DDAI-UNICC at EXIST-2026 Task 2: Human-Cognitive and Feedback-Driven Prompt Optimization for Sexism Detection

Notebook for the EXIST Lab at CLEF 2026

Anna Llinares[1,†], Marc Franco-Salvador[1,†]

[1]United Nations International Computing Centre (UNICC), Valencia, 46930, Spain

Abstract

This paper describes the participation of the DDAI-UNICC team in the EXIST 2026 shared task on sexism detection. Our best system, HC-ProTeGi-ErPt, is a sensor-enhanced, feedback-driven prompt optimization framework that uses textual gradients aggregated into reusable error patterns to refine prompts for multi-modal sexism detection. The system combines an Evaluator, a Critic, a Summarizer, and a Rewriter, and enriches the optimized prompt with centroid-distance scores derived from sensor-based embeddings of human reactions. It supports binary, multi-class, and multi-label classification settings.

According to the official EXIST 2026 shared-task rankings, we obtained first place in Task 2.1 in English, Spanish, and the global ranking. In Task 2.2, our system ranked second in all three official rankings. In Task 2.3, we achieved first place in the global and Spanish official rankings, and second place in English. These results show the effectiveness of HC-ProTeGi-ErPt for meme-based sexism detection, particularly in binary classification and multi-label categorization.

Our analysis compares the behaviour of the system across tasks, languages, model configurations, and sensor-enhanced variants, showing that summarized error patterns provide a stable optimization signal and that multi-modal information is crucial for identifying implicit, ironic, and context-dependent sexist content.

Keywords

Sexism Detection, Prompt Optimization, Multi-modal Classification, Human-Cognitive Sensor Data

1. Introduction

Online social platforms have become a major space for the production and circulation of textual, visual, and multi-modal content. This has intensified the need for robust systems capable of detecting harmful content and supporting safer online environments. Among these challenges, sexism detection has become a relevant task aimed at identifying content that targets individuals or groups on the basis of gender. However, the task remains difficult due to linguistic ambiguity, irony, implicit meaning, subjective annotation, and, in multi-modal settings, the interaction between textual and visual cues.

Recent evaluation campaigns have contributed to the development of shared benchmarks for this problem. In particular, EXIST (sEXism Identification in Social neTworks) has become a reference benchmark for sexism detection across different languages, modalities, and output spaces (https://nlp.uned.es/exist2026/). Previous editions of EXIST have progressively expanded the task from sexism identification to source intention detection, fine-grained categorization, and multi-modal content analysis [1, 2, 3]. Unlike simpler binary classification datasets, EXIST 2026 includes meme-based scenarios, as well as multi-class and multi-label settings [4, 5]. Its Learning with Disagreement setting further reflects the subjective nature of sexism perception, making the task especially challenging for automatic systems.

Recent work has addressed sexism and hate speech detection using Large Language Models (LLMs), zero-shot and few-shot prompting, Chain-of-Thought reasoning, fine-tuned transformer-based models,

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

*Corresponding author.

†This publication does not reflect the position or views of UNICC and represents only the authors' views.

llinares@unicc.org (A. Llinares); francom@unicc.org (M. Franco-Salvador)

and multi-modal architectures [6, 7, 8, 9, 10]. More recent approaches have further expanded meme-based sexism detection by incorporating richer multimodal and human-centered signals. For instance, Arcos et al. [11] proposed a human-centered multimodal fusion approach that combines meme content with eye-tracking, heart rate, and EEG signals, while Italiani et al. [12] introduced MEMEWEAVER, an inter-meme graph reasoning framework for sexism and misogyny detection. These systems show that strong performance often requires task-specific modeling decisions, external knowledge, visual-textual fusion, or carefully designed prompting strategies. At the same time, LLM-based systems remain highly sensitive to prompt formulation [13], motivating research on automatic prompt optimization. However, existing prompt optimization methods, including feedback-driven approaches such as ProTeGi [14], have been mainly studied in text-only scenarios, and their applicability to multi-modal and multi-label sexism detection remains underexplored.

This work builds on our recently proposed ProTeGi-ErPt framework (Prompt Optimization with Textual Gradients and Error Patterns), submitted to International Conference [15]. The framework supports text-only and multi-modal inputs, as well as binary, multi-class, and multi-label classification settings. It combines an Evaluator, a Critic, a Summarizer, and a Rewriter to iteratively refine prompts using structured error feedback. Unlike previous approaches that directly use complete critic outputs during prompt rewriting, ProTeGi-ErPt summarizes recurring error tendencies into compact and reusable error patterns.

In this paper, introduce HC-ProTeGi-ErPt (Human-Cognitive Prompt Optimization with Textual Gradients and Error Patterns), a human-cognitive sensor-based extension of the framework. This variant incorporates sensor-derived signals as additional contextual information during prompt optimization. Sensor embeddings are structured through clustering, and distance-based scores are provided to the optimization process in order to represent patterns of human response. This allows the framework to explore not only textual and visual evidence, but also complementary signals associated with human reactions to multi-modal content.

Our optimization process combines beam search with Upper Confidence Bound (UCB)-based candidate selection, using macro-F1 as the reward signal during prompt search. Macro-F1 was selected because it provides an interpretable and class-balanced criterion that can be applied consistently across binary, multi-class, and multi-label settings. Although ICM is the official ranking metric, it is less direct as an intermediate optimization signal, especially in multi-label scenarios where it depends on the structure and overlap of predicted label sets. Therefore, ICM and $\mathrm{ICM}_{\mathrm{norm}}$ are used for final evaluation and official comparison, while macro-F1 guides the prompt optimization process. By extending feedback-driven prompt optimization to multi-modal sexism detection and by introducing a human-cognitive sensor-based variant, our work addresses an important limitation of previous prompt optimization methods and aligns the optimization process with the complexity of current benchmarks such as EXIST 2026.

The main contributions of this work are as follows: (i) we build on the recently proposed ProTeGi-ErPt framework [15] and extend it with HC-ProTeGi-ErPt, a human-cognitive sensor-based variant that incorporates distance-based scores derived from sensor embeddings and clustering to study the effect of integrating human-cognitive sensor-derived signals into feedback-driven prompt optimization for meme-based sexism detection; (ii) we adapt the optimization process to use macro-F1 as the reward signal within beam search and UCB-based candidate selection, allowing prompt exploration across binary, multi-class, and multi-label settings; and (iii) we evaluate the resulting framework on the EXIST 2026 Task 2 benchmark in English, Spanish, and the combined multilingual setting. Experimental results show that the strongest configurations are consistently associated with HC-ProTeGi-ErPt, especially in multi-modal meme-based scenarios.

2. Background

The EXIST 2026 benchmark focuses on sexism identification in social networks. The task evaluates whether systems can detect sexist content expressed in different formats, including textual, visual, and sensor-based information. The benchmark follows the Learning with Disagreement paradigm, where

annotator disagreement is explicitly considered instead of being discarded, reflecting the subjective nature of sexism perception.

In this work, we use the EXIST 2026 dataset to evaluate our prompt optimization framework [4, 5]. The dataset contains social media instances in English and Spanish and covers different classification settings, including binary, multi-class, and multi-label sexism detection. Depending on the task, systems must identify whether a given instance is sexist, determine the intention behind the sexist content, or assign one or more fine-grained sexism categories.

The dataset is especially challenging because sexist content can be explicit or implicit, and its interpretation may depend on irony, stereotypes, contextual assumptions, or the interaction between modalities. In addition, sensor-based information introduces an additional signal related to user reactions, allowing the task to explore whether such data can support sexism detection.

During prompt optimization, we use a reduced subset of the training data in order to limit the computational cost of the iterative search process. For binary tasks, we sample 200 instances per class; for multi-class tasks, 100 instances per class; and for multi-label tasks, 500 instances. These samples are used only during prompt optimization, while final results are reported on the official EXIST 2026 test set.

The official evaluation of EXIST 2026 relies on task-specific metrics defined by the organizers [4, 5], including the Information Contrast Model (ICM) [16] and its normalized variant, $\mathrm{ICM}_{\mathrm{norm}}$, which are designed to compare predicted and gold label sets by considering their information content. Although these metrics are used for final reporting, we use macro-F1 as the optimization signal during prompt search, since it provides a more interpretable and class-balanced criterion across binary, multi-class, and multi-label settings.

For comparability, all test-set evaluations of the additional models were computed using EvALL 2.0 (Evaluate ALL 2.0; https://evall.uned.es/), the evaluation framework associated with EXIST and based on PyEvALL (https://pypi.org/project/PyEvALL/). Official shared-task scores are reported as provided by the organizers, while the remaining model scores were obtained with the same evaluation tool on the corresponding test set.

3. Proposed Framework

Figure 1 illustrates the proposed HC-ProTeGi-ErPt framework. We formulate prompt optimization as an iterative search process over natural language instructions. Starting from an initial prompt, the objective is to generate progressively better instructions that improve classification performance across different task formulations, including binary, multi-class, and multi-label settings.

The framework follows the intuition behind ProTeGi [14], where prompt refinement is guided by textual feedback derived from model errors. Instead of updating differentiable parameters, the search is carried out directly in the space of natural language prompts. Each optimization step therefore corresponds to a semantic modification of the current instruction, informed by the types of mistakes made by the model.

Once the best optimized prompt is obtained, we extend it with sensor-derived information. In this stage, sensor data associated with human reactions are first encoded into embedding representations using a neural encoder. These embeddings are then grouped through K-means clustering, producing clusters that represent different patterns of human response. For each instance, we compute centroid-distance scores that measure its relation to the learned cluster centroids. These scores are finally added to the optimized prompt as additional contextual information, allowing the final classification stage to exploit both the optimized natural language instruction and structured signals derived from human sensor reactions.

Evaluator. At each iteration, the current prompt is first used to query an LLM over a set of training instances. The generated predictions are then compared with the gold labels, and the resulting errors are used to compute the evaluation score. In this work, macro-F1 is used as the optimization objective

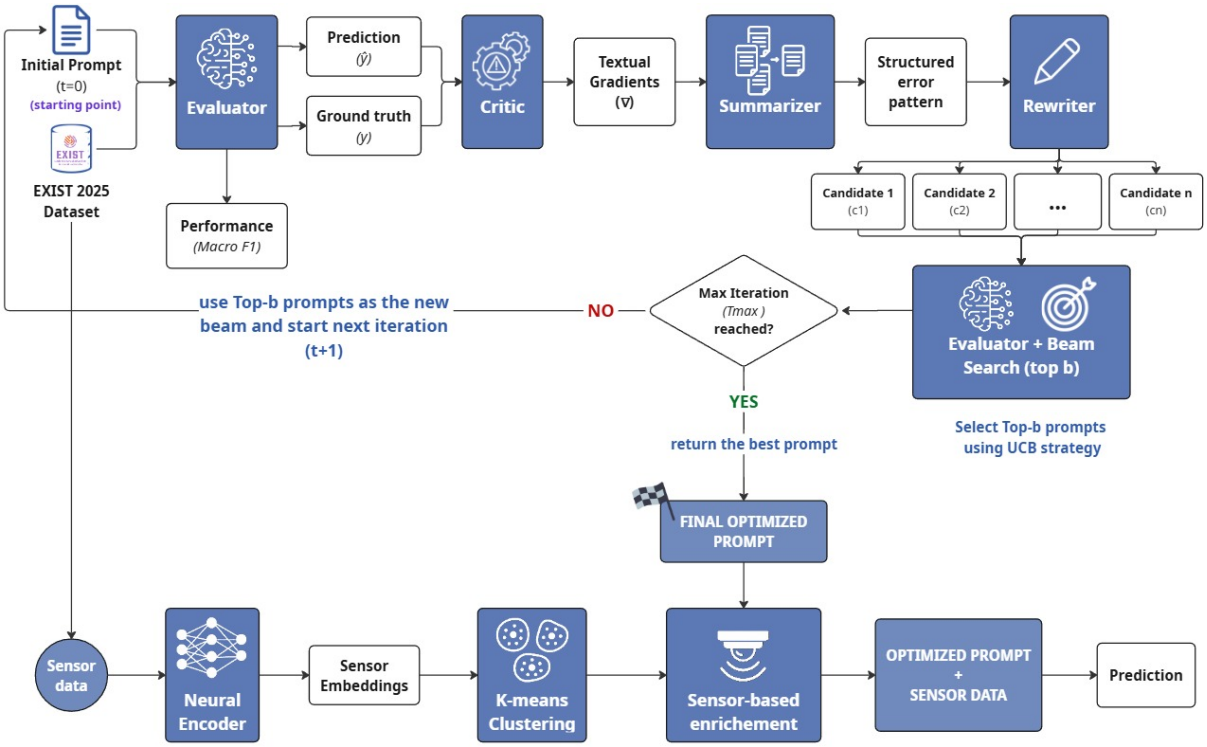

Figure 1: Overview of the proposed HC-ProTeGi-ErPt framework.

because it provides a balanced measure across classes and can be applied consistently to binary, multi-class, and multi-label scenarios.

Critic. The misclassified examples are then passed to a critique stage. For each error, an LLM analyzes the input, the expected label, and the predicted output, producing a natural language explanation of the failure. These critiques indicate which aspects of the current prompt may have contributed to the incorrect prediction and suggest how the instruction could be improved. Following the terminology of ProTeGi, we interpret these critiques as textual gradients, since they provide directional information for improving the prompt without requiring numerical gradient computation.

Summarizer. However, instance-level critiques are often redundant, noisy, or overly specific. To obtain a more stable optimization signal, ProTeGi-ErPt introduces an intermediate summarization stage. This module aggregates the individual critiques into a compact set of recurring error patterns. Each pattern describes a common failure mode and includes a corresponding recommendation for prompt refinement. This aggregation reduces the amount of feedback passed to the rewriting stage and helps the optimizer focus on systematic weaknesses rather than isolated examples.

Rewriter. The rewriting module uses these summarized error patterns to produce new candidate prompts. The candidates are generated by modifying the original instruction in the opposite semantic direction of the identified failures. In other words, the new prompts are designed to explicitly address the weaknesses detected during the critique and summarization stages. To encourage exploration, multiple prompt variants are generated at each iteration, allowing the search process to consider different ways of incorporating the feedback, such as emphasizing the global meaning of the input, clarifying label definitions, or strengthening the treatment of text–image interactions in multi-modal settings.

Beam Search Bandit Optimization. Candidate selection is performed through a beam-search procedure combined with a UCB-based bandit strategy. At each iteration, the bandit mechanism selects

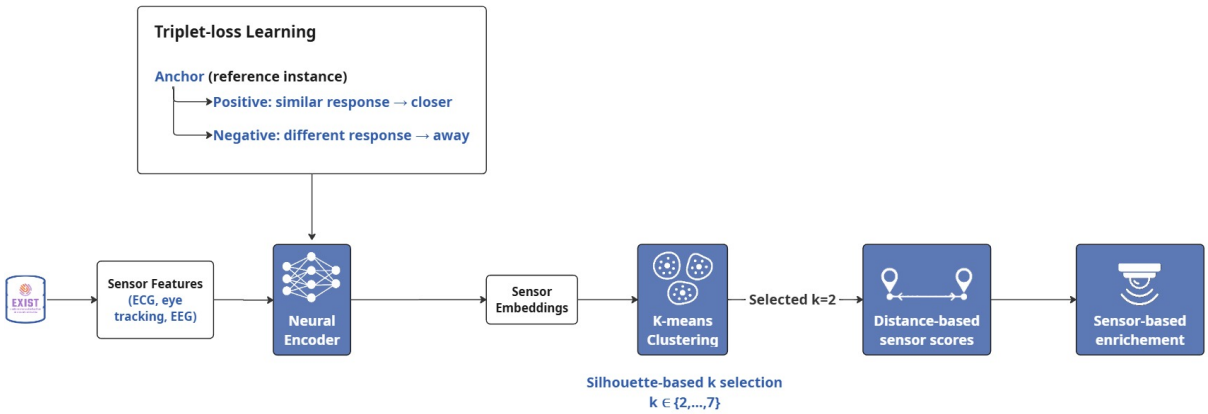

Figure 2: Detailed view of the sensor-based extension within the main HC-ProTeGi-ErPt architecture. Sensor features are encoded with a triplet-loss neural encoder, clustered using K-means with silhouette-based selection, and converted into centroid-distance scores that are later added as auxiliary prompt context.

one prompt from the current beam for expansion, balancing the exploitation of high-performing prompts with the exploration of less-tested alternatives. The selected prompt is evaluated, its errors are analyzed, and the resulting feedback is used to generate new candidates. Each candidate is then evaluated and compared with its parent prompt. The change in macro-F1 is used as the reward signal to update the bandit selector.

After each iteration, the beam is pruned by retaining only the best-performing prompts according to the predefined beam width. This process is repeated for a fixed number of iterations, without early stopping, in order to ensure a consistent comparison across optimization runs. Finally, the prompt achieving the highest macro-F1 score during the search is selected as the optimized prompt and used for the final evaluation on the test set.

Sensor-based Extension. We further investigated whether sensor-derived information could provide complementary signals for prompt-based sexism detection. This extension was applied on top of the best-performing ProTeGi-ErPt configuration, rather than being part of the core optimization loop.

Given the available sensor features, we trained a neural encoder using a triplet-loss objective. This objective encourages the model to learn an embedding space where instances with similar response profiles are closer together, while instances with dissimilar profiles are separated by a margin. The trained encoder was then used to obtain a dense embedding representation for each instance.

To identify latent response patterns, we applied K-means clustering to the learned sensor embeddings. The number of clusters was selected during prototyping using the silhouette score. We evaluated values of k from 2 to 7 for each subtask and language partition, and $k = 2$ consistently achieved the best silhouette score. This suggests that the sensor embeddings were most naturally separated into two dominant response groups. Therefore, we fixed $k = 2$ in the final HC-ProTeGi-ErPt configuration. After clustering, each instance was represented not only by its original input, but also by distance-based scores computed with respect to the learned cluster centroids.

These distance-based scores were incorporated as auxiliary contextual signals into the best ProTeGi-ErPt configuration. The purpose of this extension was to assess whether human-response patterns captured by sensor data could help the model better handle difficult cases, particularly those involving ambiguity, implicit sexism, or multi-modal interpretation.

4. Experimentation

This section describes the experimental setup and evaluation of ProTeGi-ErPt and HC-ProTeGi-ErPt. We first present the evaluated vision-language models and the optimization settings used during prompt search. We then summarize the previous ablation study of ProTeGi-ErPt and focus on the additional

contribution of the human-cognitive sensor-based extension. Finally, we report the official results for EXIST 2026 Task 2, compare them with the official shared-task rankings, analyze the evolution of optimized prompts, and discuss the main findings across model families.

4.1. Models

We evaluate both ProTeGi-ErPt and its sensor-enhanced variant, HC-ProTeGi-ErPt, across several vision-language models to assess the robustness of the framework under different model families, access settings, and capacity profiles. In both frameworks, each model is used independently and applied consistently across all framework components, namely the *Evaluator*, *Critic*, *Summarizer*, and *Rewriter*. All models support multi-modal inputs and are evaluated under the same experimental conditions.

The evaluated models include one open-source model, Qwen3-VL, and six commercial models: Gemini-3-Flash, Gemini-3.1-Flash-Lite, Gemini-2.5-Flash-Lite, GPT-5.1, GPT-5-mini, and GPT-5-nano. Qwen3-VL is included as an open-source baseline, Gemini models are considered as a natively multi-modal commercial family, and GPT models are used as a general-purpose commercial family with different capacity and cost profiles. This selection allows us to analyze whether the effectiveness of ProTeGi-ErPt and HC-ProTeGi-ErPt are stable across heterogeneous LLM backbones.

4.2. Optimization Settings

To make the comparison between configurations reliable, all experiments are performed using the same data splits and sampling strategy. This setup is kept fixed across all multi-modal tasks, so that differences in performance can be attributed to the model and prompt optimization process rather than to variations in the input data. Prompt search is carried out through a compact beam-search procedure guided by a UCB-based bandit selector. In our experiments, the beam width is set to 2, the search is limited to 5 optimization iterations, and the Rewriter produces up to 3 candidate prompts at each iteration. These values were chosen during prototyping to keep the optimization computationally feasible while still allowing the framework to explore alternative prompt formulations. The rewriting process is conditioned on the error patterns detected for each task. Since binary, multi-class, and multi-label settings produce different types of mistakes, the Rewriter applies task-specific refinement strategies. The same applies to the input modality: text-only and multi-modal tasks require different prompt adjustments, particularly when visual information must be considered. In contrast, no language-specific rewriting strategy is introduced for English and Spanish, as both languages are processed under the same task formulation. Candidate prompts are evaluated against their parent prompt using macro-F1. The reward assigned to the bandit selector corresponds to the positive gain in macro-F1 obtained by the candidate; when no improvement is observed, the reward is set to zero.

4.3. Ablation Study

A detailed ablation study was conducted in our previous work on ProTeGi-ErPt, submitted by Llinares et al. [15]. That study focused on EXIST 2025 Task 1.1, corresponding to binary sexism detection in tweets, and analyzed two aspects of the optimization process: first, whether prompt optimization improved over the initial prompt before optimization; and second, whether the use of summarized error patterns in ProTeGi-ErPt improved over the original ProTeGi optimization procedure.

The results showed that both ProTeGi and ProTeGi-ErPt improved over the initial prompt, confirming that prompt optimization was beneficial in this binary classification setting. Beyond this general gain, ProTeGi-ErPt consistently outperformed the standard ProTeGi framework. Averaged across models, ProTeGi-ErPt achieved absolute improvements over ProTeGi of 3.92 macro-F1 points and 10.54 ICM points in English, and 5.83 macro-F1 points and 13.45 ICM points in Spanish. These values were computed by subtracting the ProTeGi score from the corresponding ProTeGi-ErPt score for each model and then averaging the resulting differences across all evaluated models.

The previous analysis suggests that the improvement does not only come from iterative prompt refinement itself, but also from the way feedback is processed before rewriting. While individual

critiques can be noisy, redundant, and instance-specific, the summarization module aggregates them into higher-level error patterns, providing the Rewriter with a more compact and generalizable optimization signal.

Since the effect of the summarization module has already been analyzed in detail in that previous work, we do not repeat the full ablation table here. Instead, the present paper complements that analysis by studying the impact of the human-cognitive extension. In particular, our results show that HC-ProTeGi-ErPt further improves the strongest configurations by incorporating sensor-derived response patterns into the prompt optimization process. Therefore, while the previous ablation isolated the contribution of prompt optimization over the initial prompt and the contribution of summarized error patterns over standard ProTeGi, this work shows that adding human-cognitive information provides an additional benefit in multi-modal meme-based sexism detection.

4.4. Task 2 Evaluation: Sexism Detection in Memes

Table 1 reports the results for Task 2 in the meme-based multi-modal setting, considering both ProTeGi-ErPt and HC-ProTeGi-ErPt. Statistically significant results according to an approximate two-proportion z-test on aggregate metrics ($\alpha = 0.05$) are shown in bold.

The results show that both configurations achieve competitive performance, suggesting that prompt optimization can exploit textual and visual cues in meme-based sexism detection. In addition, the comparison with HC-ProTeGi-ErPt allows us to assess whether incorporating human-cognitive signals provides complementary information beyond the standard multi-modal input.

Task 2.1: Sexism Detection in Memes (Binary). The strongest performance is achieved by the HC-ProTeGi-ErPt extension combined with Gemini-3-Flash. In English, the results show that ProTeGi-ErPt already provides a clear improvement over the compared systems, suggesting that optimized prompts are particularly effective at exploiting the interaction between the image, the overlaid text, and the implicit meaning created by their combination. The HC extension further reinforces this behavior, achieving the best overall configuration.

For Spanish, the improvement is especially visible when human-cognitive information is incorporated. In particular, the difference between HC-ProTeGi-ErPt and the non-HC variant is especially remarkable with the strongest model, Gemini-3-Flash. This suggests that sensor-derived response patterns provide a clearer benefit in Spanish, where they appear to complement visual-textual reasoning more effectively during prompt optimization.

For the official ranking comparison, we report the three highest-ranked configurations for each language according to the official metric. The selected English systems are HC-ProTeGi-ErPt Gemini-3-Flash, ProTeGi-ErPt Gemini-3-Flash, and ProTeGi-ErPt Gemini-3.1-Flash-Lite, while the selected Spanish systems are HC-ProTeGi-ErPt Gemini-3-Flash, HC-ProTeGi-ErPt Gemini-3.1-Flash-Lite, and ProTeGi-ErPt Gemini-3-Flash.

Task 2.2: Source Intention in Memes (Multi-class). Scores are generally lower than in the binary meme classification task, reflecting the increased complexity of source-intention detection. Unlike binary sexism detection, this subtask requires distinguishing whether a meme directly expresses sexism, reports a sexist situation, or judges sexist behaviour, which demands a finer interpretation of both textual and visual cues. In this setting, the use of HC-ProTeGi-ErPt leads to a substantial improvement over the standard ProTeGi-ErPt variant in both languages. In English, the best results are obtained with HC-ProTeGi-ErPt using Gemini-3-Flash, while in Spanish the strongest configuration is HC-ProTeGi-ErPt with GPT-5.1. This indicates that human-cognitive sensor-derived signals are especially useful for source-intention classification, where subtle differences in communicative intent are harder to capture from text and image alone.

For the official ranking comparison, we report the three highest-ranked configurations for each language according to the official metric. The selected English systems are HC-ProTeGi-ErPt Gemini-3-

Table 1
Results obtained by ProTeGi-ErPt and HC-ProTeGi-ErPt on Task 2 in English (EN) and Spanish (ES), reporting Macro-F1, ICM, and ICM_{norm}.

Model	Tasks	English			Spanish		
		Macro-F1	ICM	ICM_{norm}	Macro-F1	ICM	ICM_{norm}
ProTeGi-ErPt Gemini-3-Flash	Task 2.1	**.8067**	**.5183**	**.7631**	**.7305**	.3238	**.6650**
ProTeGi-ErPt Gemini-3.1-Flash-Lite		**.7920**	**.4742**	**.7408**	**.7289**	.3159	**.6609**
ProTeGi-ErPt Gemini-2.5-Flash-Lite		.7484	.3397	.6725	**.7293**	.3208	**.6634**
ProTeGi-ErPt GPT-5.1		.7603	.3777	.6918	.7065	.2602	**.6325**
ProTeGi-ErPt GPT-5-Mini		**.7637**	.3880	.6970	**.7224**	.3000	**.6528**
ProTeGi-ErPt GPT-5-Nano		.6925	.1625	.5825	**.7220**	.3002	**.6529**
ProTeGi-ErPt Qwen3-VL		.6334	.0219	.5111	.6052	-.0324	.4835
HC-ProTeGi-ErPt Gemini-3-Flash		**.8089**	**.5279**	**.7680**	**.7586**	**.4107**	.7092
HC-ProTeGi-ErPt Gemini-3.1-Flash-Lite		**.7738**	**.4237**	**.7151**	**.7317**	.3247	.6654
HC-ProTeGi-ErPt Gemini-2.5-Flash-Lite		.7345	.3024	.6535	.6955	.2244	.6143
HC-ProTeGi-ErPt GPT-5.1		.7569	.3683	.6870	.6870	.1979	.6008
HC-ProTeGi-ErPt GPT-5-Mini		.7470	.3364	.6708	.6596	.1129	.5575
HC-ProTeGi-ErPt GPT-5-Nano		.6982	.1806	.5917	.6834	.1793	.5913
HC-ProTeGi-ErPt Qwen3-VL		.4789	-.2471	.3745	.5388	-.2240	.3859
ProTeGi-ErPt Gemini-3-Flash	Task 2.2	**.5862**	.2692	**.5934**	**.5857**	**.2199**	**.5766**
ProTeGi-ErPt Gemini-3.1-Flash-Lite		**.5815**	.1892	**.5657**	**.5688**	.1010	**.5352**
ProTeGi-ErPt Gemini-2.5-Flash-Lite		.2412	-.0929	.4677	**.5280**	.1380	.4519
ProTeGi-ErPt GPT-5.1		**.5326**	.0612	.5212	**.5965**	**.2578**	**.5898**
ProTeGi-ErPt GPT-5-Mini		**.5821**	.2422	**.5840**	**.5835**	**.1826**	**.5636**
ProTeGi-ErPt GPT-5-Nano		.5133	-.1430	.4504	**.5598**	.1226	**.5427**
ProTeGi-ErPt Qwen3-VL		.1782	-.4423	.3465	.2614	-12.5830	.0617
HC-ProTeGi-ErPt Gemini-3-Flash		**.6119**	**.3548**	**.7631**	**.5344**	.1119	.5388
HC-ProTeGi-ErPt Gemini-3.1-Flash-Lite		**.5868**	.2255	**.5782**	.5739	.1692	.5587
HC-ProTeGi-ErPt Gemini-2.5-Flash-Lite		.4450	-.1496	.4481	.4157	-.1937	.4328
HC-ProTeGi-ErPt GPT-5.1		.5076	.0339	.5118	.6053	.2995	.6043
HC-ProTeGi-ErPt GPT-5-Mini		.5586	.1733	.5600	.5475	.1591	.5552
HC-ProTeGi-ErPt GPT-5-Nano		.4903	-.2077	.4279	.5407	-.0487	.4831
HC-ProTeGi-ErPt Qwen3-VL		.2715	-.9103	.1841	.2901	-.9249	.1779
ProTeGi-ErPt Gemini-3-Flash	Task 2.3	**.5647**	**-.0138**	**.4971**	**.5680**	**-.1605**	**.4672**
ProTeGi-ErPt Gemini-3.1-Flash-Lite		**.5375**	**-.2265**	.4519	**.5272**	**-.3207**	.4344
ProTeGi-ErPt Gemini-2.5-Flash-Lite		.4658	**-.5623**	.3805	.4856	**-.5799**	.3813
ProTeGi-ErPt GPT-5.1		.5085	**-.2414**	.4487	.5119	**-.3188**	.4348
ProTeGi-ErPt GPT-5-Mini		.5117	**-.3871**	.4177	.5181	**-.4278**	.4124
ProTeGi-ErPt GPT-5-Nano		.4575	**-.6806**	.3554	.4644	**-.7690**	.3426
ProTeGi-ErPt Qwen3-VL		.2425	-3.7506	0.0000	.2921	**-2.5801**	0.0000
HC-ProTeGi-ErPt Gemini-3-Flash		**.5822**	**.0842**	**.5179**	**.5626**	**-.1095**	.4776
HC-ProTeGi-ErPt Gemini-3.1-Flash-Lite		**.5363**	**-.1200**	**.4745**	**.5366**	**-.2280**	**.4533**
HC-ProTeGi-ErPt Gemini-2.5-Flash-Lite		.3117	-2.2296	.0263	.4986	**-.5014**	**.3974**
HC-ProTeGi-ErPt GPT-5.1		.4929	**-.2574**	.4453	.5079	**-.3103**	.4365
HC-ProTeGi-ErPt GPT-5-Mini		.5023	**-.4248**	.4097	.5094	**-.4694**	.4039
HC-ProTeGi-ErPt GPT-5-Nano		.4505	-.7434	.3420	.4708	**-.7261**	.3514
HC-ProTeGi-ErPt Qwen3-VL		.4673	-.3814	.4190	.3078	**-2.6371**	0.0000

Flash, ProTeGi-ErPt Gemini-3-Flash, and ProTeGi-ErPt GPT-5-Mini, while the selected Spanish systems are HC-ProTeGi-ErPt GPT-5.1, ProTeGi-ErPt GPT-5.1, and ProTeGi-ErPt GPT-5-Mini.

Task 2.3: Sexism Categorization in Memes (Multi-label). This subtask represents the most challenging setting, since it requires interpreting meme content while assigning multiple fine-grained sexism categories. Compared with the previous subtasks, performance decreases more clearly, with several systems achieving reasonable macro-F1 values but substantially weaker ICM scores. This pattern suggests that models may partially identify relevant labels, but still fail to reproduce the

expected structure of overlapping sexism categories. The results therefore highlight the difficulty of moving from general sexism detection to multi-label categorization, where subtle visual-textual cues must be associated with several possible forms of sexist discourse. Among the evaluated configurations, the best results are obtained by HC-ProTeGi-ErPt with Gemini-3-Flash. This indicates that human-cognitive sensor-derived signals can provide useful complementary information, although the main challenge remains the accurate mapping of implicit meme meaning into multiple fine-grained labels.

For the official ranking comparison, we report the three highest-ranked configurations for each language according to the official metric. The selected English systems are HC-ProTeGi-ErPt Gemini-3-Flash, ProTeGi-ErPt Gemini-3-Flash, and HC-ProTeGi-ErPt Gemini-3.1-Flash-Lite, while the selected Spanish systems are HC-ProTeGi-ErPt Gemini-3-Flash, ProTeGi-ErPt Gemini-3-Flash, and HC-ProTeGi-ErPt Gemini-3.1-Flash-Lite.

Overall, Task 2 shows that human-cognitive information consistently benefits HC-ProTeGi-ErPt across the evaluated settings, with the clearest gains observed in binary meme classification. Nevertheless, performance becomes less stable as the output space increases in complexity, from binary sexism detection to multi-class source-intention classification and multi-label sexism categorization. This trend suggests that the framework is effective at exploiting multi-modal evidence, but that its performance is progressively constrained when the task requires finer-grained distinctions and overlapping label assignments.

4.5. Official Shared Task Evaluation

This section reports the official evaluation results obtained in the shared task for Task 2. Table 2 presents the results for the three official evaluation partitions: English, Spanish, and Global (All instances). The Global partition combines both English and Spanish test examples. Statistically significant results according to an approximate two-proportion z-test on aggregate metrics ($\alpha = 0.05$) are shown in bold.

The evaluation follows the official hard-label setting of the shared task. In this setting, each instance is evaluated against a single categorical decision derived from the annotation process, rather than against the full distribution of annotator disagreement. Therefore, the reported scores reflect the ability of each system to match the final hard labels assigned to the test instances.

The EXIST 2026 evaluation includes several metrics, including Macro-F1, ICM, and ICM_{norm}. Table 2 reports the three metrics for the official ranking systems, while ICM remains the official metric used to determine the EXIST ranking.

The EXIST 2026 shared-task rankings show that the proposed framework performs consistently well across the three Task 2 subtasks and evaluation partitions. In Task 2.1, our system ranks first in English, Spanish, and Global, showing a strong performance in binary meme-based sexism detection. In Task 2.2, the system ranks second across the three partitions, indicating that the proposed framework remains highly competitive even when the task moves from binary detection to multi-class source-intention classification. Finally, in Task 2.3, our system obtains the first position in Spanish and Global, and ranks second in English, despite the additional difficulty of assigning multiple overlapping sexism categories.

Overall, these results suggest that the proposed HC-ProTeGi-ErPt configuration is effective across different levels of task complexity. Its strong performance in both monolingual and multilingual evaluation settings indicates that the combination of prompt optimization and human-cognitive sensor-derived information provides a robust strategy for meme-based sexism detection. The results are particularly relevant because the framework remains competitive not only in binary classification, but also in more complex settings involving source intention and multi-label sexism categorization.

4.6. Manual Prompt Evolution Analysis

To better understand the effect of the optimization process, we compare the initial prompt (level 0) with the best-performing optimized prompt and analyze the main qualitative changes introduced during prompt evolution. In this setting, the initial prompt failed to capture the sexist label, whereas the optimized prompt correctly guided the model toward the sexist class. This allows us to examine how the

Table 2
Official EXIST 2026 ranking results for Task 2 under the hard-label setting, reporting rank, Macro-F1, ICM, and ICM_{norm}. Results are grouped by evaluation partition: English (EN), Spanish (ES), and Global (All instances). Rows labelled *BEST (Ours)* correspond to the best-performing configuration for each language, which in this case matches our best per-language model.

Partition	Rank	System	Macro-F1	ICM	ICM_{norm}
Task 2.1: Sexism Identification in Memes					
EN	1	HC-ProTeGi-ErPt Gemini-3-Flash (*Ours*)	**.8318**	**.5279**	**.7680**
	2	ProTeGi-ErPt Gemini-3-Flash (*Ours*)	**.8277**	**.5183**	**.7631**
	3	MRBLACK	**.8311**	**.4997**	**.7537**
ES	1	HC-ProTeGi-ErPt Gemini-3-Flash (*Ours*)	**.7855**	**.4107**	**.7092**
	2	Ordantis	**.7825**	.3469	**.6767**
	3	Ordantis	**.7820**	.3459	**.6762**
Global	1	HC-ProTeGi-ErPt Gemini-3-Flash (*Ours*)	**.8076**	**.4693**	**.7387**
	2	Ordantis	**.7911**	**.4212**	**.7142**
	3	Ordantis	**.8006**	.4079	**.7074**
Task 2.2: Source Intention in Memes					
EN	1	Ordantis	**.6179**	**.4241**	**.6472**
	2	HC-ProTeGi-ErPt Gemini-3-Flash (*Ours*)	**.6119**	.3548	**.6231**
	3	MRBLACK	**.5941**	.3459	**.6200**
ES	1	Ordantis	**.6106**	**.3176**	**.6111**
	2	HC-ProTeGi-ErPt GPT-5.1 (*Ours*)	**.6043**	**.2995**	**.6053**
	3	ProTeGi-ErPt GPT-5.1 (*Ours*)	**.5857**	.2199	**.5766**
Global	1	Ordantis	**.6158**	**.3709**	**.6289**
	2	HC-ProTeGi-ErPt *BEST* (*Ours*)	**.6081**	**.3267**	**.6136**
	3	ProTeGi-ErPt *BEST* (*Ours*)	**.5858**	.2444	**.5850**
Task 2.3: Sexism Categorization in Memes					
EN	1	MRBLACK	**.5722**	**.1225**	**.5260**
	2	HC-ProTeGi-ErPt Gemini-3-Flash (*Ours*)	**.5822**	.0842	**.5179**
	3	ProTeGi-ErPt Gemini-3-Flash (*Ours*)	**.5647**	-.0138	**.4971**
ES	1	HC-ProTeGi-ErPt Gemini-3-Flash (*Ours*)	**.5626**	**-.1095**	**.4776**
	2	ELiRF-UPV	**.5559**	**-.1334**	**.4727**
	3	ProTeGi-ErPt Gemini-3-Flash (*Ours*)	**.5680**	**-.1605**	**.4672**
Global	1	ProTeGi-ErPt Gemini-3-Flash (*Ours*)	**.5746**	**.0036**	**.5008**
	2	HC-ProTeGi-ErPt Gemini-3-Flash (*Ours*)	**.5679**	**-.0819**	**.4830**
	3	ELiRF-UPV	**.5524**	**-.1069**	**.4778**

optimized prompt modifies the model's decision criteria and improves its ability to capture multi-modal sexist cues. The analysis is conducted within the HC-ProTeGi-ErPt framework, but focuses on the prompt optimization component. We consider three representative model families: an open-source vision-language model (Qwen3-VL), a natively multi-modal LLM family (Gemini-3-Flash), and a widely used commercial proprietary LLM family (GPT-5.1). For readability, the full prompts are reported in Appendix A.

Qwen3-VL (Open-Source). For Qwen3-VL, the initial prompt was mostly a general binary classifier, with limited guidance for multi-modal meme interpretation. It did not explicitly address OCR noise, visual stereotypes, meme-specific gender tropes, or cases where the image modifies or contradicts the text.

Through the optimization process, the prompt evolved towards a more task-specific and multi-modal formulation. The optimized version introduced clearer criteria for text-image integration, meme interpretation, and the treatment of visual/OCR noise. It also refined the decision boundary by requiring explicit sexist evidence in ambiguous cases, reducing the risk of confusing general toxicity with sexism.

Overall, the evolution from the initial to the optimized prompt shows a moderate but relevant improvement. Although the final prompt remained relatively close to the base structure, it introduced more precise multi-modal decision criteria.

Gemini-3-Flash. The prompt evolution for Gemini-3-Flash shows a transition from a general binary sexism classifier to a more specialized multi-modal prompt adapted to memes. While the initial prompt already distinguished sexism from general toxicity, it provided limited guidance on how visual information should affect the final decision, especially when sexist meaning emerges from the interaction between text and image. The optimized Gemini-3-Flash prompt instructs the model to evaluate the relationship between the text and the image, considering that a neutral or friendly text can become sexist when paired with a certain visual context. It also incorporates meme-specific cues, gender-based tropes, and assertions of superiority or contempt toward a specific gender. These additions make the prompt better aligned with the characteristics of sexist memes, where the discriminatory meaning is often expressed indirectly through cultural stereotypes or visual framing.

Another important improvement is that it refines the negative class and the decision boundary by excluding non-gendered aggression, incidental gender mentions, and irrelevant visual noise. In addition, it introduces clearer logic for handling text-image contradictions and distinguishing between sexism criticism and sexist content. Overall, the evolution from the initial to the optimized prompt makes the prompt more multi-modal, meme-aware, and precise in separating sexism from general offensive content.

GPT-5.1 The prompt evolution for GPT-5.1 shows the strongest reformulation among the analyzed models. While the base prompt only provides a general binary definition and broad rules for distinguishing sexism from general toxicity, the optimized version transforms this initial formulation into a highly structured multi-modal decision prompt. It substantially expands the decision criteria by separating sexist from non-sexist content, providing a more exhaustive list of YES cases, and explicitly covering situations where the image adds sexist meaning that the text alone does not convey.

A key improvement is the refinement of the NO class and the integration rules. The optimized prompt excludes non-gendered insults, incidental gender mentions, neutral visual elements, and visual/OCR noise. This makes the decision boundary more precise and reduces false positives caused by general toxicity or irrelevant visual information.

Compared with the initial prompt, the optimized version is therefore more operational, as it replaces broad instructions with concrete decision criteria that the model can apply more consistently.

Global comparison. Overall, the qualitative analysis of the optimized prompts reveals different refinement strategies across model families. The Qwen prompt remains relatively close to the initial formulation, preserving most of the original structure and only introducing limited additions. As a result, its evolution can be interpreted as a conservative refinement rather than a substantial reformulation of the decision process.

In contrast, the Gemini-3-Flash prompt introduces clearer task-specific changes, especially through explicit evaluation rules for multi-modal integration, meme interpretation, visual-textual interaction, and the distinction between sexist content and general toxicity. However, these changes remain relatively concise and operational, focusing on the process the model should follow rather than expanding the prompt with an exhaustive list of cases.

The GPT-based prompt, on the other hand, shows the most extensive reformulation. It provides a much more detailed specification of positive and negative cases, making the prompt more explicit and comprehensive, but also considerably longer and more rigid. While this level of detail can help define the decision boundary more precisely, it may also introduce redundancy or over-constrain the model, especially when the classification requires flexible interpretation of multi-modal memes.

This pattern was also observed across the remaining optimized prompts within each model family, suggesting that prompt evolution is influenced not only by the task, but also by the characteristic refinement behavior of each underlying model.

This observation is consistent with the results obtained for Task 2.1 in English, where Gemini-3-Flash achieved the best performance, followed by GPT-5.1, while Qwen obtained the lowest results among the compared models. These results suggest that the more compact but task-oriented refinement

produced by Gemini was more effective than the highly exhaustive GPT-style prompt, while the limited evolution of Qwen was not sufficient to reach the same level of performance. This indicates that the best-performing prompt was not necessarily the longest or most detailed one, but rather the one that introduced the most relevant multi-modal decision criteria in a concise and usable way. Therefore, the qualitative comparison suggests that prompt refinement is not simply a matter of adding more instructions, but of finding an appropriate balance between specificity, clarity, and operational simplicity.

4.7. General discussion

Although commercial models achieve stronger results in Task 2, the role of open-source models should not be interpreted only in terms of lower performance. Open-source alternatives such as Qwen3-VL provide important practical advantages, including greater control over deployment, improved transparency, easier reproducibility, and reduced dependence on external API providers. These aspects are especially relevant in research settings, where experimental traceability and long-term accessibility are important. In contrast, API-based commercial models may offer higher performance, but they also introduce recurring usage costs, limited control over model updates, and dependence on proprietary infrastructures. Therefore, the results highlight a trade-off between performance and practical autonomy, as commercial systems currently provide stronger effectiveness, while open-source models remain valuable for reproducible, controllable, and more independent experimentation.

Among the evaluated configurations, HC-ProTeGi-ErPt with Gemini-3-Flash provides the most consistent performance across subtasks, which suggests that both the underlying model family and model capacity can influence the effectiveness of prompt-based optimization [17]. A plausible explanation is that Gemini's native multi-modal design may support a more effective alignment between visual and textual information, thereby benefiting meme interpretation [18]. Nevertheless, this explanation should be treated as a hypothesis, since the experimental setup does not explicitly control for architectural differences between models.

The improvement obtained with HC-ProTeGi-ErPt suggests that sensor-derived information provides complementary evidence to the textual and visual modalities. In meme-based sexism detection, the discriminatory meaning is often implicit and emerges from the interaction between the image, the overlaid text, and the reaction elicited in human observers. While the vision-language model mainly relies on the explicit multi-modal content, sensor-based representations can capture patterns related to human response, such as attention, physiological or affective activation. These signals may help the optimization process identify instances that are perceived as more ambiguous, provocative, ironic, or socially loaded.

By converting sensor embeddings into distance-based scores with respect to learned clusters, HC-ProTeGi-ErPt introduces a compact representation of human-cognitive response patterns into the prompt optimization process. This additional context can guide the framework toward prompts that better account for subjective interpretation and implicit sexism. As a result, the model is not only optimizing prompts based on classification errors, but also using auxiliary information related to how humans react to the content. This may explain why the HC extension achieves stronger results than the standard ProTeGi-ErPt configuration, especially in tasks where subjective perception and multi-modal interpretation play a central role.

This interpretation is further supported by the official competition results. HC-ProTeGi-ErPt achieves several first-place official rankings across the Task 2 evaluation partitions and obtains second place in the remaining cases. This consistent performance across languages and subtasks reinforces the effectiveness of the proposed framework and suggests that the integration of human-cognitive sensor-derived signals improves the robustness of prompt optimization for meme-based sexism detection.

5. Conclusions

This paper presented HC-ProTeGi-ErPt, a human-cognitive extension of ProTeGi-ErPt for meme-based sexism detection. The proposed framework builds on feedback-driven prompt optimization and enriches

it with sensor-derived information intended to capture patterns of human response to multi-modal content. While ProTeGi-ErPt refines prompts through an iterative process based on an *Evaluator*, a *Critic*, a *Summarizer*, and a *Rewriter*, HC-ProTeGi-ErPt extends this process by incorporating additional signals derived from human-cognitive data.

The main motivation behind this extension is that sexism in memes is often implicit, subjective, and dependent on the interaction between visual and textual elements. In these cases, textual and visual information alone may not fully capture how the content is perceived by human observers. To address this limitation, HC-ProTeGi-ErPt represents sensor data through learned embeddings, groups these representations using clustering, and converts their relation to cluster centroids into distance-based scores. These scores are then provided to the prompt optimization process as compact contextual signals.

The experimental results on Task 2 of the EXIST shared task show that the best-performing configurations are consistently associated with the HC-ProTeGi-ErPt extension. This indicates that the gains are not only due to the use of strong vision-language models or to prompt optimization alone, but also to the incorporation of human-cognitive response patterns. The official shared task results further support this interpretation, as the HC-based configurations obtain the strongest results in several evaluation partitions and remain highly competitive across the remaining ones.

The analysis also shows that the benefit of HC-ProTeGi-ErPt depends on the complexity of the subtask. The improvement is especially clear in binary meme classification, where the framework can effectively exploit multi-modal and human-centered cues. In the multi-class and multi-label settings, performance becomes less stable, reflecting the additional difficulty of identifying source intention and assigning overlapping sexism categories. Nevertheless, the HC extension remains beneficial in these more complex scenarios, suggesting that sensor-derived response patterns provide useful complementary information for subjective multi-modal classification.

The comparison between commercial and open-source models also provides an important practical insight. Commercial models achieve the strongest overall results, suggesting that larger proprietary multi-modal systems currently offer greater effectiveness for complex meme interpretation. However, open-source models remain valuable because they provide greater control, transparency, reproducibility, and independence from external providers. Therefore, the results show a trade-off between the higher performance of commercial models and the greater control offered by open-source alternatives.

This distinction was also reflected in the manual prompt evolution analysis. Qwen3-VL produced simpler and more conservative rewrites, keeping the optimized prompt close to the initial formulation and introducing only limited additions. In contrast, the commercial models generated more substantial reformulations. Gemini tended to produce concise but task-focused refinements, adding clearer multi-modal decision criteria without making the prompt excessively long. GPT-based models, produced the most extensive rewrites, with more detailed positive and negative decision rules. This qualitative pattern helps contextualize the quantitative results, showing that each model family not only differs in performance, but also in the way it reformulates and operationalizes the prompt during optimization.

Overall, the results support the usefulness of integrating human-centered signals into feedback-driven prompt optimization. HC-ProTeGi-ErPt offers a way to combine error-pattern-based prompt refinement with information related to human reactions, making the optimization process better suited to ambiguous and socially grounded multi-modal tasks.

Future work will focus on improving the stability of the optimization process, exploring stronger strategies for representing sensor-derived signals, and further analysing how human-cognitive information can support prompt-based classification in subjective tasks.

Acknowledgements

The work developed in this publication will be presented as part of the MSc Thesis of the first author, Anna Llinares, at the Universitat Politècnica de València (UPV), Valencia, Spain.

Declaration on Generative AI

During the preparation of this work, the authors used large language and vision-language models as part of the experimental framework, for meme classification, error analysis, summarization of error patterns, and prompt rewriting. In addition, OpenAI's ChatGPT was used as an auxiliary writing support tool to assist with grammar and spelling checking, sentence rephrasing, and improving the clarity of some sentences. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization (Extended Overview), in: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS, 2023. URL: https://ceur-ws.org/Vol-3497/paper-070.pdf.

[2] 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, CEUR Workshop Proceedings, CEUR-WS, 2024. URL: https://ceur-ws.org/Vol-3740/paper-87.pdf.

[3] 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, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2025, pp. 266–289.

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

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

[6] L. S. Mut Altin, H. Saggion, Investigating large language models' (LLMs) capabilities for sexism detection on a low-resource language, in: G. Angelova, M. Kunilovskaya, M. Escribe, R. Mitkov (Eds.), Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2025, pp. 771–779. URL: https://aclanthology.org/2025.ranlp-1.89/.

[7] E. C. Verdú, M. Franco-Salvador, Ecorbi-upv at exist 2025: Large language models and embedding strategies for sexism detection in tweets, Working Notes of CLEF (2025).

[8] T. Alcántara, 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: CLEF 2025 Working Notes, 2025.

[9] I. Arcos, Arcosgpt at exist 2025: Identifying sexism in memes with multimodal deep learning: Fusing text and visual cues, in: CLEF 2025 Working Notes, 2025.

[10] N. Nowakowski, L. Calogiuri, E. Egyed-Zsigmond, D. Nurbakova, J. Erbani, S. Calabretto, Grootwatch at exist 2025: Automatic sexism detection on social networks - classification of tweets and memes, in: CLEF 2025 Working Notes, 2025.

[11] I. Arcos, P. Rosso, E. Gomis, Human-centered multimodal fusion for sexism detection in memes with eye-tracking, heart rate, and eeg signals, in: Proceedings of the Language Resources and Evaluation Conference (LREC 2026), 2026.

[12] P. Italiani, D. Gimeno-Gómez, L. Ragazzi, G. Moro, P. Rosso, Memeweaver: Inter-meme graph reasoning for sexism and misogyny detection, in: Findings of the European Chapter of the Association for Computational Linguistics (EACL 2026), 2026.
[13] B. Pecher, M. Spiegel, R. Belanec, J. Cegin, Revisiting prompt sensitivity in large language models for text classification: The role of prompt underspecification, arXiv preprint arXiv:2602.04297 (2026). URL: https://arxiv.org/abs/2602.04297. doi:10.48550/arXiv.2602.04297.
[14] R. Pryzant, D. Iter, J. Li, Y. T. Lee, C. Zhu, M. Zeng, Automatic prompt optimization with "gradient descent" and beam search, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2023, pp. 7957–7968. doi:10.18653/v1/2023.emnlp-main.494.
[15] A. Llinares, M. Franco, P. Rosso, Feedback-driven prompt optimization for sexism detection, in: International Conference (submitted), 2026.
[16] E. Amigo, 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), 2022, pp. 5809–5819.
[17] A. M. Sarvazyan, J. Á. González, P. Rosso, M. Franco-Salvador, Supervised machine-generated text detectors: Family and scale matters, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2023, pp. 121–132.
[18] G. D. Team, Gemini: A family of highly capable multimodal models, arXiv preprint arXiv:2312.11805 (2023).

A. Prompt Evolution Examples

A.1. Initial Prompt: Level 0 (Same prompt for all models)

```
You are an expert classifier in sexism detection.

Binary task:
- YES: the content is sexist, describes a sexist situation,
  or criticizes sexist behavior.
- NO: the content has no sexist component, even if it includes
  insults, an aggressive tone, or vulgar language.

Evaluate the text and the image jointly.

Decision rules:
1. Label YES if there are gender stereotypes, objectification,
   inferiority, contempt, or discrimination based on being a woman
   (or another gender), or if a sexist behavior is described/reported.
2. Label NO if the attack is not based on gender, even if it mentions
   "woman/women" or uses offensive language.
3. Do not confuse general toxicity with sexism.
4. In case of doubt, prioritize explicit evidence of a sexist component.

Text:
{text}

Image:
{image}

Respond ONLY with YES or NO.
```

A.2. Qwen3-VL Optimized Prompt

```
You are an expert classifier in sexism detection.
```

```
Binary task:
- YES: the content is sexist, describes a sexist situation,
  or criticizes sexist behavior.
- NO: the content has no sexist component, even if it includes
  insults, an aggressive tone, or vulgar language.

Evaluate the text and the image jointly.

Decision rules:
1. Ignore visual/OCR noise and focus on significant visual elements
   that align with the text.
2. Detect meme tropes such as 'Boys vs Girls' or other gendered clichés.
3. Consider if the image contradicts a 'friendly' or benign text.
4. Label YES if there are gender stereotypes, objectification,
   inferiority, contempt, or discrimination based on being a woman
   (or another gender), or if a sexist behavior is described/reported.
5. Label YES if the image explicitly reinforces or contradicts the text
   in a way that indicates sexist content.
6. Label NO if the attack is not based on gender, even if it mentions
   "woman/women" or uses offensive language.
7. Do not confuse general toxicity with sexism.
8. In ambiguous cases or weak evidence, default to NO and assign YES
   only when there is explicit and unambiguous sexist content.

Text:
{text}

Image:
{image}

Respond ONLY with YES or NO.
```

A.3. Gemini-3-Flash Optimized Prompt

```
You are an expert multi-modal classifier specializing in the detection of
sexist content within memes and social media posts. Your task is to perform
a binary classification by integrating both the textual content and the
visual information provided.

Task:
- YES: The content is sexist, promotes gender stereotypes, objectifies,
  or displays contempt based on gender.
- NO: The content has no sexist component, even if it is aggressive,
  vulgar, or includes general toxicity.

Evaluation Rules:
1. Multi-modal Integration: You must evaluate the relationship between
   the text and the image. Sometimes a seemingly neutral text becomes
   sexist when paired with an image, or an image subverts a "friendly"
   text to convey a sexist message.

2. Consider as sexist:
   - Use of gender-based tropes, such as "Boys vs. Girls" memes that
     depict one gender as inherently inferior, irrational, or incompetent.
   - Dehumanization, objectification, or the reduction of individuals to
     their physical appearance or gendered roles.
   - Assertions of superiority or contempt directed at a specific gender.
```

```
   - Situations where the visual context provides the sexist intent that
     the text alone might mask.

3. Do not consider as sexist:
   - General toxicity, insults, or aggressive language that does not target
     a person or group specifically because of their gender.
   - Content that mentions gender but does so without relying on stereotypes,
     inferiority, or prejudice.
   - Noise or irrelevant visual elements, such as background clutter,
     watermarks, or poor image quality, that do not contribute to the message
     of the meme.

4. Decision Logic:
   - If the text and image contradict each other, analyze the combined
     semantic meaning to determine the intent.
   - If in doubt, prioritize explicit evidence of sexist ideology or systemic
     prejudice.
   - Distinguish between a critique of sexism, which is YES, and the use of
     sexism to entertain, which is also YES.

Text:
"{text}"

Image:
[Process the provided image visual features in conjunction with the text]

Respond ONLY with YES or NO.
```

A.4. GPT-5.1 Optimized Prompt

```
You are an expert multi-modal classifier in sexism detection.
You can see both the text and the image, including any embedded
text in the image. Evaluate the two channels jointly and decide
whether the combined meaning expresses sexism.

Binary task:
- YES: the content is sexist, describes a sexist situation, or
  endorses, normalizes, or criticizes sexist behavior based on gender.
- NO: the content has no sexist component, even if it contains insults,
  an aggressive tone, vulgar language, or targets a non-gendered trait.

Consider as sexist (label YES) when any of the following applies:
1. The text or image, or their combination, expresses or endorses
   gender stereotypes, such as "women are bad drivers" or
   "men should not show emotion", objectification, sexualization,
   or treats a gender as inferior, contemptible, or inherently
   different in a derogatory way.

2. The content describes, supports, or normalizes discrimination,
   exclusion, harassment, or violence against people because of
   their gender.

3. The image adds sexist meaning that the text alone does not,
   or vice versa. For example, seemingly neutral or "friendly" text
   paired with an image that depicts ridicule, demeaning captions,
   or stereotypical imagery targeted at a gender.
```

4. The text and image together create a sexist comparison or hierarchy, such as "Boys vs Girls" memes that mock or demean one gender, or "women are this, men are that" tropes.

5. The content uses gendered slurs, demeaning metaphors, or sexual objectification tied to being a woman, man, or other gender.

6. The content reports sexist behavior or attitudes in a way that endorses or minimizes them rather than condemning them.

Do not consider as sexist (label NO) when:
1. The attack or insult is based on non-gender attributes, such as intelligence, appearance unrelated to gender, political affiliation, race, or religion, even if it mentions gender words incidentally.

2. The language is toxic, vulgar, or aggressive but does not target someone because of their gender or does not rely on gendered stereotypes.

3. The image contains neutral photos, memes, or drawings that do not add sexist meaning, and the text alone is non-sexist.

4. Ambiguous or noisy visual elements, such as blurred text, small incidental words, or artifacts from image compression, should be ignored unless they are legible and clearly relevant.

Decision and integration rules:
1. Integrate both channels: label YES only when the gendered meaning is present in text, image, or emerges from their interaction. If one channel is friendly or neutral but the other is clearly sexist, treat the overall content as sexist.

2. Detect contradictions: if the text seems non-sexist but the image portrays stereotypical or demeaning visuals, or vice versa, consider the combination. A hostile image with neutral text can convey sexism.

3. Handle meme tropes: watch for "Boys vs Girls", "expectation vs reality", or template-based memes that compare genders. These often encode stereotypes even when the words are mild.

4. Ignore visual/OCR noise and incidental words that are illegible or clearly accidental. Only use readable, intentional text in the image to determine meaning.

5. Do not conflate general toxicity or insult with sexism. The insult must target or rely on gender to be sexist.

6. If evidence is unclear or equivocal, prioritize explicit, unambiguous gender-based content. When in doubt, require clear gendered targeting to label YES.

Guiding examples:
- YES: image showing a woman depicted as incompetent alongside the caption "Women cannot drive" (stereotype + image).
- NO: vulgar insult directed at a specific person for their behavior that makes no reference to gender.

Text of the instance:
"{text}"

Image:
consider the visible image and any clear embedded text.

Respond ONLY with YES or NO.