alBERT-319 at EXIST 2026: Calibrating a Multimodal Sexism Classifier With Physiological Signals

Notebook for the EXIST Lab at CLEF 2026

Alberto Rey[1,*]

[1]Universidad Nacional de Educación a Distancia (UNED), 28040 Madrid, Spain

Abstract

This work describes the participation of team alBERT-319 in Subtask 2.1 of EXIST 2026. We propose three classification models, developed incrementally to quantify the impact of integrating physiological signals as input. Aggregated metrics from eye-tracking, heart rate, and EEG, collected during the annotation of each meme, serve as input to a temperature scaling module that dynamically calibrates the predictions of a vision-language classifier. The text-only baseline (XLM-RoBERTa-large over OCR) achieves 0.5867 ICM-Hard-Norm on the development set. The multimodal system (Qwen3-VL-8B-Instruct fine-tuned with QLoRA over image and OCR) yields 0.6607. When physiological calibration is incorporated into Run 3, the result is preserved on the development partition but improves on test ($+0.0135$), correcting overly pessimistic low-confidence predictions, albeit at the expense of introducing minor miscalibration in previously well-calibrated regions. The soft evaluation shows a slight worsening with almost identical magnitude on dev (-0.0085) and test (-0.0081). The per-modality analysis identifies heart rate as the only standalone calibration signal, achieving the best Expected Calibration Error (ECE) in this study on its own. The two multimodal systems achieve top-10 in the official rankings: Run 3 secures the 9$^{\text{th}}$ position out of 214 systems in the hard evaluation, and Run 2 the 7$^{\text{th}}$ out of 141 in the soft evaluation.

Keywords

EXIST 2026, CLEF 2026, sexism identification, memes, multimodal classification, temperature scaling, ECE, QLoRA, physiological signals, eye-tracking, heart rate, EEG

1. Introduction

The spread of sexist content on social media has become a growing concern, particularly regarding memes. These often mask offensive and discriminatory stereotypes against women under a humorous guise, which hinders their detection. Recent advances in vision-language models and physiological signal acquisition have opened new directions for both detecting and calibrating predictions in this type of content. The EXIST campaign series [1] has driven research in this field, both in English and Spanish, since 2021, incorporating memes into its corpus from the 2024 edition onwards.

This work describes the participation of team alBERT-319 in Subtask 2.1 of binary sexism identification over a bilingual (English/Spanish) meme corpus within EXIST 2026 [2, 3]. One of the novelties of the lab in this edition is the incorporation of physiological signals (*sensorial signals*) collected in a process in which a group of subjects view each meme. The collected modalities are eye-tracking, heart rate, and EEG. The training corpus (3,984 memes) available to participants is labeled by six annotators, with a complete demographic profile for each. In contrast, the test set (1,053 memes) is released without labels but maintains the same coverage of physiological data.

In this context, a single hypothesis is proposed and contrasted analytically through the three proposed approaches:

> *The physiological signals aggregated per meme contain useful information about the uncertainty in the classifier's output when evaluating that meme, beyond what can be inferred from the content itself.*

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

*Corresponding author.

arey319@alumno.uned.es (A. Rey)

The analysis is built upon three systems (*runs*) evaluated incrementally. The first (Run 1) is a text-only classifier over the meme's OCR; the second (Run 2) is a multimodal vision-language classifier that incorporates the image; and the third (Run 3) adds on top of Run 2 a temperature scaling calibration module fed by a vector of 10 physiological features. The comparison between Run 2 and Run 3 captures the main contribution of the paper.

The remainder of this paper is structured as follows. Section 2 provides a brief review of the state of the art of the 2024–2025 editions of the lab and contrasts it with the paper's proposal. Section 3 describes the data and the metrics employed. Section 4 presents the three applied systems and the bootstrap-based threshold tuning protocol. Section 5 analyzes the results on the dev partitions together with the two additional comparisons, by physiological modality and by language. Finally, Section 6 closes with the discussion, the limitations, and the reproducibility resources.

2. State of the Art on EXIST 2024–2025 (Subtask 2.1)

EXIST 2024 introduced for the first time sexism identification in memes as a subtask of the lab [4]. The subsequent edition, EXIST 2025 [1], retained this subtask, enabling the categorization of participating systems into three distinct architectural families.

Systems based on VLM (fine-tuned or zero-shot). This family is characterized by using a single vision-language model. The top-performing system in the hard evaluation, CogniCIC [5], reaches 0.6877 using Claude 3.7 in zero-shot (a proprietary multimodal system). In 2025, GrootWatch [6] obtained second place (ICM-Hard-Norm $= 0.6825$) by fine-tuning Qwen2.5-VL-32B with LoRA (they also fine-tuned the 7B variant and evaluated 7B, 32B, and 72B without fine-tuning). I2C-UHU-Altair [7] reported 0.4474 using BLIP2 and Qwen without fine-tuning and guided by a prompt.

Systems based on graphs or prompt chains. TrankilTwice [8] obtains 0.5848 in the hard evaluation with a graph-based representation over the meme and language model prompts applied to each node.

Modular multimodal systems. These systems employ a pre-trained visual encoder and a pre-trained text encoder combined modularly rather than an end-to-end VLM. In EXIST 2024, NICA [9] applied CLIP to combine image and text representations. In 2025, ArcosGPT [10] reaches 0.6627 by combining descriptions generated with BLIP and GPT-4o with a subsequent fusion of ViT and RoBERTa. UMUTeam [11] reaches 0.3496 by combining XLM-RoBERTa with ViT in a multimodal architecture with soft-label learning. CLTL [12] reaches 0.3206 with a hard majority voting ensemble that integrates Swin Transformer V2, RoBERTa, BERTweet, and an SVM with stylometric features.

Taken together, the three families confirm that any competitive proposal on sexism identification in memes requires a multimodal component. In this regard, the 16 systems that surpassed the majority-class baseline are all multimodal, with no exclusively text-based system entering the evaluation.

Conversely, the soft evaluation track remains underexplored, with only five teams reporting ICM-Soft-Norm metrics for Subtask 2.1 in 2025 [1]. The three with the best scores are TrankilTwice (0.5110), surrey-mm-group (0.3865), and I2C-UHU-Altair (0.3624). This margin leaves ample room for assessing calibration mechanisms on the soft track.

The system proposed in this work builds upon the previous observations. Run 1 uses XLM-RoBERTa-large [13] fine-tuned on the meme's OCR with KL loss against the vote distribution. It serves as a text-only reference to measure the gain when visual information is added. Run 2 follows the "VLM + LoRA" family but introduces changes in the implementation. First, the base model is replaced by Qwen3-VL-8B-Instruct [14], and 4-bit quantization is also applied through QLoRA [15], which allows training the model on a single GPU ($1\times$A100 40GB) compared to the $2\times$A100 used by GrootWatch. The main contribution of the paper lies in Run 3, where a temperature scaling calibration module is defined on the outputs of Run 2, whose parameter τ is computed per meme from a vector of 10 aggregated physiological features (eye-tracking, heart rate, and EEG) collected during annotation. The physiological signals are a novelty of the 2026 corpus and constitute a resource still underused in classifier calibration, which motivates their use as the central contribution of this work.

3. Dataset and Metrics

3.1. Corpus

The corpus `EXIST2026_training.json` contains 3,984 memes (1,979 in Spanish and 2,005 in English). Each entry includes the identifier (`id_EXIST`), the language, the transcribed OCR, the path to the JPEG image, the `labels_task2_1` labels from the 6 annotators (each one "YES" or "NO"), the five parallel lists of per-annotator demographic profile (gender, age, ethnicity, education level, and country), and the sensorial block with the modalities available for the meme. The majority-vote distribution on train is 2,003 YES / 1,367 NO / 614 ties. The ties (15.4 %) are discarded from the hard evaluation by task specification.

The test set (`EXIST2026_test_clean.json`) contains 1,053 memes (540 ES and 513 EN) and keeps the rest of the schema except for the labels. Both the demographic profile and the sensorial block are fully available.

The sensorial block contains, per meme and subject, three modalities. Eye-tracking contributes 24 variables (left and right pupil diameter with their mean/std/min/max statistics, counts and durations of fixations, saccades and blinks, and a global reaction time). Heart rate contributes 4 variables (mean/std/min/max from Garmin). EEG contributes 80 variables (16 channels across 5 bands: δ, θ, α, β, γ). The number of subjects per meme is variable (typically 2–4) and the data availability per modality is never complete for all subjects, so the aggregation must tolerate NaN values.

3.2. Train/Dev Split

A subset comprising 10 % of the training data was allocated for development, maintaining stratification across the combination of language and majority label. Memes with tied votes go entirely to training and do not contribute to the stratification. The `random_state` seed is set to 42 and the partition is computed only once to guarantee comparability among the three systems. Table 1 shows the partitions used in the work.

Table 1
Stratified partitions used in the work.

Partition	ES	EN	Total	Use
Original train	1,979	2,005	3,984	Split into dev + training fit
Dev (held-out)	167	170	337	Selection, thresholds, ECE
Training (fit)	1,812	1,835	3,647	Backpropagation
Test	540	513	1,053	Final prediction

3.3. Evaluation Metrics

The official lab metrics [16] are used, computed with the PyEvALL library [17].

- For the hard evaluation: ICM, ICM-Hard-Norm, and F-measure (on the YES class). The main metric for the official ranking is ICM-Hard-Norm. Since Subtask 2.1 is a binary classification, the ICM hierarchy parameter does not need to be specified.
- For the soft evaluation: ICM-Soft, ICM-Soft-Norm, and CrossEntropy. The main ranking metric in this evaluation is ICM-Soft-Norm.

In addition, for the paper's analysis, the Expected Calibration Error (ECE) is computed over the soft probabilities [18]. We use 10 equal-width bins and report both the aggregated ECE and the per-bin breakdown (number of memes n and gap $\overline{p} - \overline{y}$, where $\overline{p}$ is the mean predicted probability in the bin and $\overline{y}$ is the observed YES frequency across 6 annotators). The ECE directly contrasts the calibration quality among the three systems and is threshold-independent. Although ECE is not part of the official EXIST panel evaluation, it serves as the central metric for testing the paper's hypothesis.

4. Systems

The three systems are designed incrementally, ensuring that each architectural addition can be evaluated in isolation. The comparison between Run 1 and Run 2 measures the gain brought by visual information over the OCR text. Building on that, Run 3 adds only the physiological calibration module, which allows isolating its effect without changing any other component of the system, enabling the evaluation of the work's hypothesis.

Run 1. XLM-RoBERTa-large [13] fine-tuned with KL loss against the vote distribution over the meme's OCR.

Run 2. Qwen3-VL-8B-Instruct [14] fine-tuned with LoRA [19] over (image, OCR) in 4-bit quantization [15]. The same objective function as in Run 1 is used.

Run 3. On top of Run 2, a temperature scaling calibration module is added, which adjusts the parameter τ of each meme from a vector of 10 aggregated physiological features.

4.1. Run 1: XLM-RoBERTa-large Over OCR

Run 1 fine-tunes XLM-RoBERTa-large [13] as a binary classifier over the OCR of each meme. The tokenizer input is constructed as "`<lang> ocr_text`" (with `<es>` or `<en>` as language tags placed at the beginning), truncated to 128 tokens. As the objective function, the KL divergence against the observed vote distribution $(p_{\text{YES}}, p_{\text{NO}}) = (n_{\text{YES}}/6,\ n_{\text{NO}}/6)$ is used, computed with the `F.kl_div` function over the `log_softmax` of the two head logits.

Training runs for 4 epochs with batch 16 and AdamW optimizer (`lr`=10^{-5}, `weight_decay`=0.01, `warmup_ratio`=0.1, and gradient clipping at 1.0). The checkpoint is selected by the best ICM-Soft-Norm value on dev. The prediction on test is obtained as $p_{\text{YES}} = \text{softmax}(\text{logits})_{\text{YES}}$.

4.2. Run 2: Qwen3-VL-8B-Instruct Fine-Tuned With QLoRA

Run 2 starts from the Qwen3-VL-8B-Instruct model [14] and loads it with the `bitsandbytes` library in 4-bit quantization (`nf4` format with double quantization and BF16 `compute dtype`) [15]. The fine-tuning is then performed with LoRA [19] only on the language model, leaving the visual tower untrained. The LoRA configuration is `r`=16, `lora_alpha`=32, and `lora_dropout`=0.05, with target modules `q_proj/k_proj/v_proj/o_proj`, which yields approximately 50M trainable parameters. To reduce activation memory consumption, gradient checkpointing is also enabled.

The classifier input combines the meme image, inserted as a visual content block, with the following user text:

```
OCR text: "<ocr>".
Is this meme sexist? Answer YES or NO.
```

Extracting the probability is non-trivial, as the tokens `YES` and `NO` may be subject to varying tokenization rules depending on preceding whitespace. To resolve this, the two possible variants ("`YES`" and "` YES`") are compared over a small set of training examples and the one that best separates the model's scores for YES and NO is selected. The probability is then obtained as:

$$p_{\text{YES}} = \frac{\exp(\ell_{\text{YES}})}{\exp(\ell_{\text{YES}}) + \exp(\ell_{\text{NO}})},$$

where ℓ_{YES} and ℓ_{NO} are the scores assigned by the model to YES and NO at the last input position. The loss is the same KL divergence as in Run 1, applied only to these two options.

The training hyperparameters included three epochs, a batch size of 1 with 16 gradient accumulation steps (effective batch size of 16), and the AdamW optimizer with `lr`=2×10^{-4}, `weight_decay`=0.01, and `warmup_ratio`=0.05. The image is resized to 448 pixels per side before being passed through the visual encoder.

Training was executed on Google Colab Pro+ with an A100 40GB GPU; each epoch takes around 43 minutes, placing the full training at slightly more than two hours. At the end of training, the Run 2 logits over the full training and test sets are persisted, together with their identifiers, to serve as input to the calibration module in Run 3.

4.3. Run 3: Temperature Scaling Calibration Module Over Physiological Signals

Run 3 adds a calibration module fed by the physiological signals associated with each meme. Due to the high dimensionality of the raw sensorial block (108 variables) and the variability in subject coverage, a heuristic selection of 10 representative features was performed. We prioritized global statistics (means and counts) that could be reliably aggregated across different subjects per meme, focusing on core metrics of eye activity, heart rate, and brain activity. This low-dimensional representation also serves as a regularizer for the subsequent MLP training. From the sensorial block of each meme, a fixed vector of 10 components is constructed:

Table 2
Components of the 10-dimensional physiological vector aggregated per meme.

Modality	Feature	Unit
ET	Reaction time	seconds
	Left pupil diameter (mean)	—
	Right pupil diameter (mean)	—
	Number of fixations	—
	Mean fixation duration	nanoseconds
	Number of saccades	—
	Number of blinks	—
HR	Heart rate (mean)	bpm
	Heart rate (std. dev.)	bpm
EEG	Spectral power in α band (16 channels, averaged)	—

The measurements from different subjects are averaged with the `numpy.nanmean` function, ignoring missing values. Each component is then standardized to zero mean and unit variance, replacing missing values with the training mean. The resulting normalization is persisted to disk and reused to obtain the results on dev and test.

4.3.1. Calibration Module

Once the physiological vector for each meme is available, the next step is to use it to adjust the confidence with which the classifier emits each prediction. For this purpose, temperature scaling [18] is applied, with its parameter τ computed per meme from the physiological vector, allowing the probabilities to be softened or sharpened according to the cognitive context of each meme. The calibrated probability is defined as:

$$p_{\text{YES}}^{\text{cal}} = \sigma\left(\frac{\ell_{\text{YES}} - \ell_{\text{NO}}}{\tau(\mathbf{x})}\right),$$

where σ is the sigmoid function and $\tau(\mathbf{x})$ is obtained from a multi-layer perceptron (MLP) with one hidden layer of 32 neurons, ReLU activation, and sigmoid output mapped to the range $[0.5,\ 2.0]$:

$$\tau(\mathbf{x}) = 0.5 + 1.5 \cdot \sigma(\text{MLP}(\mathbf{x})).$$

To prevent the random initialization from biasing the learning, the bias of the MLP's final layer is fixed at $-\log 2$. With this initial value, $\sigma(\cdot) \approx 1/3$ and therefore $\tau \approx 1.0$, rendering Run 3 functionally identical to Run 2 at initialization. Consequently, any learned deviation in τ is strictly driven by the gradient signal from the physiological features.

4.3.2. MLP Training

The MLP is fit on the training fit set (3,647 memes), so that the dev set (337 memes) is entirely held out of training and reserved for model selection and quality checks. The objective function is binary cross-entropy (BCE) between the calibrated probability $p_{\text{YES}}^{\text{cal}}$ and the observed YES vote frequency over the 6 annotators $y = n_{\text{YES}}/6$. The optimizer is AdamW with `lr`=10^{-3}, `weight_decay`=10^{-4}, and batch 64. Training runs for 50 epochs and, at the end of each one, the model is evaluated on dev and the state with the lowest BCE is kept.

It is worth noting that a check has been performed on the calibration module to verify that the parameter τ actually varies across memes and reflects the information in the physiological vector. If all memes received the same τ, Run 3 would be equivalent to Run 2 in probabilities and the paper's hypothesis would lack empirical support. To rule out this scenario, $\text{std}(\tau_{\text{dev}}) > 0.05$ is required on the best checkpoint, a threshold representing 3.3 % of the range $[0.5,\ 2.0]$ where τ varies. On Run 3 with the 10 features, the best checkpoint is obtained at the fifth epoch with $\text{mean}(\tau_{\text{dev}}) = 1.083$ and $\text{std}(\tau_{\text{dev}}) = 0.094$, surpassing the acceptance criterion.

4.4. Threshold Tuning

Soft probabilities require a threshold to generate hard predictions. By default, 0.5 is selected, but the optimal threshold on dev could lie at a different value. This choice is especially relevant for Run 3 because temperature scaling is monotonic and preserves the logit's sign. Formally, $\sigma(\ell/\tau) \geq 0.5 \iff \ell \geq 0$ as long as $\tau > 0$. Consequently, with a fixed value of 0.5, Run 3's hard predictions coincide by definition with Run 2's, and only with a different threshold does the effect of physiological calibration become observable in the hard evaluation. This makes threshold tuning a key step for this comparison.

The tuning is performed in three stages:

1. **Coarse sweep**. Threshold in $[0.40,\ 0.60]$ with step 0.02 (11 points). This search matches the initial version of the protocol and is preserved in the results for traceability.
2. **Fine sweep**. Threshold in $[0.40,\ 0.65]$ with step 0.005 (51 points). The interval is extended because the optimal threshold from the initial search for Runs 2 and 3 lies exactly at the upper boundary of the range.
3. **Bootstrap validation**. To assess robustness against sampling variability, $K = 50$ subsamples of size 280 (approximately $83\,\%$ of dev) are drawn without replacement from the development set. For each subsample, the ICM-Hard-Norm is computed across all fine-sweep thresholds and the local optimum is identified. The final threshold is selected as the one exhibiting the highest *pick rate*, defined as the frequency with which it constitutes the local optimum across the K subsamples; ties are resolved by the highest mean ICM-Hard-Norm.

This validation guards against false optima caused by a single mispredicted meme. The selected thresholds are 0.585 for Run 1 (pick rate 48 %), 0.625 for Run 2 (pick rate 72 %), and 0.610 for Run 3 (pick rate 76 %). In all three cases, the fine sweep improves the ICM-Hard-Norm over the best solution from the coarse sweep: $+0.84$ pp on Run 1, $+1.35$ pp on Run 2, and $+1.61$ pp on Run 3 (on dev).

5. Results and Analysis

Table 3 presents, for the three systems with optimized thresholds, the official metrics and the ECE evaluated on dev. The thresholds are those selected by the bootstrap protocol. The metric ICM-Hard-Norm$_{0.5}$ corresponds to the hard metric applied with the default threshold of 0.5. The ECE is computed using 10 equal-width bins.

The progression from Run 1 to Run 2 corroborates existing literature on meme classification. The vision-language classifier outperforms the text-only system on all three metrics, as a joint result of the change in architecture, modality, and training scheme.

Table 3
Results of the three systems on the development set (337 memes).

System	Thr	ICM-Hard-Norm$_{0.5}$	ICM-Hard-Norm$_{\text{thr}}$	Pick	ICM-Soft-Norm	ECE
Run 1	0.585	0.5572	0.5867	48%	0.4877	0.0392
Run 2	0.625	0.6000	0.6607	72%	0.5198	0.0341
Run 3	0.610	0.6000	0.6607	76%	0.5113	0.0360

The step from Run 2 to Run 3 isolates exclusively the effect of physiological calibration and shows a mixed balance on dev. The ICM-Hard-Norm matches that of Run 2 (0.6607) and the difference in the bootstrap mean (0.6628 vs. 0.6622) falls within its own margin of variability (± 0.016). On the soft branch and on the aggregated ECE, however, Run 3 falls slightly below Run 2 (-0.0085 and $+0.0019$ respectively). These aggregated figures do not capture the full behavior of the system. The per-bin breakdown of the ECE, presented next, reveals that physiological calibration redistributes the error in a non-monotonic way, which is consistent with the paper's hypothesis and justifies a specific analysis.

5.1. ECE Per-Bin Breakdown

Table 4 presents the gap $\Delta = \overline{p} - \overline{y}$ per bin for the three systems. A positive value indicates that the model overestimates the probability and a negative value that it underestimates it.

Table 4
Per-bin ECE gap (10 bins of width 0.1) on dev. n is the number of memes in the bin and the gap is $\overline{p} - \overline{y}$.

	Run 1		Run 2		Run 3	
Center	n	Δ	n	Δ	n	Δ
0.15	20	-0.051	16	-0.109	15	-0.027
0.25	38	-0.039	19	$+0.064$	22	$+0.018$
0.35	41	-0.080	26	$+0.025$	24	$+0.039$
0.45	31	-0.060	42	-0.009	44	-0.002
0.55	33	$+0.041$	46	$+0.099$	51	$+0.090$
0.65	39	$+0.015$	52	$+0.034$	58	$+0.032$
0.75	88	$+0.031$	78	$+0.000$	81	-0.035
0.85	46	$+0.017$	49	-0.013	36	-0.020

Run 1 systematically underestimates in the low-probability bins, with gaps between -0.04 and -0.08 in the ranges 0.15 to 0.45. A purely text-based classifier assigns low probabilities to memes that in reality receive a notable share of YES votes. The incorporation of visual information in Run 2 substantially improves calibration in these bins, with the exception of the 0.15 bin, where the gap reaches -0.109 over 16 memes. That is, the memes that Run 2 considers unlikely to be sexist ($p \approx 0.15$) actually accumulate $26\,\%$ of YES votes, which constitutes the largest per-bin calibration error of this system.

Run 3 reduces this error appreciably, going from -0.109 to -0.027 ($+0.082$ absolute improvement). The mechanism is consistent with the calibration module, since the MLP learns τ values greater than 1 on $81\,\%$ of the memes in this bin (local mean $\tau = 1.079$, in line with the global mean on dev of 1.083), which softens the classifier's output and brings the predicted probability closer to the observed frequency. The interpretation is that the physiological signals collected during viewing contribute relevant information in cases where the classifier is underconfident, since that information allows the module to correct the underestimation.

Run 3's balance is not uniform, since the 0.75 bin, where Run 2 was practically calibrated ($+0.000$ over 78 memes), shifts to -0.035. The MLP applies equivalent softening in this region ($83\,\%$ of memes receive $\tau > 1$, with local mean $\tau = 1.087$) even though the previous calibration did not require it, which moves the gap away from zero. This effect explains why Run 3's aggregated ECE remains slightly

above Run 2's, since the 0.75 bin weighs more in the global computation ($n \approx 80$) than the 0.15 bin ($n \approx 16$) and the aggregated metric does not capture the local improvement. More specifically, the global ECE underestimates the correction provided by the module in the most vulnerable region of the previous classifier.

In this regard, temperature scaling conditioned on physiological signals acts not as a uniform calibration enhancer, but rather as an error redistribution mechanism. This redistribution is nevertheless informative, since it corrects the region where the model errs most at the cost of slightly worsening another region that was already well-calibrated.

The two subsections that follow detail the behavior of the calibration module and form the basis on which the discussion rests. Both analyses are executed non-destructively, since their results are stored separately and do not modify either the Run 3 checkpoints or the files submitted to the panel.

5.2. Per-Modality Analysis

With the goal of identifying which signal contributes most to the calibration, the MLP is trained separately for each physiological modality. In each case, the input vector is restricted to its corresponding components: eye-tracking (7 features, indices 0–6), heart rate (2 features, indices 7–8), and EEG in the α band (1 feature, index 9). The rest of the training process is identical to the system with the 10 features.

Table 5
Per-modality analysis on dev. ICM-Hard-Norm is evaluated at the threshold 0.610 of Run 3 with the 10 features. Values marked with † indicate collapse of the module (std$(\tau_{\text{dev}}) < 0.05$).

Subset	# features	Best epoch	ICM-Hard-Norm$_{0.610}$	ICM-Soft-Norm	ECE	std(τ_{dev})
`all` (10 feat.)	10	5	**0.6607**	0.5113	0.0360	0.094
`et`	7	1	0.6344	0.5198	0.0341	$0.037^{\dagger}$
`hr`	2	50	0.6548	0.5100	**0.0336**	0.086
`eeg`	1	16	0.6344	0.5198	0.0341	$0.041^{\dagger}$

The `et` and `eeg` subsets do not pass the variability criterion, with values std$(\tau_{\text{dev}}) = 0.037$ and 0.041 respectively, both below the threshold 0.05. In these two variants the module converges to an approximately uniform τ on dev, which makes their ECE equal to that of Run 2 without calibration and reduces the ICM-Hard-Norm at threshold 0.610 to 0.6344 in both cases.

The `hr` subset, formed only by the mean and the standard deviation of heart rate, does pass the criterion and produces the best ECE in the work (0.0336), below both Run 2 (0.0341) and Run 3 with the 10 features (0.0360). Its ICM-Hard-Norm at the same threshold is 0.6548, higher than those of the collapsed subsets but lower than that of the full Run 3.

It is worth noting that the best checkpoint of `hr` is reached at the last epoch of the available budget (epoch 50, with the loss still decreasing), while the system with the 10 features converges at the fifth. One possible interpretation is that the vector reduced to 2 features forces the MLP to extract the two available signals more deeply, which slows but stabilizes convergence.

The per-modality analysis identifies heart rate as the signal with the greatest calibration capacity within the physiological vector considered. Eye-tracking and EEG, evaluated in isolation, do not produce a valid module and only contribute in combination with heart rate.

5.3. Per-Language Thresholds

The second analysis examines whether the optimal decision threshold depends on the meme's language. To this end, the three-stage protocol is run independently on the Spanish dev subset (167 memes) and the English dev subset (170 memes), with the bootstrap proportionally adapted to the size of each subset ($K = 50$ subsamples of $139/167$ and $141/170$ respectively, without replacement, seed 42). The per-language thresholds obtained are reported in Table 6.

Table 6
Optimal per-language thresholds on dev and comparison with the global threshold from Table 3. `ES thr` and `EN thr` are the thresholds obtained by the bootstrap protocol applied to each language; ICM-Hard-Norm$_{es/en}$ are the metrics on each subset. The column Δ_{full} reports the difference, on the full dev, between applying the per-language thresholds and the global threshold. Best value per column in bold.

System	ES thr	ICM-Hard-Norm$_{es}$	EN thr	ICM-Hard-Norm$_{en}$	Δ_{full}
Run 1	0.535	0.5353	0.585	0.6366	$+0.0052$
Run 2	0.620	0.6757	0.555	0.6649	$+0.0087$
Run 3	0.605	**0.6844**	0.550	0.6649	$+\mathbf{0.0129}$

The analysis reveals that the per-language optimal thresholds differ systematically. The model's predictions appear slightly more overestimated on Spanish, since the optimal threshold on ES is higher than on EN both for Run 2 (0.620 vs. 0.555) and for Run 3 (0.605 vs. 0.550). Run 1 shows the opposite pattern.

On the Spanish subset, Run 3 reaches the best absolute figure in the work, with an ICM-Hard-Norm of 0.684 at threshold 0.605. Run 3's improvement over Run 2 with per-language thresholds is concentrated exclusively on Spanish ($+0.0087$ on dev_es vs. $+0.0000$ on dev_en), which suggests that physiological calibration is especially useful on the Spanish subcorpus. The per-language bootstrap protocol reaches selection frequencies between $62\,\%$ and $90\,\%$ in the multimodal systems.

The delta on full dev is positive for the three systems, but the per-language thresholds have not been applied to the final submission to the panel. The reason is that, in this particular split of dev_es, there are no memes whose probability falls in the gap between the per-language threshold and the global one ($[0.605,\ 0.610]$ for Run 3 and $[0.620,\ 0.625]$ for Run 2, both of width 0.005), so the improvement on full dev is attributable entirely to the change on the English subset. On test (540 Spanish memes compared to 167 in dev_es), the persistence of the per-language optima cannot be guaranteed; accordingly, under the conservative criterion "each language must strictly improve and the combination must not worsen," the global threshold is retained. The per-language adaptation is left for future investigation.

5.4. Official Results on Test

This subsection presents the results on the test set (1,053 memes) released by the EXIST 2026 organizers after the close of the submission window. Table 7 reports the positions of the three systems on the combined set, and Table 8 breaks them down by language. The rankings are computed over 214 systems on the hard branch and 141 on the soft branch, after filtering out invalid submissions.

Table 7
Official test results, combined evaluation (ALL). CE: CrossEntropy. Best value per column in bold; for rankings, a lower position is better.

Run	Hard rank	ICM-Hard-Norm	F1(YES)	Soft rank	ICM-Soft-Norm	CE
Run 1 (XLM-R)	45 / 214	0.5613	0.7033	22 / 141	0.4658	0.9124
Run 2 (Qwen3-VL + QLoRA)	12 / 214	0.6459	0.7488	**7 / 141**	**0.5135**	0.8479
Run 3 (Run 2 + τ-MLP)	**9 / 214**	**0.6594**	**0.7571**	9 / 141	0.5054	0.8485

Table 8
Official test results broken down by language. For each cell the ranking and the official metric are reported.

Run	Hard ES	Hard EN	Soft ES	Soft EN
Run 1	35 · 0.5472	59 · 0.5749	16 · 0.4570	36 · 0.4725
Run 2	13 · 0.6296	16 · 0.6622	**7 · 0.4991**	**9 · 0.5262**
Run 3	**9 · 0.6412**	**12 · 0.6775**	8 · 0.4902	13 · 0.5188

The two multimodal systems reach top-10 in the official rankings. Run 3 does so in both evaluations (hard #9, soft #9), and Run 2 does so in the soft evaluation (#7), placing at #12 in the hard evaluation. Run 1, purely text-based, sits at notably lower positions (#45 in hard, #22 in soft), which empirically confirms the state-of-the-art observation that systems based only on OCR are not competitive on memes 2.1.

Run 3 outperforms Run 2 in the hard evaluation on test by $+0.0135$ absolute (0.6594 vs. 0.6459), whereas on dev both systems were tied on ICM-Hard-Norm (0.6607 with their respective thresholds). This margin appears when moving from 337 to 1,053 memes, consistent with a set five times larger where the sample variance is smaller and the real differences between systems have a higher probability of being observed.

In the soft evaluation, the calibration module preserves the slight worsening observed on dev, yielding -0.0085 on dev and -0.0081 on test. The pattern reproduces across the two partitions with almost identical direction and magnitude, indicating that the asymmetry between hard and soft evaluations is an effect of the system rather than of the specific partition evaluated.

Run 3's improvement over Run 2 in the hard evaluation is consistent across the two languages, yielding an improvement of $+0.0116$ on the Spanish subset (0.6412 vs. 0.6296) and $+0.0153$ on the English subset (0.6775 vs. 0.6622). The English subset reaches a higher absolute figure than Spanish across the three systems, which suggests that the subtask is relatively easier on EN.

6. Discussion and Conclusions

The hypothesis of the work is tested through the comparison of the three systems, with an asymmetric result between the hard and the soft evaluations. On dev, the ICM-Hard-Norm is identical between Run 2 and Run 3 (0.6607 with their respective thresholds), while the ICM-Soft-Norm shows a slight worsening (-0.0085) and the aggregated ECE does as well ($+0.0019$). On the official test set, the behavior diverges. Run 3 outperforms Run 2 in the hard evaluation by $+0.0135$ absolute (0.6594 vs. 0.6459, position #9 vs. #12 out of 214 systems), while the soft worsening persists with a magnitude almost identical to that of dev (-0.0081 on test, -0.0085 on dev). The asymmetry between hard and soft evaluations reproduces across the two partitions, indicating that the effect of the physiological calibration module concentrates on the binary decision and does not transfer to the full probability distribution.

The per-bin breakdown of the ECE on dev allows a finer reading that is consistent with the results on test. The module corrects Run 2's worst calibration error (bin 0.15, absolute improvement of $+0.082$), but worsens a bin where Run 2 was practically calibrated (bin 0.75, -0.035). Physiological calibration is therefore not a homogeneous improvement but a redistribution of the error. This redistribution adjusts the probabilities near the decision threshold, which explains why the hard evaluation improves and the soft one does not, since the improvement in the binary decision does not automatically translate into a better calibration of the full probability distribution. This pattern is confirmed on test, where the difference in the hard evaluation grows when moving from 337 to 1,053 memes (from a tie on dev to $+0.0135$ on test), while the difference in the soft evaluation remains stable.

In conclusion, two elements suggest that this redistribution of the error is linked to the problem addressed and is not attributable to training. As to the first, the improvement appears exactly where the model was most pessimistic, in memes with low predicted probability and a higher observed YES frequency, which is consistent with the hypothesis that the physiological signals collected during meme viewing carry information about its judgment difficulty. As to the second, the per-modality analysis isolates heart rate (mean and standard deviation) as the primary driver of calibration, since with only these two variables it alone is sufficient to produce the best ECE in this study, while eye-tracking and EEG are not able to do so in isolation and only contribute when combined with it.

6.1. Limitations and Future Work

The following limitations of this study highlight key areas for future research and system refinement.

1. The EXIST 2026 organizers release the aggregated rankings on test but not the individual labels, so the ECE, the per-bin breakdown, and the per-modality analysis can only be computed on dev. The validation of these analyses on test rests on the consistency of the hard/soft asymmetry observed in both partitions.
2. The physiological vectors are aggregated per meme by averaging over the 2–4 available subjects, discarding individual variation. A per-subject approach, instead of a mean vector, could better exploit the heterogeneity of the corpus.
3. A single calibration mechanism is applied, namely temperature scaling. The comparison with alternatives such as Platt scaling, Beta calibration, vector scaling, or histogram binning is left as immediate future work.

An interesting future direction would be to explore non-monotonic calibration transformations conditioned on the physiological vector, capable of improving the soft branch without sacrificing the gain on the hard branch observed in this work.

6.2. Reproducibility

Run 2's training was executed on Google Colab Pro+ on an A100 40GB GPU and took approximately two hours (43 minutes per epoch, 3 epochs). Run 1 was trained on a local CPU (Ryzen 7 with 32 GB of RAM) and also supports low-cost GPUs (L4 or T4) without code changes. The Run 3 calibration module is trained on CPU in under a minute on the fit and development partitions. All seeds in the pipeline (stratified dev/train partition, threshold bootstrap, and random MLP initialization) are set to 42 to guarantee the reproducibility of the random results.

Declaration on Generative AI

During the preparation of this work, the author used Gemini in order to: Text Translation, Paraphrase and reword and Grammar and spelling check. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] L. Plaza, J. Carrillo-De-Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos (Extended Overview), in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 1690–1735.

[2] 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.

[3] 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.

[4] L. Plaza, J. Carrillo-De-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes (Extended Overview), in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 908–941.

[5] T. Alcantara, O. Garcia-Vazquez, H. Calvo, J. E. Valdez-Rodríguez, CogniCIC at EXIST 2025: Identifying Sexist Content in Text and Visual Media using Transformers and Generative AI Models, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 1761–1776.

[6] N. Nowakowski, L. Calogiuri, E. Egyed-Zsigmond, D. Nurbakova, J. Erbani, S. Calabretto, Groot-Watch at EXIST 2025: Automatic Sexism Detection on Social Networks – Classification of Tweets and Memes, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 2058–2083.

[7] M. Guerrero-García, F. Carrillo-García, J. Mata-Vázquez, V. Pachón-Álvarez, I2C-UHU-Altair at EXIST2025: Multimodal Sexism Detection and Classification Using Advanced Vision-Language Models BLIP2 and Qwen, Large Language Models, and Learning with Disagreement Frameworks, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 1978–1993.

[8] P. Italiani, F. Maqbool, D. Gimeno-Gómez, E. Fersini, C.-D. Martínez-Hinarejos, TrankilTwice at EXIST2025: Detecting Sexism in Memes under Multi-Lingual Settings, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 2012–2022.

[9] A. Naebzadeh, M. Nobakhtian, S. Eetemadi, NICA at EXIST CLEF Tasks 2024, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 1122–1134.

[10] I. Arcos, ArcosGPT at EXIST 2025: Identifying Sexism in Memes with Multimodal Deep Learning: Fusing Text and Visual Cues, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 1801–1808.

[11] R. Pan, T. Bernal-Beltrán, J. A. García-Díaz, R. Valencia-García, UMUTeam at EXIST 2025: Multimodal Transformer Architectures and Soft-Label Learning for Sexism Detection, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 2084–2099.

[12] A. Britez, I. Markov, CLTL at EXIST 2025: Identifying Sexist Memes Using an Ensemble of Shallow and Transformer Models, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 1828–1839.

[13] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747.

[14] Qwen Team, Qwen3-VL: A Family of Open Vision-Language Models, https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct, 2025. Technical report; model card.

[15] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient Finetuning of Quantized LLMs, Advances in Neural Information Processing Systems 37 (2023).

[16] E. Amigó, A. Delgado, Evaluating Extreme Hierarchical Multi-label Classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5809–5819.

[17] UNED LENAR Group, PyEvALL: A Library for the Evaluation of NLP Systems, https://github.com/UNEDLENAR/PyEvALL, 2024.

[18] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On Calibration of Modern Neural Networks, in: Proceedings of the 34th International Conference on Machine Learning (ICML), volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 1321–1330. URL: https://proceedings.mlr.press/v70/guo17a.html.

[19] 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, CoRR abs/2106.09685 (2021). URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.