LLM-Based Zero-Shot Pipeline with Physiological Sensor Fusion for Sexism Detection in Memes

Notebook for the EXIST Lab at CLEF 2026

Aryan[1,*], Somnath Banerjee[2]

[1]National Institute of Technology, Agartala, India
[2]Institute of Computer Science, University of Tartu, Estonia

Abstract

This paper addresses automatic sexism detection and characterisation in meme images under a Learning with Disagreement evaluation framework, as part of the EXIST 2026 shared task at CLEF 2026. A fully local zero-shot pipeline is proposed, built around large language models operating on consumer hardware without any cloud services, and applied to the meme subtasks covering binary sexism identification, source intention detection, and multi-label sexism categorisation. Three progressively richer configurations are investigated: a standard prompt approach on extracted meme text, an enhanced version with Chain-of-Thought reasoning and annotator demographic context, and a novel physiological sensor fusion mechanism that post-processes soft confidence scores using eye-tracking, heart rate, and EEG signals recorded from real human subjects, aggregated into an uncertainty-weighted arousal score. Results show that Chain-of-Thought prompting with demographic context improves soft label calibration over the standard configuration, while sensor fusion achieves the best confidence alignment on sexism identification and the strongest multi-label categorisation ranking, but introduces a directional bias on intention detection identified as the primary failure mode. These findings establish zero-shot prompting on extracted meme text as a viable approach to the Learning with Disagreement setting, and demonstrate that grounding model confidence in human physiological responses is a promising direction for human-centered sexism detection.

Keywords

sexism detection, meme analysis, large language models, zero-shot classification, CoT prompting, physiological sensor fusion, learning with disagreement (LeWiDi)

1. Introduction

This section introduces the EXIST 2026 challenge, the motivation behind our participation, and the research objectives that shaped the design of our pipeline.

Sexism in online media has become an increasingly urgent problem for platforms, policymakers, and researchers alike. Automated detection of sexist content has drawn growing research attention, with studies demonstrating that transformer-based models can achieve strong performance on text-based sexism classification [10, 11, 12]. Memes, however, represent a substantially harder challenge because their meaning rarely comes from the words alone; it emerges from how image, text, irony, cultural references, and implied context interact, making memes substantially harder to classify than plain text [5].

The EXIST@CLEF 2026 (sEXism Identification in Social neTworks) shared task [5, 6] continues a series of evaluation campaigns focused on detecting sexism in social media since 2021 [1, 2, 4]. This edition concentrates entirely on multimedia content, specifically memes and TikTok videos, and introduces a novel feature: physiological and behavioural sensor data recorded from real human subjects who viewed each meme during a controlled lab experiment [8]. These multimodal signals, covering eye tracking (ET), heart rate (HR), and electroencephalography (EEG), offer a Human-Centered AI perspective on how people respond unconsciously to sexist content, independent of what they consciously report.

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany
*Corresponding author.
aryanagrahari215@gmail.com (Aryan); somnath.banerjee@ut.ee (S. Banerjee)
https://kodu.ut.ee/~somnath/ (S. Banerjee)

This paper describes our participation in EXIST@CLEF-2026, addressing exclusively the three meme subtasks: binary sexism identification (2.1), source intention classification (2.2), and multi-label sexism categorisation (2.3).

Research Objectives: Our primary objective was to present a cost-effective, privacy-preserving solution for sexism detection in memes. Rather than relying on a typical cloud-based proprietary setup requiring external API access, we developed a fully local pipeline that ensures data privacy, cost efficiency, and complete control over the processing workflow. This design choice accepts a modest decline in raw performance relative to cloud-scale models in exchange for reproducibility, accessibility, and independence from third-party services. All inference runs entirely on consumer hardware with no data transmitted externally.

Submissions: We submitted three runs forming a principled ablation study:

- **Run 1**: Qwen2.5-VL 7B [17], a vision-language model, with standard prompts applied to extracted meme text only. This configuration serves as the reference point for the ablation.
- **Run 2**: Quantised Llama 3.1 8B [18] with enhanced Chain-of-Thought prompts and annotator demographic metadata. This configuration isolates the contribution of richer prompting and demographic context.
- **Run 3**: Gemma-4B [19] extended with a novel physiological sensor fusion mechanism, where ET, HR, and EEG signals are used to adjust soft confidence scores via a derived arousal score. This configuration isolates the contribution of physiological signal calibration.

The paper is organised as follows. Section 2 reviews related work. Section 3 describes the task and dataset. Section 4 presents the prompt-driven LLM framework. Section 5 describes the experimental setup. Section 6 presents results and analysis. Section 7 concludes.

2. Related Work

This section demonstrates how prior work on sexism detection in social media, learning with disagreement, and physiological signal analysis collectively motivates the design of our three-run pipeline.

2.1. Sexism Detection in Social Media

The EXIST shared task series has played a central role in advancing automated sexism detection, evolving from text-based settings to multimodal and disagreement-aware evaluation.

Automated sexism detection has been studied across a range of content types and languages. Early work demonstrated that fine-tuned transformer models such as BERT achieve strong performance on text-level sexism classification in tweets and social media posts [10, 11]. Samory et al. [12] further showed that the choice of annotation guidelines and the framing of the classification task substantially affect what models learn, motivating the move toward richer, disagreement-aware annotation frameworks. In the meme domain, the challenge is compounded by the visual and cultural context that text alone cannot capture, as demonstrated by recent multimodal approaches combining image and text encoders [9].

The EXIST shared task series has directly shaped evaluation standards for sexism detection. The EXIST 2023 edition [1] established the multi-level classification hierarchy covering binary identification, source intention, and category attribution that is retained in 2026. The EXIST 2024 edition [2] introduced the Learning with Disagreement (LeWiDi) paradigm, which treats annotator disagreement not as label noise to be discarded but as a meaningful signal of genuine ambiguity. This shift required systems to produce calibrated soft probability distributions over label classes rather than single hard predictions, fundamentally changing evaluation from accuracy-based to distribution-aligned metrics such as ICM-Soft [7]. The EXIST 2025 edition [4] extended the scope to memes and TikTok videos, introducing multimodal content where meaning emerges from the interaction of image, text, and cultural context. These editions established that text-only models face a systematic bottleneck on meme content due to the short and often decontextualised nature of extracted meme text.

The EXIST 2026 edition [5, 6] builds on these foundations by introducing physiological and behavioural sensor recordings from real human subjects as an additional data modality. Subtask 2.1 (binary sexism identification), subtask 2.2 (source intention classification), and subtask 2.3 (multi-label sexism categorisation) each require systems to reason not only about linguistic content but also about the soft distribution of human judgements across six annotators with diverse demographic backgrounds. Our approach addresses all three subtasks under this framework, using zero-shot LLM prompting as the classification backbone and physiological signals as an external calibration source.

2.2. Physiological Signals and Sensor-Grounded NLP

Research on physiological signals and perception of offensive content has established that human bodily responses carry information beyond what conscious annotation alone can capture. Alaçam et al. [15] show that eye-tracking fixation patterns and dwell times correlate with the subjective perception of hate speech, demonstrating that gaze behaviour differs measurably when annotators encounter disturbing content. Complementing this, Zhang et al. [14] demonstrate that combining EEG signals with eye-tracking enables word-level neural state classification, with Theta-band power serving as a reliable indicator of emotional engagement during reading. Ogğuz et al. [16] further show that heart rate and ECG signals encode emotion-relevant information that is recoverable through standard machine learning methods. Together, these works establish the scientific basis for our use of reaction time, pupil dilation, fixation count, EEG Theta power, and heart rate variability as the five components of our arousal score.

On the side of sensor-grounded language model inference, Yoon et al. [13] show that physiological sensor readings can be used as visual prompts to ground multimodal LLM outputs, improving alignment between model predictions and human perceptual responses. In the specific context of the EXIST 2026 task, Arcos et al. [8] present the dataset of eye-tracking, heart rate, and EEG recordings collected during meme viewing, confirming that physiological arousal signals correlate with sexism perception in a controlled experimental setting. Our Run 3 sensor fusion mechanism builds directly on these findings, translating the arousal signal into a bounded soft score adjustment that is strongest when the LLM is most uncertain, while leaving confident predictions essentially unchanged.

3. Task and Dataset Description

This section describes the three meme subtasks we participated in and provides a brief overview of the EXIST 2026 Memes Dataset.

3.1. EXIST Subtasks

The evaluation is organised into three subtasks, each targeting a different aspect of sexist content in memes.

EXIST 2026 [5] includes six subtasks across two modalities. We participated exclusively in the three meme subtasks (2.1, 2.2, 2.3).

Subtask 2.1 — Sexism Identification is a binary classification task. Given a meme, the system decides whether it is sexist (*YES*) or not (*NO*).

Subtask 2.2 — Source Intention classifies the communicative intent behind a sexist meme as either *DIRECT* (the meme directly conveys sexism) or *JUDGEMENTAL* (the meme critiques or describes sexism).

Subtask 2.3 — Sexism Categorisation is a multi-label task assigning sexist memes to one or more of five categories: *IDEOLOGICAL-INEQUALITY*, *STEREOTYPING-DOMINANCE*, *OBJECTIFICATION*, *SEXUAL-VIOLENCE*, and *MISOGYNY-NON-SEXUAL-VIOLENCE*.

3.2. Dataset

We used the EXIST 2026 Memes Dataset [5], provided by the EXIST organisers for training and testing across each subtask. The dataset contains more than 5,000 labelled memes in English and Spanish. The training set contains 3,984 memes and the test set 1,053 memes. Language distribution is balanced: 2,005 English and 1,979 Spanish memes in training, and 513 English and 540 Spanish in the test set. We did not employ any other dataset to train our classification models for any subtask.

The dataset is annotated under the Learning with Disagreement (LeWiDi) paradigm [5, 2], which treats annotator disagreement as meaningful signal rather than noise. Each meme is labelled by six independent human annotators. Hard labels are derived by majority vote; memes with an exact three-to-three split (15.4% of training data) are excluded from hard evaluation. Soft labels are the raw vote fractions per class. The training dataset is nearly perfectly balanced across language, annotator gender (50% female, 50% male), and age groups. Of the 3,984 training memes, 50.3% received a majority YES label, 34.3% received NO, and 15.4% were ambiguous.

4. Prompt-Driven LLM Framework

This section presents the prompt-based LLM framework used for the EXIST 2026 subtasks. The approach performs zero-shot classification through structured prompts and deterministic output processing. Three configurations are evaluated: (i) Text-Only Prompt Framework (TOPF), (ii) Reasoning-Enhanced Prompt Framework (REPF), and (iii) Multimodal Prompt Fusion Framework (MPFF).

4.1. Text-Only Prompt Framework (TOPF)

This configuration uses Qwen2.5-VL 7B [17] with standard prompts applied to the textual content extracted from memes, with no image, metadata, or sensor information. The prompts directly instruct the model to predict labels for sexism detection, source intention, and category classification. This setup serves as a basic prompt-based approach without any additional contextual information.

Qwen2.5-VL 7B was managed locally through Ollama [22]. The *temperature* was set to 0.2 for output consistency. Three sequential prediction calls are made per meme (subtasks 2.1, 2.2, then 2.3). Memes predicted as non-sexist skip subtasks 2.2 and 2.3, receiving default labels directly.

4.2. Reasoning-Enhanced Prompt Framework (REPF)

This configuration uses a quantised Llama 3.1 8B [18, 20] model served via Ollama [22], with *temperature* set to 0.2. It extends TOPF with three additions: a structured four-step analytical protocol per subtask (surface analysis, contextual analysis, label determination, and calibrated confidence scoring); explicit calibration anchor examples to ground soft probability estimates; and annotator demographic context (gender, age, ethnicity) injected per meme, aligning with the LeWiDi philosophy [5]. The subtask cascade structure from TOPF is preserved. The detailed prompt design for each subtask is described in Section 4.6.

4.3. Multimodal Prompt Fusion Framework (MPFF)

This configuration uses Gemma-4B GGUF Q4_K_XL [19, 21] served via Ollama [22], at a higher temperature of 0.65–0.7 to allow more varied soft estimates for subsequent calibration. MPFF takes the REPF soft probability outputs as its starting point and post-processes them using a derived arousal score computed from eye-tracking, heart rate, and EEG signals recorded from real human subjects who viewed each meme. This arousal score adjusts the REPF soft probabilities upward or downward using a bounded, uncertainty-weighted formula. The complete prompt augmentation and sensor fusion mechanism are described in Sections 4.6 and 4.7 respectively.

4.4. Subtask Cascade and Output Structure

The three subtasks are organised as a hierarchical prediction pipeline, where the output of one stage constrains the predictions generated at subsequent stages.

Figure 1 illustrates the subtask cascade. Subtask 2.1 runs first and its output label cascades into subtasks 2.2 and 2.3. Memes predicted as non-sexist skip model inference in subtasks 2.2 and 2.3, receiving default labels and zero probabilities directly. This design is both computationally efficient and semantically correct across all three configurations.

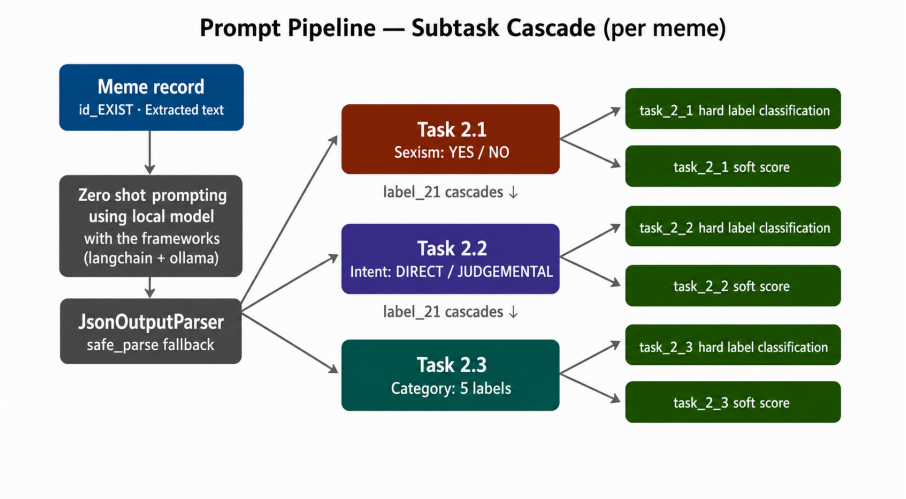

Figure 1: Subtask cascade per meme. Subtask 2.1 runs first; its label cascades into 2.2 and 2.3. Non-sexist memes skip LLM inference in 2.2 and 2.3.

4.5. Shared Prompt System Base

A base-prompt applicable to all configurations (i.e., TOPF, REPF, and MPFF) is designed that includes useful information such as employed LLM's role, task specific definitions and taxonomies, and output requirements. This ensures that any differences among the configurations are due only to their specific additional requirements.

```
You are an expert sociolinguistic annotator for sexism detection in online memes.
You receive TEXT extracted from a meme (no image provided).

CONTEXT: Memes use irony, sarcasm, and cultural references.
Sexist meaning may appear through:
  - Implied stereotypes, sarcastic "compliments"
  - Jokes normalising harassment or violence
  - Dog-whistles and internet slang

SEXISM: Any expression that degrades, stereotypes, objectifies,
or promotes violence against women, including direct insults,
coded language, objectification, and dismissal of gender equality.

NOT SEXIST: Neutral gender mentions, factual statements,
criticism of an individual not based on gender, or
positive statements about women.

OUTPUT: Respond ONLY with a valid structured object.
No explanation outside it.
```

4.6. Prompt Design per Subtask and Configuration

The prompts used in each configuration are adapted to the requirements of the three subtasks while preserving a consistent prediction framework.

4.6.1. TOPF: Standard Classification Prompts

Subtask 2.1:

```
MEME TEXT: """{meme_text}"""
Step 1 - Surface: Identify explicit sexist language.
Step 2 - Context: Consider irony, sarcasm, implied meaning.
Step 3 - Label: YES (sexist) or NO (not sexist).
Step 4 - Confidence: soft_score from 0.0 to 1.0.
Return: {"label": "YES"/"NO", "soft_score": <float>}
```

Subtask 2.2:

```
MEME TEXT: """{meme_text}"""
TASK 2.1: label={label_21}, soft_score={soft_21}
If label_21="NO": return label="NO", all probs=0.0
If label_21="YES": classify as DIRECT or JUDGEMENTAL.
  p_direct + p_judgemental must equal 1.0.
Return: {"label": "DIRECT"/"JUDGEMENTAL"/"NO",
         "p_direct": <f>, "p_judgemental": <f>, "p_no": 0.0}
```

Subtask 2.3:

```
MEME TEXT: """{meme_text}"""
TASK 2.1: label={label_21}, soft_score={soft_21}
If label_21="NO": return labels=[], all five probs=0.0
If label_21="YES": assign independent probabilities (0-1)
  per category. Labels = categories >= 0.5.
Return: {"labels": [...], "p_ideological_inequality": <f>,
  "p_stereotyping_dominance": <f>, "p_objectification": <f>,
  "p_sexual_violence": <f>,
  "p_misogyny_non_sexual_violence": <f>}
```

4.6.2. REPF: Enhanced Chain-of-Thought Prompts

The prompt for REPF extends the TOPF prompts with demographic context and a structured reasoning protocol.

```
[Extends TOPF prompt for all subtasks with:]

DEMOGRAPHICS: gender={g}, age={a}, ethnicity={e}

SUBTASK 2.1 - Additional Calibration Anchors:
  >=0.95: unambiguous sexism
  0.75-0.94: clear but subtle
  0.50-0.74: ambiguous
  <0.25: likely non-sexist

SUBTASK 2.2 - Additional Intent Markers:
  DIRECT: imperatives, slurs, overt objectification.
  JUDGEMENTAL: generalisations, ironic framing.

SUBTASK 2.3 - Additional Category Guidance:
  Use 0.10-0.20 for faintly present categories (not 0.0).
```

4.6.3. MPFF: Sensor Context Addition

The prompt for MPFF appends physiological sensor readings and the derived arousal score to each REPF subtask prompt.

```
SENSOR DATA (avg over {n_subjects} subjects):
  ET:  reaction_time={rt}ms, pupil={pup}mm,
       fixations={fix}, saccades={sac}, blinks={blk}
  HR:  mean={hr_mean}bpm, std={hr_std}
  EEG: delta={d} theta={t} alpha={a} beta={b} gamma={g}
  AROUSAL = {arousal:.4f}
    (0=calm, 0.5=neutral, 1=highly disturbed)
HIGH (>0.65): increase sexist confidence slightly.
LOW  (<0.35): decrease sexist confidence slightly.
MEDIUM: keep score mostly from text.
Do not move a confident prediction by >0.10.
```

4.7. Physiological Sensor Fusion

The physiological responses are integrated into MPFF to calibrate model confidence and account for variations in human perception.

For each meme, seven eye-tracking features, four heart-rate features, and five EEG band-power features are extracted and averaged across participants to form a 16-dimensional feature vector. Five signals — reaction time, heart rate variability [16], pupil dilation [15], EEG Theta power [14], and fixation count — are individually rescaled to $[0, 1]$ using physiologically motivated reference values and averaged into a single arousal score $a \in [0, 1]$, where 0 represents a calm response and 1 represents a highly disturbed response.

The REPF soft probability p_{yes} is subsequently adjusted as follows:

$$\Delta = \alpha \cdot (a - 0.5) \cdot 2 \cdot p_{\text{yes}}(1 - p_{\text{yes}}), \qquad p'_{\text{yes}} = \text{clip}(p_{\text{yes}} + \Delta,\ 0.001,\ 0.999) \tag{1}$$

where $\alpha = 0.35$. The term $p_{\text{yes}}(1 - p_{\text{yes}})$ ensures that the adjustment is largest when the model is uncertain and becomes progressively smaller as confidence increases [13]. For Subtask 2.2, higher arousal values increase the likelihood of the DIRECT class. For Subtask 2.3, the same procedure is applied independently to each category using $\alpha = 0.21$.

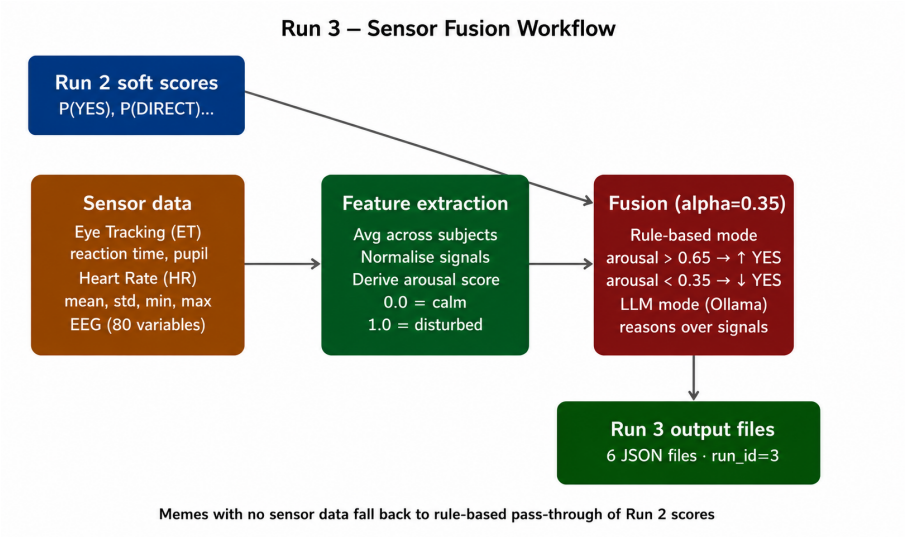

Figure 2: MPFF sensor fusion workflow. ET, HR, and EEG signals are aggregated to compute an arousal score, which is then used to refine REPF soft predictions through a bounded uncertainty-weighted adjustment.

5. Experimental Setup

This section describes the hardware configuration and the process used to select the language models for each run.

5.1. System Setup

All experiments were executed entirely on a single consumer-grade local machine with no cloud computing resources or external inference services involved at any stage. Model serving for all three runs was handled by Ollama [22], a local inference framework that supports GGUF-quantised and standard model formats, enabling consistent structured prediction output within a constrained memory budget. Prompt orchestration across all three subtasks was managed using LangChain [23], which provided a modular chain-based interface for sequential subtask execution and output parsing. Hardware acceleration through the available neural engine and unified GPU cores enabled practical inference throughput on quantised models ranging from 3 to 7 GB in size, without requiring dedicated server-grade GPU infrastructure. All dependencies were managed through a Conda environment running Python 3.10. Figure 3 shows the full technology stack.

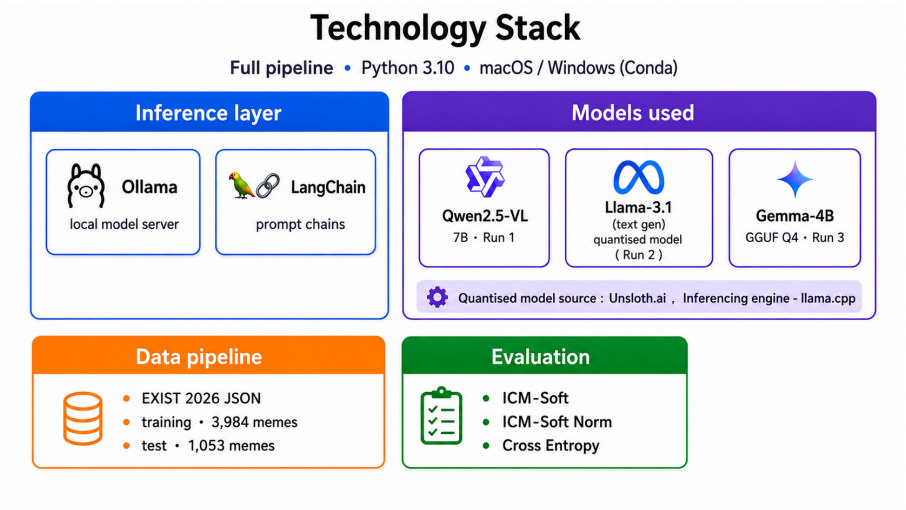

Figure 3: Technology Stack

5.2. LLM Selection

In order to demonstrate how the three language models were selected from a wider set of candidates and the criteria used to evaluate them. To fulfil our research objectives of fully local, privacy-preserving inference, we evaluated a range of open-source instruction-following models against three core requirements: (i) fully local execution within a constrained memory budget; (ii) reliable structured prediction output via Ollama [22]; and (iii) strong instruction-following and reasoning capability for zero-shot classification without any task-specific fine-tuning. Table 1 presents the comparative evaluation of all candidate models.

The three selected models and the rationale for each are as follows:

(i) **Qwen2.5-VL 7B** [17] for Run 1: a vision-language model selected for its strong instruction-following and multimodal reasoning capabilities. Although the framework operates on extracted meme text, meme interpretation often depends on visual context, cultural references, and layout cues, making a vision-language model a more suitable baseline than a text-only model of comparable size.

Table 1
Model comparison for local meme sexism classification. Reasoning quality refers to multi-step instruction following with structured output. Sizes refer to quantised versions.

Model	Size	Quant	Reasoning	Output	Used in
Qwen2.5-VL 7B [17]	7B	No	Good	Reliable	Run 1
Llama 3.1 8B [18]	8B	Q4 [20]	Very Good	Reliable	Run 2
Gemma-4B (Unsloth) [19, 21]	4B	Q4_K_XL	Good	Reliable	Run 3
Mistral 7B [24]	7B	Yes	Good	Moderate	Not selected
LLaVA 13B [25]	13B	No	Moderate	Variable	Not selected
Phi-3 Mini [26]	3.8B	Yes	Moderate	Unreliable	Not selected
DeepSeek-R1 7B [27]	7B	Yes	Very Good	Good	Not selected

(ii) **Llama 3.1 8B** [18] for Run 2: a text generation model with strong instruction-following at 8B scale. Q4 quantisation [20] reduces its footprint to approximately 4.5 GB. Preferred over Mistral 7B [24] (inconsistent output formatting) and DeepSeek-R1 7B [27] (slower inference).

(iii) **Gemma-4B GGUF Q4_K_XL** [19] for Run 3: at approximately 3 GB using the Unsloth quantisation format [21], its smaller footprint allows simultaneous sensor feature computation and LLM inference without memory pressure.

Table 2 summarises the experimental setups for all runs.

Table 2
Side-by-side comparison of the three experimental runs.

	Run 1 (TOPF)	Run 2 (REPF)	Run 3 (MPFF)
Model	Qwen2.5-VL 7B [17]	Llama 3.1 8B Q4 [18, 20]	Gemma-4B Q4_K_XL [19, 21]
Size	∼5 GB	∼4.5 GB	∼3 GB
Prompts	Standard	Chain-of-Thought	Chain-of-thought + sensor context
Temperature	0.2	0.2	0.65–0.7
Input data	Extracted text	Text + demographics	Text + demographics + ET/HR/EEG
Sensor fusion	None	None	Arousal score ($\alpha = 0.35$)

5.3. Baselines

We compare our results against the baselines provided by the EXIST 2026 organisers [5] for both hard-hard (non-probabilistic) and soft-soft (probabilistic) evaluations:

- **Gold**: This baseline provides the best possible reference.
- **Majority class**: classifies all instances as the majority class — a non-informative upper-bound reference for systems that predict without any reasoning.
- **Minority class**: classifies all instances as the minority class — a non-informative lower-bound reference.

5.4. Experimental Algorithm

The overall workflow implemented across the three configurations is outlined in Algorithm

Algorithm 1 EXIST 2026 Three-Run Meme Sexism Pipeline

Require: Test data $\mathcal{T}$ (1,053 memes), Run 2 outputs $\mathcal{R}_2$
Ensure: 18 prediction files (6 per run)

// **Run 1 (TOPF)**
Load Qwen2.5-VL 7B [17] via Ollama [22] (T=0.2)
for each meme $m \in \mathcal{T}$ **do**
    $r_{21} \leftarrow$ LLM(standard prompt, m.text)
    if r_{21}=YES **then** run subtasks 2.2 and 2.3
    else assign default labels and zero probabilities
    end if
end for
Write 6 prediction files (run_id=1)
// **Run 2 (REPF)**
Load Llama 3.1 Q4 [18, 20] via Ollama [22] (T=0.2)
for each meme $m \in \mathcal{T}$ **do**
    $r_{21} \leftarrow$ LLM(enhanced prompt, m.text, m.demographics)
    if r_{21}=YES **then** run subtasks 2.2 and 2.3
    else assign default labels
    end if
end for
Write 6 prediction files (run_id=2) $\rightarrow \mathcal{R}_2$
// **Run 3 (MPFF)**
Load Gemma-4B Q4_K_XL [19, 21] via Ollama [22] (T=0.65–0.7)
for each meme $m \in \mathcal{T}$ **do**
    $\mathbf{f}_m \leftarrow$ ExtractSensorFeatures(m) ▷ avg across subjects
    $a_m \leftarrow$ ComputeArousal($\mathbf{f}_m$)
    $\Delta \leftarrow \alpha(a_m - 0.5) \cdot 2 \cdot p^{(2)}_{\text{yes}}(1 - p^{(2)}_{\text{yes}})$ (Eq. 1)
    $p^{(3)}_{\text{yes}} \leftarrow \text{clip}(p^{(2)}_{\text{yes}} + \Delta,\ 0.001,\ 0.999)$
    Adjust subtasks 2.2 and 2.3 (α'=0.21)
end for
Write 6 prediction files (run_id=3)

5.5. Evaluation Metrics

All runs were evaluated following the official metrics proposed by the EXIST 2026 organisers [5]. Two types of evaluation are performed for each subtask: soft-soft evaluation (probabilistic predictions) and hard-hard evaluation (non-probabilistic predictions). The primary metric for both evaluations is the Information Contrast Measure (ICM) [7], which generalises Pointwise Mutual Information to hierarchical and multi-label settings. A normalised version, ICM-Soft-Norm [7], maps scores to $[0, 1]$, where 1.0 represents perfect agreement with the gold distribution and 0.0 represents the minority-class baseline. CrossEntropy measures the divergence between predicted and gold probability distributions; lower values indicate better calibration. Table 3 summarises the evaluation configuration per subtask.

Table 3
Evaluation settings per subtask. The Hierarchy column shows the class structure used for ICM computation.

Subtask	Type	Hard metric	Soft metric	Hierarchy
2.1	Binary, mono-label	ICM, ICM-Norm	ICM-Soft, CrossEntropy	None
2.2	Multiclass, mono-label	ICM, ICM-Norm	ICM-Soft	YES → {DIRECT, JUDGEMENTAL}, NO
2.3	Multiclass, multi-label	ICM, ICM-Norm	ICM-Soft	YES → {5 categories}, NO

6. Results and Analysis

This section presents the official evaluation results and analyses the three-run progression against the task baselines.

6.1. Official Results

The results presents the complete official scores across all three runs and the task-provided baselines.

Table 4 reports the official scores released by the EXIST 2026 organisers [5]. For reference, the gold upper bound (perfect system) and the majority-class and minority-class baselines are included for each subtask. The ICM-Soft-Norm scale runs from 0.0 (minority-class baseline) to 1.0 (gold), making it straightforward to position each run relative to the task difficulty range. Lower CrossEntropy indicates better probability calibration.

6.2. Results Analysis

The experimental results are analysed with reference to the official task baselines across all three subtasks.

All three configurations exceed the majority-class baseline on subtask 2.1, with REPF achieving the best ICM-Soft-Norm (0.3794) and MPFF achieving the best CrossEntropy (1.3139), suggesting that sensor fusion improves probability calibration while introducing some distributional shift. On subtask 2.2, REPF performs best (ICM-Soft-Norm 0.2033, rank 85/114), while MPFF degrades sharply to 0.0000 due to the arousal-based bias toward DIRECT intent, The primary failure mode of the current sensor fusion design. On subtask 2.3, all three configurations achieve ICM-Soft-Norm of 0.0000, consistent with the difficulty of multi-label soft evaluation observed in prior EXIST editions [2, 4]; however, MPFF achieves the best raw ICM-Soft score (-10.1289) and ranking (76/115), indicating a marginal benefit from per-category arousal adjustment.

Table 4
Official evaluation results with task baselines. ICM-Soft-Norm ranges from 0.0 (minority class) to 1.0 (gold). Lower CrossEntropy is better. Rankings are out of total submitted systems per subtask.

System	Subtask	Rank	ICM-Soft	ICM-Soft-Norm	CrossEntropy
Reference bounds					
Gold (upper bound)	2.1	0/141	3.1107	1.0000	0.5852
Majority class	2.1	140/141	−2.3568	0.1212	4.4015
Minority class	2.1	141/141	−3.5089	0.0000	5.5672
Our submissions — Subtask 2.1					
Run 1 (TOPF / Qwen2.5-VL 7B)	2.1	103/141	−0.7714	0.3760	1.4447
Run 2 (REPF / Llama 3.1 Q4)	2.1	101/141	−0.7501	0.3794	1.4397
Run 3 (MPFF / Gemma-4B+Sensor)	2.1	114/141	−0.8753	0.3593	1.3139
Reference bounds					
Gold (upper bound)	2.2	0/114	4.7018	1.0000	0.9325
Majority class	2.2	112/114	−5.0745	0.0000	5.5565
Minority class	2.2	114/114	−18.9382	0.0000	8.0245
Our submissions — Subtask 2.2					
Run 1 (TOPF / Qwen2.5-VL 7B)	2.2	92/114	−3.1586	0.1614	2.7707
Run 2 (REPF / Llama 3.1 Q4)	2.2	85/114	−2.7905	0.2033	4.0678
Run 3 (MPFF / Gemma-4B+Sensor)	2.2	113/114	−5.2134	0.0000	2.9626
Reference bounds					
Gold (upper bound)	2.3	0/115	9.4343	1.0000	—
Majority class	2.3	74/115	−9.8173	0.0000	—
Minority class	2.3	114/115	−50.0353	0.0000	—
Our submissions — Subtask 2.3					
Run 1 (TOPF / Qwen2.5-VL 7B)	2.3	79/115	−10.5257	0.0000	—
Run 2 (REPF / Llama 3.1 Q4)	2.3	77/115	−10.8377	0.0000	—
Run 3 (MPFF / Gemma-4B+Sensor)	2.3	76/115	−10.1289	0.0000	—

7. Conclusion

This paper presented a fully local, zero-shot LLM pipeline for meme sexism detection submitted to the EXIST 2026 shared task [5, 6], addressing binary sexism identification, source intention detection, and multi-label sexism categorisation through three progressively complex configurations. The results confirm that the Chain-of-Thought prompting with annotator demographic context provides the most consistent improvement over the standard configuration under the Learning with Disagreement evaluation, while physiological sensor fusion demonstrates measurable gains in confidence alignment on sexism identification and multi-label categorisation, despite introducing a directional bias on intention detection that remains the primary failure mode of the current design. Collectively, these findings show that zero-shot prompting on extracted meme text is a viable and resource-efficient approach to human-centred sexism detection, and that grounding model confidence in real human physiological responses is a scientifically principled direction worthy of further investigation.

Limitations and Future Work

The current framework relies solely on extracted meme text and does not incorporate visual information. In addition, the arousal score is derived from manually defined physiological thresholds rather than learned from labelled data, which may limit its generalisability across different content types and populations. Future work will explore the integration of vision-language models, data-driven sensor fusion strategies, fine-tuning with LeWiDi soft labels, and explainable AI techniques to improve performance, calibration, and interpretability.

References

[1] Laura Plaza, Jorge Carrillo-de-Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, and Paolo Rosso. Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2023*, LNCS vol. 14163, Springer, 2023.

[2] Laura Plaza, Jorge Carrillo-de-Albornoz, Víctor Ruiz, Alba Maeso, Berta Chulvi, Paolo Rosso, Enrique Amigó, Julio Gonzalo, Roser Morante, and Damiano Spina. Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes. In L. Goeuriot et al. (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 93–117. Springer, 2024.

[3] Laura Plaza, Jorge Carrillo-de-Albornoz, Víctor Ruiz, Alba Maeso, Berta Chulvi, Paolo Rosso, Enrique Amigó, Julio Gonzalo, Roser Morante, and Damiano Spina. Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes (Extended Overview). *Working Notes of CLEF 2024*, CEUR Working Notes Vol-3740, pages 908–941, 2024.

[4] Laura Plaza, Jorge Carrillo-de-Albornoz, Iván Arcos, Paolo Rosso, Damiano Spina, Enrique Amigó, Julio Gonzalo, and Roser Morante. Overview of EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2025*, LNCS vol. 16089, Springer, 2025.

[5] Laura Plaza, Jorge Carrillo-de-Albornoz, Iván Arcos, María Aloy-Mayo, Paolo Rosso, Enrique García-Arias, and Damiano Spina. Overview of EXIST 2026: Physiological Data for Multimodal Sexism Characterization in Social Media. *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026)*, 2026.

[6] Laura Plaza, Jorge Carrillo-de-Albornoz, Iván Arcos, María Aloy-Mayo, Paolo Rosso, Enrique García-Arias, and Damiano Spina. Overview of EXIST 2026: Physiological Data for Multimodal Sexism Characterization in Social Media (Extended Overview). *CLEF 2026 Working Notes*. Editors: Eva Sánchez Salido, Alberto Barrón-Cedeño, Alba García Seco de Herrera, Sean MacAvaney, Julia Maria Struß, 2026.

[7] Enrique Amigó and Agustín Delgado. Evaluating Extreme Hierarchical Multi-label Classification. In *Proceedings of ACL 2022*, pages 5809–5820, 2022.

[8] Iván Arcos, Paolo Rosso, and Elena Gomis-Vicent. Human-Centered Multimodal Fusion for Sexism Detection in Memes with Eye-Tracking, Heart Rate, and EEG Signals. In *Proceedings of LREC 2026*, http://arxiv.org/abs/2602.23862, 2026.

[9] Iván Arcós and Paolo Rosso. Sexism Identification on TikTok: A Multimodal AI Approach. In *Proceedings of CLEF 2024*, LNCS 14958, pages 61–73, 2024.

[10] Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. An Multilingual BERT for Hate Speech Detection in Multilingual Datasets. In *Proceedings of ECAI 2020/2021*, 2021.

[11] Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of SemEval 2023*, pages 2193–2210, 2023.

[12] Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. "Call me sexist, but...": Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples. In *Proceedings of ICWSM 2021*, 2021.

[13] Hyeon Yoon, Bontu Tesfaye Tolera, Tao Gong, Kimin Lee, and Sung-Ju Lee. By My Eyes: Grounding Multimodal LLMs with Sensor Data via Visual Prompting. In *Proceedings of EMNLP 2024*, pages 2219–2241, 2024.

[14] Yonghao Zhang, Qianhui Li, Shivam Nahata, Talha Jamal, Shuyi Cheng, Gert Cauwenberghs, and Tzyy-Ping Jung. Integrating LLMs, EEG, and Eye-Tracking for Word-Level Neural State Classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:3465–3475,

2024.

[15] Özge Alaçam, Sara Hoeken, and Sina Zarrieß. Eyes Don't Lie: Subjective Hate Annotation and Detection with Gaze. In *Proceedings of EMNLP 2024*, pages 187–205, 2024.

[16] Fatma Elif Oğuz, Aydın Alkan, and Thomas Schöler. Emotion detection from ECG signals with different learning algorithms and automated feature engineering. *Signal, Image and Video Processing*, 17(7):3783–3791, 2023.

[17] Qwen Team, Alibaba Group. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[18] Meta AI. Llama 3.1: Open Foundation and Fine-Tuned Chat Models. https://ai.meta.com/blog/meta-llama-3-1/, 2024.

[19] Gemma Team, Google DeepMind. Gemma: Open Models Based on Gemini Research and Technology. *arXiv preprint arXiv:2403.08295*, 2024.

[20] Georgi Gerganov. GGUF: A File Format for GGML Quantised Models. https://github.com/ggerganov/llama.cpp, 2023.

[21] Daniel Han and Michael Han. Unsloth: Efficient LLM Fine-tuning and Quantisation. https://github.com/unslothai/unsloth, 2024.

[22] Ollama Team. Ollama: Local Large Language Model Serving. https://ollama.com, 2023.

[23] Harrison Chase. LangChain: Building Applications with LLMs through Composability. https://github.com/langchain-ai/langchain, 2022.

[24] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[26] Microsoft Research. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint arXiv:2404.14219*, 2024.

[27] DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*, 2025.