended to outperform on F1, suggesting that fine-tuning and direct inference optimize different aspects of classification. Our best pipelines ranked 10th in Hard vs Hard ES for Subtask 2.1, 13th in Hard vs Hard ALL for Subtask 2.1, and 14th in Hard vs Hard ES for Subtask 2.1, signaling that multimodal models are a promising approach for sexism detection in memes.

Keywords

Large Language Models, Sexism, Finetuning, Image Classification

1. Introduction

Sexism remains a persistent form of discrimination in digital environments, where harmful content can spread rapidly through multimedia formats such as memes. Automatic sexism detection remains a significant hurdle, particularly when relevant semantic information is embedded within images and is not directly available as processable text. EXIST 2026 addresses this challenge by promoting automatic approaches for sexism identification in multimedia content on social media.

Memes combine visual and textual information to convey meanings, stereotypes, and discriminatory messages. Therefore, multimodal approaches constitute a relevant strategy for subsequent Natural

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

†These authors contributed equally.

luisheml16@comunidad.unam.mx (L. A. H. Miranda); monse_arriaga@ciencias.unam.mx (E. M. Arriaga Santana); aline.flores112@comunidad.unam.mx (A. Flores Reyes); luis_ed@comunidad.unam.mx (L. E. Quintero Cuevas); francisco.lopez.ponce@ciencias.unam.mx (F. F. López-Ponce); gsierram@iingen.unam.mx (G. Sierra)

0000-0003-0280-3854 (L. A. H. Miranda); 0009-0006-5458-2546 (E. M. Arriaga Santana); 0009-0003-0555-5363 (A. Flores Reyes); 0009-0003-6932-9883 (L. E. Quintero Cuevas); 0009-0007-4265-1070 (F. F. López-Ponce); 0000-0002-6724-1090 (G. Sierra)

Language Processing tasks and automatic content analysis. Since computational models require structured representations to efficiently analyze multimedia information, it is necessary to develop systems capable of simultaneously processing different types of information present in this kind of content.

We evaluated three Qwen2.5-VL vision-language systems: Qwen 2.5 72B VL FT, Qwen 2.5 72B VL BASE, and Qwen 2.5 32B VL. Qwen 2.5 72B VL FT followed the complete training and prediction pipeline and was adapted through QLoRA, whereas Qwen 2.5 72B VL BASE and Qwen 2.5 32B VL were evaluated only through direct inference without additional parameter updates. Therefore, the 32B model was not fine-tuned in this work. The objective was to automatically classify whether a meme contained sexist expressions and, if so, identify the category of sexism present in the content. Our experiments explore the capability of specifically fine-tuned models and direct inference systems for automatic sexism identification and categorization tasks in multimedia content within the EXIST 2026 framework.

2. Background

In present-day online environments, the volume of information circulating daily on social media platforms has become overwhelming, and its impact on users' daily lives is undeniable and profound. Due to this massive data flow, automating content moderation emerges as an efficient strategy for the identification and filtering of harmful online behavior, such as sexism.

Evidence gathered from *Bluesky* [1] shows that both proprietary and open-weight models achieve specificity values between 91% and 100% when classifying threats and offensive language, as well as agreement with decisions made by the platform's moderation service. Another study analyzing the *Reddit* ecosystem reports a true negative rate of 92.3% [2]. These results indicate that LLMs can support the processing of large volumes of text while distinguishing non-harmful content with precision. This capability may help reduce operational demands and limit the exposure of moderation teams to harmful material [3].

However, performance of purely linguistic tools decays drastically when confronting policy violations that do not depend on explicit text strings. Audits conducted on rule-violating posts on *Reddit* documented a low true positive rate of 43.1% due to the model's tendency to prioritize keyword-matching over the comprehension of highly complex contexts [2].

Despite the robustness of LLMs in the syntactic and semantic analysis of text, graphical components in contemporary digital communication constitutes their most critical limitation in technical terms. In social networks, harmful content is frequently disguised in hybrid formats (memes, screenshots, or video thumbnails) where the isolated text remains neutral but the visual stimulus that surrounds it transforms the message into violent, fraudulent, or discriminatory content [4]. Hence, the integration of multimodal variants of LLMs represents a fundamental paradigm shift in digital platform governance. By implementing this methodology, the gap toward a real pragmatic understanding of interactions diminishes and contributes to creating a safer and healthier environment in social media platforms.

To address the challenges of modern digital safety, utilizing an advanced open-weight vision-language architecture like Qwen 2.5 72B VL BASE provides a robust and targeted solution for detecting online sexism, misogyny, and complex harmful content. Within online environments, cybersexism frequently propagates through online memes, where derogatory messaging—ranging from gender stereotyping and body shaming to objectification and explicit violence—is deliberately embedded in both image and textual components [5]. Because perpetrators intertwine these hostile narratives through a combined visual-textual matrix. The harmful intent is often heavily masked, allowing the content to easily bypass traditional text-only filters that analyze single components in isolation. Consequently, modern content moderation demands multi-modal safety infrastructure capable of processing spatial layouts and decoding semantic interactions. The deployment of a vision-language model can identify masked patterns of hostility without degrading the original structural integrity of the digital media [6].

2.1. Qwen 2.5 72B VL BASE & QLoRA as an optimal adaptation technique

Qwen 2.5 72B VL BASE is a vision-language model [7]. Built as a multimodal architecture, it is engineered to perceive, reason over, and synthesize both textual and graphic input. Rather than degrading an image by forcing it into a fixed, cropped grid, the model preserves the input's original resolution and aspect ratio. This technical capability enhances performance by allowing the system to accurately map the spatial layout of visual elements and cross-reference them with complex textual layers, such as slang and dynamic internet argots. Consequently, the model excels at exposing hidden ironies, hostile inferences, and masked harassment, where the visual context is what weaponizes the text.

While processing these deep multimodal inputs traditionally demands massive, cost-prohibitive infrastructure, the integration of Quantized Low-Rank Adapters (QLoRA) fundamentally shifts the operational efficiency of the system without sacrificing classification quality [8, 9]. In practice, this efficiency allows localized community servers and independent research teams to adapt rapidly to evolving trends. Safety systems can be retrained locally and on the fly to intercept digital toxicity, achieving a highly responsive, low-latency system that effectively reduces false positives and handles ambiguous cultural content at scale.

The model's visual fidelity and computational efficiency of the parameter adaptation, were the reasons why this technique was selected as a framework for deployment. Choosing Qwen 2.5 72B VL BASE allows us to bypass the severe classification drops that happen when memes are downsampled, capturing background details and text overlays that determine whether an image is benign or harmful. The viability offered by QLoRA meant the research could be executed without expensive computing clusters, enabling retraining rounds.

3. Methodology

3.1. Data

The 2026 edition of the EXIST task introduced two multimedia formats: memes and short-form videos from *TikTok* [10, 11]. Since our participation focused exclusively on the meme-based subtasks, only the Meme Dataset was used in the experiments reported in this paper. This dataset contains 3,984 training images and 1,053 test memes distributed across English and Spanish.

The datasets are provided in a JSON format, as in the example provided in Figure 5. Each instance includes metadata such as language, text automatically extracted from the media, paths, and demographic annotator arrays. A major methodological shift in this edition is the integration of a Human-Centered AI framework, where, beyond standard socio-demographic labels from human annotators, the dataset introduces continuous physiological and behavioral tracking data collected from subjects exposed to the multimedia material. These signals include Eye Tracking (ET), Heart Rate (HR), and Electroencephalography (EEG), which can serve as proxies for visual attention, cognitive load, and emotional arousal when users are exposed to sexist content.

Although these physiological signals provide valuable information, they were not used in the present work because our system was designed as a vision-language classification pipeline focused on meme images and textual content. Incorporating ET, HR, and EEG would require a different multimodal fusion architecture and additional preprocessing steps to align physiological time-series data with each meme instance. For this reason, our experiments focus exclusively on the image and textual components of the meme dataset, leaving the integration of physiological signals for future work.

3.2. Data Overview

The dataset provided for this study comprises 5,037 annotated records balanced across two languages: English (2,518) and Spanish (2,519). The data is divided into two partitions: one for training the model and the other for testing. The training split contains 3,984 entries, of which 50.33% are in English, as shown in Table 1. Conversely, the test split consists of 48.72% English entries. The training dataset

includes labels and information structured according to the three classification subtasks established in the EXIST 2026 challenge guidelines. To provide a clearer and more illustrative representation of the corpus distribution, the following data rely on a multi-label mode approach rather than the sum of individual annotator votes for each meme. This method consolidates the multiple evaluations for a single record by assigning it only the most frequently selected label (the mode).This statistical treatment was applied across all three classification tasks to stablish a single definitive category per entry, allowing us to observe consolidated discourse trends without the distortion of vote multiplicity.

Table 1
Language Distribution in EXIST 2026 Datasets

Dataset Split	Language	Records	%
Training (Train)	Spanish (ES)	1,979	49.67%
	English (EN)	2,005	50.33%
	Total	**3,984**	**100.00%**
Test (Test)	Spanish (ES)	540	51.28%
	English (EN)	513	48.72%
	Total	**1,053**	**100.00%**

3.2.1. Sub-task 2.1: Sexism detection

The first task focuses on identifying whether the content is sexist or not. As shown in Figure 1, the dataset exhibits a slight prevalence of sexist content across both languages. In the English subset, 1,114 records were labeled as YES (sexist) and 891 as NO (non-sexist). Similarly, the Spanish subset contains 1,172 YES labels and 807 NO labels.

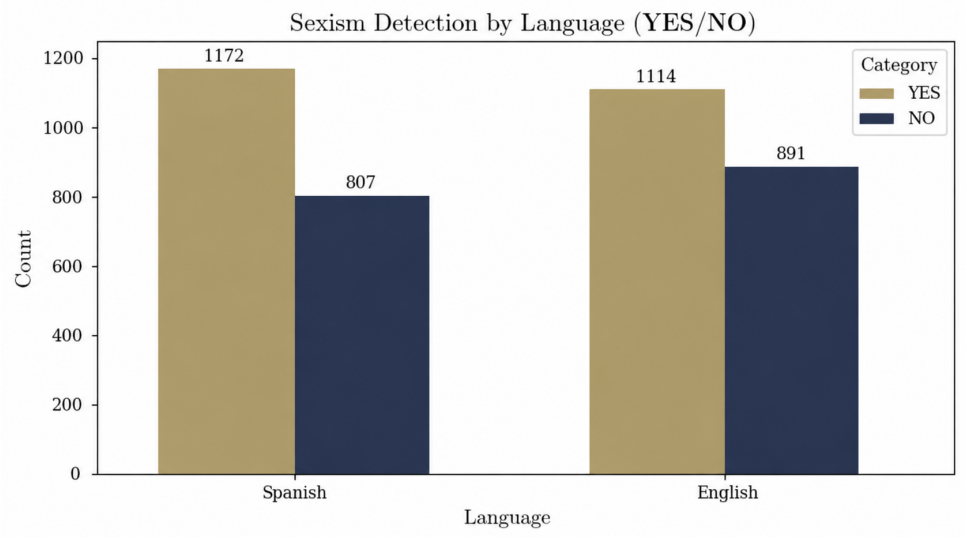

Figure 1: Class Distribution by Language — Sub-Task 2.1.

3.2.2. Subtask 2.2: Source Intention

Subtask 2.2 focuses on identifying the author's intention once a meme has been classified as sexist. Specifically, the task distinguishes whether the meme expresses sexism directly (DIRECT) or criticizes sexist attitudes (JUDGEMENTAL). Non-sexist memes are assigned the label NO. The distribution reveals a notable cross-linguistic nuance: while English-language discourse contains a more balanced

distribution of these categories (742 direct and 371 judgemental), Spanish-language records exhibit a significantly higher prevalence of direct sexism (889 direct and 283 judgemental).

Figure 2 illustrates these proportions, highlighting the higher incidence of direct sexist expressions in Spanish compared to English, which demonstrates a relatively higher concentration of judgmental sexist remarks.

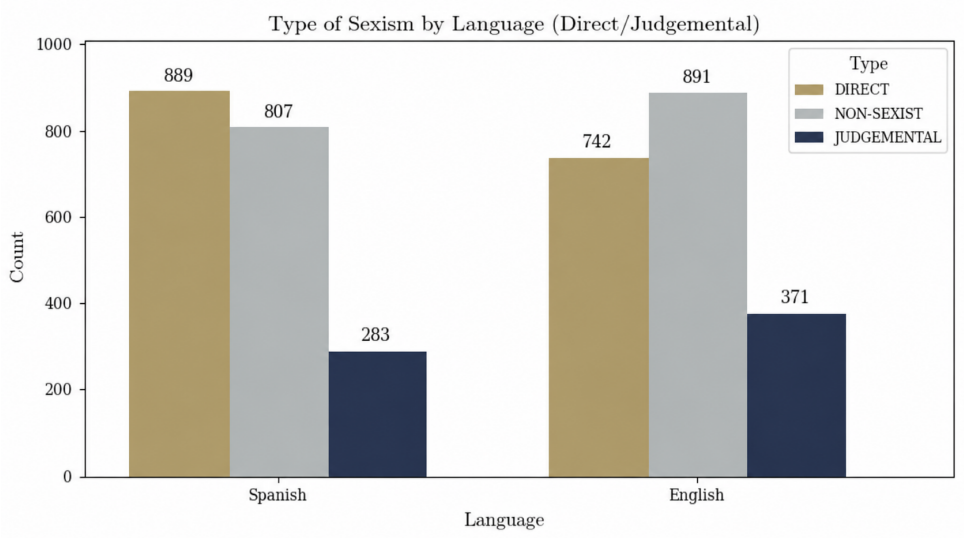

Figure 2: Class Distribution by Language — Sub-Task 2.2.

3.2.3. Subtask 2.3: Sexism Categorization

Subtask 2.3 extends the previous analysis by classifying sexist content into six categories:

- **Ideological and inequality:** Promotes beliefs or social structures that justify the inferiority or subjugation of women.
- **Stereotyping and dominance:** Relies on regressive tropes or asserts power imbalances to marginalize individuals based on gender.
- **Objectification:** Reduces a person to a physical object, body parts, or a commodity for others' gratification.
- **Sexual violence:** Contains explicit threats of assault or the trivialization/glorification of sexual abuse.
- **Misogyny and non-sexual violence:** Direct expressions of hatred, hostility, or non-sexual abuse targeted at individuals due to their gender.
- **Non-sexist:** Instances that, upon deeper inspection, do not contain identifiable sexist content.

These distributions, as shown in Figure 3, highlight the importance of cultural and linguistic nuances in digital discourse. By categorizing instances, we establish a diagnostic baseline that reveals not just the presence of sexism, but also the specific social biases and tropes employed in each corpus. This categorization is essential for understanding how automated systems must adapt to identify diverse and culturally situated manifestations of harmful content.

This dataset provides a balanced and robust foundation for comparative analysis, with linguistic diversity, distinguishable between universal sexist patterns and language-specific rhetorical trends. By categorizing these instances through a three-tiered approach, we establish the necessary empirical basis for evaluating model performance and address the nuances of harmful content across English and Spanish digital discourse.

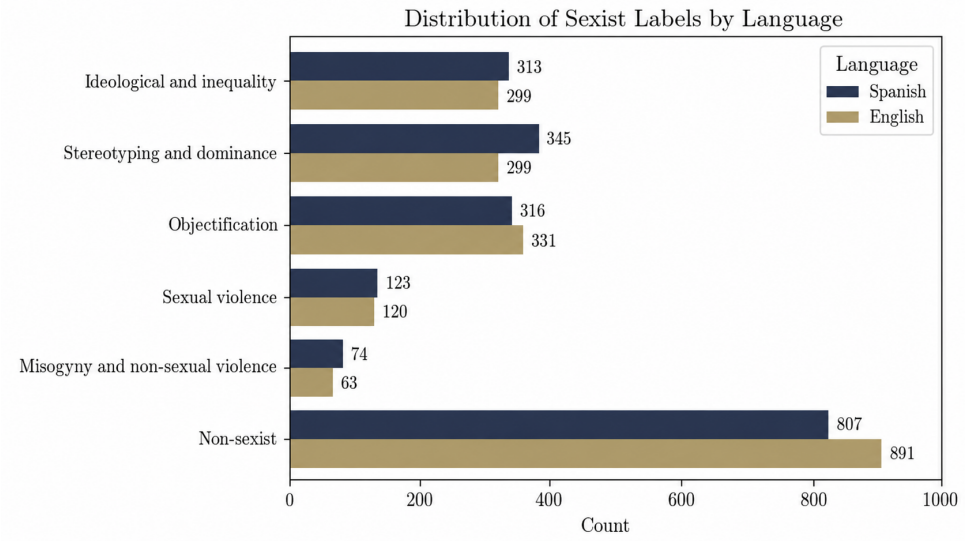

Figure 3: Class Distribution by Language — Sub-Task 2.3.

3.3. System Overview

Our system was designed to evaluate three Qwen-based vision-language systems for the EXIST 2026 meme-based subtasks: Qwen 2.5 72B VL FT, Qwen 2.5 72B VL BASE, and Qwen 2.5 32B VL. The input to all systems consisted of a meme image and task-specific textual information, while the output was a structured JSON object adapted to each subtask: a binary label for Subtask 2.1, a source-intention label for Subtask 2.2, and one or more sexism categories for Subtask 2.3.

Only Qwen 2.5 72B VL FT followed the complete experimental workflow, including data preparation, supervised fine-tuning through QLoRA, adapter validation, upload to the Hugging Face Hub, and prediction on the official test set. In contrast, Qwen 2.5 72B VL BASE and Qwen 2.5 32B VL were evaluated only through the direct inference pipeline, without additional training or parameter updates. This design allowed us to compare the effect of parameter-efficient fine-tuning against the original capabilities of pretrained vision-language models while maintaining the same task formulation, prompt structure, output schema, and evaluation protocol.

The same general prompt structure was used during both fine-tuning and classification. During fine-tuning, the prompt was paired with the expected gold-standard JSON response to train the model. During classification, the same prompt was used without the expected response, allowing the model to generate the predicted JSON output. The complete workflow, including fine-tuning, Hugging Face upload, and prediction, is summarized in Figure 4.

3.4. Fine-Tuning Pipeline

3.4.1. Data Preparation

The fine-tuning pipeline was applied exclusively to Qwen 2.5 72B VL FT. The process began with the validation of the training dataset to ensure the availability of both annotations and meme images. Records containing missing labels or unavailable image files were removed from the training data.

After validation, each meme image was loaded and converted to RGB format to ensure compatibility with the Qwen2.5-VL processor. The annotations corresponding to each EXIST subtask were transformed into structured JSON outputs. For Subtask 2.1, the expected output was a binary YES/NO label. For Subtask 2.2, the expected output was one of three source-intention labels: NO, DIRECT, or JUDGEMENTAL. For Subtask 2.3, the expected output was a set of sexism categories or NO when no sexism category was identified.

Different prompts were designed for Subtasks 2.1, 2.2, and 2.3 while maintaining a consistent

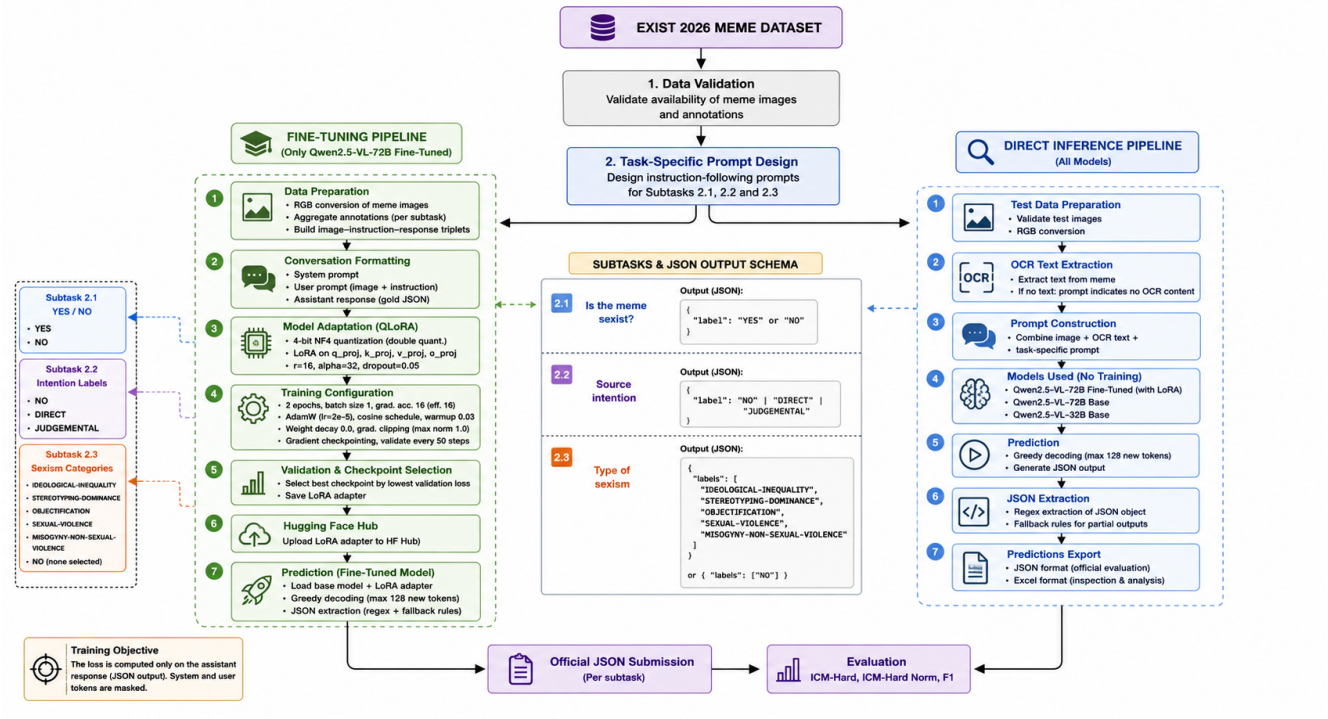

Figure 4: Overview of the experimental pipelines used for Qwen-based meme classification in EXIST 2026.

instruction-following structure. Each training example was represented as an image–instruction–response triplet. The image corresponded to the meme, the instruction contained the task-specific prompt, and the response contained the expected JSON output. These triplets were subsequently converted into the conversational format required by Qwen2.5-VL, where the system and user messages provided the input context and the assistant message represented the target response.

3.4.2. Model Adaptation

Qwen 2.5 72B VL FT was adapted using Quantized Low-Rank Adaptation (QLoRA). For efficiency, the base model was loaded in 4-bit precision using NF4 quantization with double quantization and bfloat16 computation, reducing memory requirements while preserving most of the model's capabilities. Instead of updating all parameters, LoRA adapters were attached to the attention projection layers q_proj, k_proj, v_proj, and o_proj, with rank 16, alpha 32, and dropout 0.05.

Fine-tuning was performed for two epochs with a batch size of 1 per device and gradient accumulation of 16 steps, yielding an effective batch size of 16. We used the fused AdamW optimizer with a learning rate of 2×10^{-5}, a cosine learning-rate schedule, a warmup ratio of 0.03, weight decay of 0.0, and gradient clipping with a maximum norm of 1.0. Gradient checkpointing was enabled to reduce memory usage.

During training, the model learned to generate the expected JSON output associated with each task. The loss was computed only over the assistant response, while the system and user tokens were masked. This strategy encouraged the model to learn the target classification behavior and the required output format simultaneously. Validation was performed every 50 steps, and the checkpoint with the lowest validation loss was selected as the final model. The resulting LoRA adapter was saved and uploaded to the Hugging Face Hub, where it was later retrieved for the prediction stage.

3.5. Direct Inference Pipeline

3.5.1. Data Preparation

The direct inference pipeline was applied to the official test set and was used by all evaluated models. First, the test images were validated to ensure that each file was available and readable. Each meme image was then loaded and converted to RGB format.

To provide additional textual context, Optical Character Recognition (OCR) was applied to extract text embedded within the meme. The extracted text was incorporated into the corresponding task-specific prompt together with the image. If no text was detected, the prompt explicitly indicated that no OCR content was available.

Unlike the fine-tuning stage, no gold-standard response was provided during inference. Instead, the model received only the image and the task instruction and was required to generate the corresponding prediction.

3.5.2. Prediction

Three vision-language systems were evaluated during inference: Qwen 2.5 72B VL FT, Qwen 2.5 72B VL BASE, and Qwen 2.5 32B VL. The fine-tuned system was loaded together with its trained LoRA adapter, whereas the two base systems were used directly without additional adaptation.

For each meme, the image and prompt were submitted to the model. Predictions were produced through greedy decoding with a maximum of 128 newly generated tokens and `do_sample=False`. Since stochastic sampling was disabled, `temperature` and `top_p` were set to `None`, and `top_k` was not used during inference. For reproducibility, a random seed of 42 was set for Python random, NumPy, PyTorch, the stratified data split, and the training procedure. The output was constrained to the JSON schema defined for each subtask: a binary label for Subtask 2.1, a source-intention label for Subtask 2.2, and an array of sexism categories for Subtask 2.3.

The generated outputs were parsed to extract the predicted JSON object. A regular expression was used to identify the JSON object containing the predicted label or labels. When necessary, fallback parsing rules were applied to recover valid labels from partially structured outputs. Finally, the predictions were exported in JSON format for official evaluation and in spreadsheet format for manual inspection and error analysis.

This procedure enabled a direct comparison between the fine-tuned model and the original pretrained models while maintaining the same prompt structure, multimodal inputs, output schema, and evaluation protocol.

3.6. Prompt Design

The following prompts were used to instruct the model during fine-tuning and inference. The same prompt structure was used during both fine-tuning and classification. During fine-tuning, the prompt was paired with the expected gold-standard response to train the model. During classification, the same prompt was used without the expected response, allowing the model to generate the predicted JSON output.

3.6.1. Subtask 2.1

```
System prompt:
You are a panel of 5 diverse judges evaluating memes for sexism.
Consider multiple cultural and personal perspectives.
Output MUST be valid JSON with key: label.
Format: {"label": "YES"} or {"label": "NO"}

User prompt:
MEME ANALYSIS - 5 PERSPECTIVES PANEL
```

```
EXTRACTED TEXT FROM MEME: {OCR_TEXT}

Each judge votes. Majority rules.

JUDGE 1 (Teenager on TikTok): Would this get canceled? -> YES if sexist
JUDGE 2 (HR Manager): Would this create a hostile workplace? -> YES if so
JUDGE 3 (Stand-up Comedian): Is it punching UP or DOWN on gender? -> YES if down
JUDGE 4 (Sociology Professor): Does this reflect systemic inequality? -> YES if so
JUDGE 5 (Your Mom): Would she be disappointed? -> YES if so

FINAL DECISION (majority vote):
Return valid JSON: {"label": "YES"} or {"label": "NO"}
```

3.6.2. Subtask 2.2

```
System prompt:
You are a social media analyst detecting meme intent.
Output MUST be valid JSON with key: label.
Final output: {"label": "NO"} or {"label": "DIRECT"} or {"label": "JUDGEMENTAL"}

User prompt:
THE MIRROR TEST - Who is the audience?

CONTENT:
- Visual: {CAPTION}
- Text: {OCR_TEXT}

STEP 1: IS THIS MEME ABOUT GENDER/SEXISM AT ALL?
- If NO (unrelated to gender, no stereotypes, no sexism references) -> NO
- If YES -> Continue to Step 2

STEP 2: WHAT IS THE INTENTION?

IMAGINE 3 PEOPLE SEEING THIS MEME:

PERSON A (Sexist): "Haha, so true! Women ARE like that!"
- If Person A would GENUINELY agree and feel validated -> DIRECT

PERSON B (Feminist): "This perfectly shows how society treats women!"
- If Person B would share it to CRITICIZE sexism -> JUDGEMENTAL

THE WHISPER TEST:
- "I'm making fun of women" -> DIRECT
- "I'm making fun of sexists" -> JUDGEMENTAL
- "Gender is irrelevant here" -> NO

FINAL: {"label": "NO"} or {"label": "DIRECT"} or {"label": "JUDGEMENTAL"}
```

3.6.3. Subtask 2.3

```
System prompt:
You are a panel of 5 specialists categorizing sexism types.
Output MUST be valid JSON with key: labels (array).
Format: {"labels": ["CAT1", "CAT2"]} or {"labels": ["NO"]}

User prompt:
EXPERT PANEL - SEXISM CATEGORIZATION

EXTRACTED TEXT: {OCR_TEXT}

Each specialist votes on their area. ALL votes are included.
```

```
SPECIALIST 1 (Political Scientist): IDEOLOGICAL-INEQUALITY
"Does this undermine gender equality as a social goal?"
- YES -> Include "IDEOLOGICAL-INEQUALITY"

SPECIALIST 2 (Sociologist): STEREOTYPING-DOMINANCE
"Does this reinforce traditional gender hierarchy?"
- YES -> Include "STEREOTYPING-DOMINANCE"

SPECIALIST 3 (Media Critic): OBJECTIFICATION
"Does this reduce women to decorative or sexual value?"
- YES -> Include "OBJECTIFICATION"

SPECIALIST 4 (Victim Advocate): SEXUAL-VIOLENCE
"Does this normalize or trivialize sexual assault?"
- YES -> Include "SEXUAL-VIOLENCE"

SPECIALIST 5 (Hate Speech Researcher): MISOGYNY-NON-SEXUAL-VIOLENCE
"Does this express hatred or advocate violence toward women?"
- YES -> Include "MISOGYNY-NON-SEXUAL-VIOLENCE"

NO SPECIALIST FLAGS -> {"labels": ["NO"]}

PANEL REPORT: {"labels": ["..."]}
```

The persona-based prompts were designed as heuristic devices to introduce multiple evaluative perspectives into the classification process. However, these personas should not be interpreted as stable or universal social categories. Roles such as a teenager on TikTok, an HR manager, a sociology professor, or a family figure may carry different cultural meanings across regions such as Mexico, Spain, and the United States. Therefore, the prompts were used to operationalize perspective diversity rather than to represent fixed demographic or cultural identities. This limitation should be considered when interpreting the results, since culturally situated references may influence how the model frames sexist content.

3.7. Evaluation metrics

Following the competition guidelines, all submitted systems were assessed under the Hard vs Hard evaluation setting, where both predictions and reference labels are represented as hard classifications.

The official metrics used in the evaluation were ICM-Hard, ICM-Hard Norm, and F1-score. The ICM-Hard metric measures the agreement between the predicted and reference labels while taking into account the structure of the classification task. ICM-Hard Norm corresponds to a normalized version of this metric, allowing comparisons across different experimental settings. Additionally, the F1-score was used to evaluate the balance between precision and recall, providing an overall measure of classification performance.

Results were reported separately for the multilingual dataset (ALL) as well as for the Spanish (ES) and English (EN) subsets. In addition, the official ranking provided by the organizers was used to compare the proposed systems against other participating approaches.

4. Results

Table 2 summarizes the official leaderboard context for the Hard vs Hard setting. The official ranking is ordered according to the ICM-Hard metric. For each subtask and language setting, the table reports the top three systems, the organizer baselines, the last-ranked system, and the best GIL-UNAM submission.

Table 2: Official leaderboard context for the EXIST 2026 Hard vs Hard evaluation setting. Rankings are ordered by ICM-Hard.

Subtask	Setting	Top 3 systems	Organizer baselines	Last-ranked system	Best GIL-UNAM system
2.1	ALL	DDAIUNICC_1 (1, 0.4693); DDAIUNICC_2 (2, 0.4212); Ordantis_2 (3, 0.4079)	Majority (202, -0.4038); Minority (214, -0.6468)	Minority (214, -0.6468)	Qwen 2.5 72B VL FT (13, 0.2849)
2.1	ES	DDAIUNICC_1 (1, 0.4107); Ordantis_2 (2, 0.3469); Ordantis_1 (3, 0.3459)	Majority (198, -0.4001); Minority (211, -0.6557)	MRBLACK_1 (214, -0.6557)	Qwen 2.5 72B VL FT (10, 0.2734)
2.1	EN	DDAIUNICC_1 (1, 0.5279); DDAIUNICC_2 (2, 0.5183); MRBLACK_1 (3, 0.4997)	Majority (202, -0.4076); Minority (213, -0.6381)	PolskaGurom_1 (214, -0.6381)	Qwen 2.5 72B VL BASE (19, 0.3046)
2.2	ALL	Ordantis_3 (1, 0.3709); DDAIUNICC_1 (2, 0.3267); DDAIUNICC_2 (3, 0.2444)	Majority (167, -1.0445); Minority (183, -2.0637)	Minority (183, -2.0637)	Qwen 2.5 72B VL FT (97, -0.4785)
2.2	ES	Ordantis_3 (1, 0.3176); DDAIUNICC_1 (2, 0.2995); DDAIUNICC_2 (3, 0.2199)	Majority (173, -1.0504); Minority (183, -2.0866)	Minority (183, -2.0866)	Qwen 2.5 72B VL FT (102, -0.5513)
2.2	EN	Ordantis_3 (1, 0.4241); DDAIUNICC_1 (2, 0.3548); MRBLACK_1 (3, 0.3459)	Majority (168, -1.0385); Minority (183, -2.0410)	Minority (183, -2.0410)	Qwen 2.5 72B VL FT (84, -0.4052)
2.3	ALL	DDAIUNICC_1 (1, 0.0036); DDAIUNICC_2 (2, -0.0819); ELiRF-UPV_3 (3, -0.1069)	Majority (140, -2.0711); Minority (174, -3.3135)	TheAntisexists_2 (184, -4.7287)	Qwen 2.5 72B VL FT (11, -0.4622)
2.3	ES	DDAIUNICC_1 (1, -0.1095); ELiRF-UPV_3 (2, -0.1334); DDAIUNICC_2 (3, -0.1605)	Majority (132, -2.1173); Minority (173, -3.2599)	PolskaGurom_1 (184, -5.1342)	Qwen 2.5 72B VL FT (11, -0.5685)
2.3	EN	MRBLACK_1 (1, 0.1225); DDAIUNICC_1 (2, 0.0842); DDAIUNICC_2 (3, -0.0138)	Majority (144, -2.0015); Minority (175, -3.3506)	TheAntisexists_2 (184, -4.7769)	Qwen 2.5 72B VL FT (12, -0.3797)

Table 3 presents the performance results for Subtask 2.1, which is a binary image classification task. As seen before, every model is evaluated using three metrics: ICM-Hard, ICM-Hard Norm, and F1 score for hard classifications (YES/NO) across three arrangements: ALL, which includes English and Spanish images; ES, which only contains Spanish text in the images; and EN, with English text in the images.

Qwen 2.5 72B VL FT obtained the best ICM-Hard and ICM-Hard Norm scores in the ALL and Spanish settings. In second place, Qwen 2.5 72B VL BASE obtained the best performance for English image classification across the three metrics: ICM-Hard, ICM-Hard Norm, and F1 score. These results suggest that Qwen 2.5 72B VL FT was more competitive according to the official ICM-based metrics, while Qwen 2.5 72B VL BASE showed stronger performance in terms of F1 score.

Table 3: Task 2.1 Performance of GIL-UNAM systems in the EXIST 2026 hard-hard evaluation setting.

Hard vs Hard ALL				
System	**Ranking**	**ICM-Hard**	**ICM-Hard Norm**	**F1 YES**
EXIST2025-test_gold	0	0.9832	1.000	1.000
Qwen 2.5 72B VL FT	13	**0.2849**	**0.6449**	0.7580
Qwen 2.5 72B VL BASE	15	0.2650	0.6384	**0.7671**
Qwen 2.5 32B VL	65	0.0612	0.5311	0.6794
Hard vs Hard ES				
System	**Ranking**	**ICM-Hard**	**ICM-Hard Norm**	**F1 YES**
EXIST2025-test_gold	0	0.9815	1.000	1.000
Qwen 2.5 72B VL FT	10	**0.2734**	**0.6393**	0.7559
Qwen 2.5 72B VL BASE	14	0.2252	0.6147	**0.7572**
Qwen 2.5 32B VL	81	-0.0431	0.4780	0.6245
Hard vs Hard EN				
System	**Ranking**	**ICM-Hard**	**ICM-Hard Norm**	**F1 YES**
EXIST2025-test_gold	0	0.9848	1.000	1.000
Qwen 2.5 72B VL BASE	19	**0.3046**	**0.6547**	**0.7778**
Qwen 2.5 72B VL FT	20	0.2957	0.6501	0.7603
Qwen 2.5 32B VL	51	0.1677	0.5851	0.7303

Table 4 summarizes the results for Subtask 2.2, which addresses source intention for multiple image classification. Across the three subsets, Qwen 2.5 72B VL FT achieved the highest ICM-Hard and ICM-Hard Norm scores among our submitted systems, obtaining the best official ranking in ALL, ES and EN. However, Qwen 2.5 72B VL BASE obtained the highest F1 score in the three settings. This pattern indicates a divergence between the official ICM-based evaluation and the F1 metric: while the fine-tuned model better aligned with the official scoring criterion, the base model showed stronger performance according to F1. This divergence may be related to the different aspects captured by each metric. While F1 measures the balance between precision and recall at the label level, ICM-Hard and ICM-Hard Norm place greater emphasis on the agreement between predicted and reference labels according to the official evaluation framework of the task. As a result, Qwen 2.5 72B VL FT appears to be better aligned with the criteria used by EXIST 2026, whereas Qwen 2.5 72B VL BASE preserves stronger performance according to traditional classification measures. Although a systematic qualitative error analysis was not conducted, the observed divergence between ICM-based metrics and F1 suggests different error profiles across systems. Qwen 2.5 72B VL FT may have learned to better reproduce the annotation structure and expected label combinations of the EXIST 2026 task, which could explain its stronger performance under ICM-Hard and ICM-Hard Norm. However, this task-specific adaptation may also have increased sensitivity to patterns in the training data, leading to less balanced precision and recall in some settings. In contrast, Qwen 2.5 72B VL BASE may have preserved broader multimodal generalization, which could explain its higher F1 scores despite lower official ICM-based rankings. Future work should include a manual inspection of false positives, false negatives, and malformed JSON outputs to better characterize these differences.

Table 4: Task 2.2 Performance of GIL-UNAM systems in the EXIST 2026 hard-hard evaluation setting.

Hard vs Hard ALL				
System	**Ranking**	**ICM-Hard**	**ICM-Hard Norm**	**F1**
EXIST2025-test_gold	0	1.4383	1.0000	1.0000
Qwen 2.5 72B VL FT	97	**-0.4785**	**0.3337**	0.3078
Qwen 2.5 72B VL BASE	109	-0.5259	0.3172	**0.4312**
Qwen 2.5 32B VL	131	-0.6316	0.2804	0.4151
Hard vs Hard ES				
System	**Ranking**	**ICM-Hard**	**ICM-Hard Norm**	**F1**
EXIST2025-test_gold	0	1.4356	1.0000	1.0000
Qwen 2.5 72B VL FT	102	**-0.5513**	**0.3080**	0.2871
Qwen 2.5 72B VL BASE	122	-0.6471	0.2746	**0.4041**
Qwen 2.5 32B VL	145	-0.7795	0.2285	0.3834
Hard vs Hard EN				
System	**Ranking**	**ICM-Hard**	**ICM-Hard Norm**	**F1**
EXIST2025-test_gold	0	1.4409	1.0000	1.0000
Qwen 2.5 72B VL FT	84	**-0.4052**	**0.3594**	0.3264
Qwen 2.5 72B VL BASE	85	-0.4054	0.3593	**0.4588**
Qwen 2.5 32B VL	108	-0.4841	0.3320	0.4460

Table 5 shows the results for Subtask 2.3, which focuses on sexism categorization in memes. Unlike Subtask 2.1, this task goes beyond binary identification and evaluates the ability of the models to assign sexism-related categories to each image. The systems were evaluated under the hard-hard setting using ICM-Hard, ICM-Hard Norm, and F1 score for the ALL, ES, and EN subsets.

Qwen 2.5 72B VL FT obtained the best ICM-Hard and ICM-Hard Norm scores across the three subsets, achieving the highest official ranking among our systems in ALL, ES, and EN. However, Qwen 2.5

72B VL BASE reached the highest F1 score in all three settings. This result follows the same pattern observed in Subtask 2.2: the fine-tuned model was better aligned with the official ICM-based evaluation, whereas the base model showed stronger performance according to F1 score.

For Subtask 2.3, the difference between ICM-based metrics and F1 may also reflect the difficulty of assigning multiple sexism categories to the same meme. Unlike Subtask 2.2, where each instance receives a single intention label, this task requires the model to decide not only whether sexism is present, but also which specific categories co-occur in the image. The higher F1 obtained by the base model suggests that it may have produced category assignments with a better balance between false positives and false negatives. In contrast, the fine-tuned model obtained better official scores, which may indicate a closer adaptation to the annotation structure and the expected label combinations of the task. This reinforces the need to analyze multilabel performance beyond a single global score.

Table 5: Task 2.3 Performance of GIL-UNAM systems in the EXIST 2026 hard-hard evaluation setting.

Hard vs Hard ALL				
System	**Ranking**	**ICM-Hard**	**ICM-Hard Norm**	**F1**
EXIST2025-test_gold	0	2.4100	1.0000	1.0000
Qwen 2.5 72B VL FT	11	**-0.4622**	**0.4041**	0.4689
Qwen 2.5 72B VL BASE	19	-0.6686	0.3613	**0.4903**
Qwen 2.5 32B VL	23	-0.7502	0.3444	0.4859
Hard vs Hard ES				
System	**Ranking**	**ICM-Hard**	**ICM-Hard Norm**	**F1**
EXIST2025-test_gold	0	2.4432	1.0000	1.0000
Qwen 2.5 72B VL FT	11	**-0.5685**	**0.3837**	0.4651
Qwen 2.5 72B VL BASE	18	-0.8004	0.3362	**0.4805**
Qwen 2.5 32B VL	21	-0.8741	0.3211	0.4727
Hard vs Hard EN				
System	**Ranking**	**ICM-Hard**	**ICM-Hard Norm**	**F1**
EXIST2025-test_gold	0	2.3532	1.0000	1.0000
Qwen 2.5 72B VL FT	12	**-0.3797**	**0.4193**	0.4666
Qwen 2.5 72B VL BASE	21	-0.5711	0.3787	**0.4932**
Qwen 2.5 32B VL	26	-0.6636	0.3590	0.4902

5. Conclusion

This paper presented the participation of UNAM_GIL Lab in the EXIST 2026 meme-based subtasks for sexism detection, source intention classification, and sexism categorization. We evaluated three Qwen-based vision-language systems: Qwen 2.5 72B VL FT, Qwen 2.5 72B VL BASE, and Qwen 2.5 32B VL. Qwen 2.5 72B VL FT was adapted through QLoRA and followed a complete training and prediction pipeline, while Qwen 2.5 72B VL BASE and Qwen 2.5 32B VL were used only for direct classification.

Across the three subtasks, Qwen 2.5 72B VL FT generally achieved the best results according to the official ICM-Hard and ICM-Hard Norm metrics, especially in the ALL and ES settings. In Subtask 2.1, Qwen 2.5 72B VL FT ranked 13th in ALL and 10th in ES, while Qwen 2.5 72B VL BASE ranked 19th in EN. In Subtask 2.2, Qwen 2.5 72B VL FT obtained the best official ranking among our systems in all settings, ranking 97th in ALL, 102nd in ES, and 84th in EN. In Subtask 2.3, Qwen 2.5 72B VL FT achieved its strongest relative performance, ranking 11th in ALL, 11th in ES, and 12th in EN.

These rankings suggest that parameter-efficient fine-tuning helped Qwen 2.5 72B VL FT align more closely with the official evaluation criteria. However, Qwen 2.5 72B VL BASE often obtained higher F1

scores, particularly in the English subset and across Subtasks 2.2 and 2.3. This indicates that fine-tuning and direct inference may optimize different aspects of classification performance depending on the metric used.

This discrepancy is one of the most relevant findings of our work. At a technical level, fine-tuning may have helped Qwen 2.5 72B VL FT internalize the expected JSON format, the task-specific label structure, and the annotation logic of the shared task. However, this adaptation may also have made the model more sensitive to the patterns present in the training data and to the specific prompt structure used during supervision. In contrast, Qwen 2.5 72B VL BASE did not receive task-specific training, but it may have preserved a broader generalization capacity, which could explain its stronger F1 performance in several settings.

This behavior may also reflect the interaction between the training data, the prompt design, and the prior biases learned by Qwen 2.5 72B VL BASE during pretraining. While fine-tuning exposes Qwen 2.5 72B VL FT to the specific annotation patterns of the shared task, Qwen 2.5 72B VL BASE relies more heavily on its general multimodal priors. Therefore, the observed differences between ICM-based metrics and F1 may result from the way each system balances task-specific adaptation with its internal pretrained representations.

At the task level, these results also show that evaluating sexism in memes is not only a matter of predicting the correct label, but also of determining which type of agreement with the reference annotations is being rewarded. ICM-based metrics appear to capture alignment with the official structure of the task, whereas F1 reflects a more traditional balance between precision and recall. Therefore, the ranking of systems can vary depending on whether the evaluation emphasizes task-specific agreement or conventional classification performance.

Our results ranked competitively in the shared task, particularly for Subtask 2.3, where Qwen 2.5 72B VL FT achieved the best official ranking among our submitted systems. Overall, the results show that vision-language models are a promising approach for detecting and categorizing sexism in memes, especially when visual and textual information must be interpreted jointly. Furthermore, the results obtained by Qwen 2.5 72B VL FT suggest that parameter-efficient adaptation techniques such as QLoRA can support the deployment of large multimodal models using accessible computational resources. This finding reinforces the feasibility of developing moderation systems capable of adapting to the evolving visual and textual characteristics of online communities without requiring extensive computational infrastructure. At the same time, the differences between ICM-based metrics and F1 highlight the need to evaluate multimodal systems from multiple perspectives, as different evaluation criteria may capture different aspects of model behavior and task alignment. Future work should explore soft-label evaluation, video-based subtasks, and the integration of sensor-based signals to better capture annotator disagreement, user responses, and the multimodal nature of sexist content.

6. Limitations

The main limitation of this work is that our experiments were restricted to the meme-based subtasks, excluding the video subtasks. In addition, we evaluated only hard-label predictions, leaving the soft-label learning-with-disagreement setting for future work. The proposed systems also depended on prompt-based interaction with vision-language models, making the results sensitive to prompt formulation and model decoding behavior. Moreover, the differences observed between ICM-based metrics and F1 suggest that model performance varied depending on the evaluation criterion.

Another limitation is that we did not use the physiological and behavioral signals included in the EXIST 2026 dataset, namely Eye Tracking (ET), Heart Rate (HR), and Electroencephalography (EEG). These signals were excluded because the present system was implemented as a vision-language pipeline centered on image and text classification. Future work should explore the integration of these signals through multimodal fusion strategies, such as late fusion, where predictions derived from visual-textual inputs can be combined with physiological indicators of attention, arousal, or cognitive processing. This could help capture dimensions of user response that are not directly observable from the meme

image or text alone.

We also did not evaluate the systems on memes outside the EXIST 2026 dataset. Therefore, the extent to which the models generalize to out-of-distribution memes, different visual styles, or other cultural contexts remains an open question. Future work should test the proposed systems on external meme corpora to assess their robustness beyond the annotation patterns and visual characteristics of the shared task dataset.

Finally, although fallback parsing rules were used to recover valid predictions from partially structured outputs, we did not quantify the rate of malformed JSON outputs across models. As a result, we cannot directly compare whether Qwen 2.5 72B VL FT produced fewer formatting errors than Qwen 2.5 32B VL or Qwen 2.5 72B VL BASE. Future evaluations should report malformed-output rates, parsing failures, and recovery rates to better understand the reliability of structured generation in this task.

7. Acknowledgments

Luis A. H. Miranda (CVU: 1145972) and Francisco F. López-Ponce (CVU: 2045472) gratefully acknowledge the support of the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI). The authors also acknowledge the support of the Grupo de Ingeniería Lingüística (GIL) and the Universidad Nacional Autónoma de México (UNAM) through PAPIIT project IG400725. ¡Por mi raza, hablará el espíritu!

8. Declarations on generative AI

During the preparation of this work, GPT-5 models were used exclusively for grammar correction and language refinement. After using these tools, the authors reviewed, edited, and approved the final manuscript and take full responsibility for its content.

References

[1] H.-Y. Chou, W. Naveed, S. Zhou, X. Yang, Are open-weight llms ready for social media moderation? a comparative study on bluesky, 2026. URL: https://arxiv.org/abs/2602.05189. arXiv:2602.05189.

[2] M. Kolla, S. Salunkhe, E. Chandrasekharan, K. Saha, Llm-mod: Can large language models assist content moderation?, in: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA '24, Association for Computing Machinery, New York, NY, USA, 2024. URL: https://doi.org/10.1145/3613905.3650828. doi:10.1145/3613905.3650828.

[3] R. Spence, J. DeMarco, Content moderator mental health and associations with coping styles: Replication and extension of previous studies, Behavioral Sciences 15 (2025). URL: https://www.mdpi.com/2076-328X/15/4/487. doi:10.3390/bs15040487.

[4] A. Bonagiri, L. Li, R. Oak, Z. Babar, M. Wojcieszak, A. Chhabra, Towards safer social media platforms: Scalable and performant few-shot harmful content moderation using large language models, 2025. URL: https://arxiv.org/abs/2501.13976. arXiv:2501.13976.

[5] E. Fersini, F. Gasparini, S. Corchs, Detecting sexist meme on the web: A study on textual and visual cues, in: 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), 2019, pp. 226–231. doi:10.1109/ACIIW.2019.8925199.

[6] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, CoRR abs/2005.04790 (2020). URL: https://arxiv.org/abs/2005.04790. arXiv:2005.04790.

[7] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, J. Zhou, Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL: https://arxiv.org/abs/2308.12966. arXiv:2308.12966.

[8] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates,

Inc., 2023, pp. 10088–10115. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf.

[9] X. Zhang, N. Rajabi, K. Duh, P. Koehn, Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA, in: P. Koehn, B. Haddow, T. Kocmi, C. Monz (Eds.), Proceedings of the Eighth Conference on Machine Translation, Association for Computational Linguistics, Singapore, 2023, pp. 468–481. URL: https://aclanthology.org/2023.wmt-1.43/. doi:10.18653/v1/2023.wmt-1.43.

[10] L. Plaza, J. C. 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.

[11] L. Plaza, J. C. 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. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.

Appendix


```
"310001": {
  "id_EXIST": "310001",
  "lang": "es",
  "text": "Soy como la madre de mi hermano ",
  "meme": "310001.jpeg",
  "path_memes": "memes/310001.jpeg",
  "number_annotators": 6,
  "annotators": ["Annotator_888", "Annotator_889", "..."],
  "gender_annotators": ["M", "M", "M", "F", "F", "F"],
  "age_annotators": ["46+", "23-45", "18-22", "46+", "18-22", "23-45"],
  "ethnicities_annotators": ["White or Caucasian", "..."],
  "study_levels_annotators": ["Master's degree", "..."],
  "countries_annotators": ["Italy", "Spain", "Portugal", "..."],
  "split": "TEST-MEME_ES",
  "sensorial": {
    "users": ["ES1", "ES3"],
    "modalities": {
      "ET": {
        "by_user": {
          "ES1": {
            "reaction_time": 8066.0,
            "3d_eye_states_pupil diameter left [mm]_mean": 2.97,
            "fixations_count": 25.0,
            "saccades_count": 24.0,
            "...": "..."
          },
          "ES3": { "..." }
        }
      },
      "HR": {
        "by_user": {
          "ES1": { "garmin_hr_mean": 71.625, "garmin_hr_std": 0.9922, "...": "..." },
          "ES3": { "garmin_hr_mean": 69.3125, "..." }
        }
      },
      "EEG": {
        "by_user": {
          "ES1": {
            "EXG_Channel_0_mean": -7.3333,
            "EXG_Channel_0_Delta_power": 0.3171,
            "EXG_Channel_0_Beta_power": 1.0865,
            "...": "...",
            "EXG_Channel_15_mean": -1.3866,
            "EXG_Channel_15_Gamma_power": 1.7579
          },
          "ES3": { "..." }
        }
      }
    }
  }
}
```


Figure 5: Abbreviated JSON test instance configuration for the EXIST 2026 data analysis framework, showcasing text attributes, anonymized evaluator demographics, and raw values for the multi-user `sensorial` recording block.