Gemini-Enriched Probabilistic Modeling for Multimodal Sexism Characterization in Memes

Sergio Ortiz Montesinos[1,2,*], Fernando Martínez Gómez[1,2]

[1]Universitat Politècnica de València, València, Spain
[2]Ordantis, Spain

Abstract

This paper describes the Ordantis submission to EXIST 2026 Task 2, which focuses on sexism characterization in bilingual memes enriched with physiological data. The task includes three subtasks: binary sexism detection, source-intention classification, and multi-label sexism categorization. The system is based on a probabilistic learning-with-disagreement formulation. Instead of collapsing the six human annotations available for each meme into a single majority label, the models are trained with empirical annotator distributions. Instead of relying on raw image embeddings as the main visual signal, Gemini flash 3.1 is used as a visual and semantic interpreter of the meme: it generates descriptions, sexism analyses, reasoning fields, intention cues, irony summaries, category hints and zero-shot probabilities. These outputs are combined with OCR and encoded with XLM-RoBERTa or a multilingual Longformer, then fused with EEG and Ekman emotion features. The official results show strong performance in Tasks 2.1 and 2.2, especially in the soft setting, where the best runs ranked first. Task 2.3 remained more challenging, particularly for hard multi-label prediction. Beyond the numerical results, the experiments indicate that large language models are useful in multimodal meme understanding not only as zero-shot classifiers, but also as semantic mediators that transform visual and pragmatic content into structured text for supervised models.

Keywords

Sexism Detection, Memes, Multimodal Learning, Learning with Disagreement, Soft Labels, Large Language Models, Gemini, XLM-RoBERTa, Longformer, Physiological Signals, EXIST 2026

1. Introduction

Sexism refers to prejudice, stereotyping, or discrimination directed at people on the basis of their sex or gender, most commonly against women. It spans a wide spectrum, from explicit hostility and demeaning language to subtler forms such as benevolent stereotypes, the objectification of the body, or the normalization of rigid gender roles through humor [1]. Because sexism is grounded in social norms and shared cultural assumptions, it is often expressed implicitly: the same utterance can read as offensive or innocuous depending on context, tone, and the relationship between speaker and audience. This context dependence is precisely what makes it both socially harmful and computationally hard to detect.

On social media, sexist content is amplified by scale and informality, and it is frequently conveyed through memes. A meme pairs a short piece of text with an image, and its meaning emerges from the interaction between the two: visual framing, humor, irony, stereotypes and cultural references all contribute to an interpretation that neither modality fully carries on its own. A caption that is harmless in isolation can become demeaning once combined with a particular image, and a neutral image can acquire a sexist reading from the text placed over it. This multimodal nature has direct consequences for automatic detection. A system that only reads the text transcribed from the meme may miss the visual cues, while a system that only processes the pixels may fail to recover the pragmatic meaning that arises from the combination.

The technologies available for this kind of problem have advanced rapidly in recent years. Text classification is now dominated by the Transformer architecture [2], whose self-attention mechanism

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany
*Corresponding author.
sortmon@etsinf.upv.es (S. O. Montesinos); fmargom1@etsinf.upv.es (F. M. Gómez)

replaced recurrent models and enabled large-scale pretraining. Encoder models such as BERT [3] and its more robustly optimized variant RoBERTa [4] established the pretrain-then-fine-tune paradigm that remains the standard for sentence- and document-level classification. For multilingual settings, XLM-RoBERTa [5] extends this approach to many languages within a single shared model, while long-document encoders such as Longformer [6] relax the quadratic cost of attention so that longer inputs can be processed without aggressive truncation.

Image classification has followed a parallel trajectory. Deep convolutional networks such as ResNet [7] long defined the state of the art, and were later challenged by the Vision Transformer (ViT) [8], which applies the same self-attention machinery directly to sequences of image patches. The convergence of text and image models around a common architecture also enabled vision-language models that align both modalities in a shared representation: CLIP [9], for example, learns a joint image-text embedding space from large collections of captioned images and supports zero-shot classification through natural-language prompts. The specific difficulty of multimodal meme understanding was popularized by the Hateful Memes Challenge [10], which showed that unimodal models struggle on examples deliberately constructed so that only the joint reading of text and image reveals the harmful meaning. More recent work has extended this line of research specifically to sexism in memes, combining visual and textual features with physiological signals such as eye-tracking, heart rate and EEG to better capture how humans perceive and react to harmful content [11].

This work builds on a more recent capability of large multimodal language models: rather than encoding an image as an opaque vector, they can describe it, reason about it, and produce natural-language interpretations of its content. The central idea explored here is to use such a model not as a final classifier but as a semantic mediator that translates the visual and pragmatic content of a meme into structured text, which is then encoded by strong multilingual text models and fused with auxiliary signals. The resulting system is the Gemini-Enriched Multimodal Fusion (GEMF) architecture: Gemini converts each meme into descriptions, analyses and intermediate cues; these are concatenated with the transcribed text and processed by XLM-RoBERTa or Longformer; and the resulting representation is combined with physiological and affective features before a final probabilistic decision. This study was developed in the context of the EXIST 2026 shared task on multimodal sexism characterization in social media [12, 13].

A second design decision concerns supervision. Each meme is annotated by several people, and sexism perception is inherently subjective. Collapsing these annotations into a single majority label would discard the disagreement that is itself part of the phenomenon, so empirical annotator distributions are used as training targets, consistent with the Learning with Disagreement paradigm and with the soft evaluation setting of the shared task. This report describes the full pipeline, the design choices, the submitted runs, the official results and the main lessons learned. The emphasis is not only on what was submitted, but also on why each decision was taken and what the results suggest about the potential and limits of LLM-based semantic enrichment for multimodal meme classification.

2. Task Definition and Data

EXIST 2026 Task 2 is centered on memes in English and Spanish. Each item includes the image, OCR text, language metadata and six human annotations. The training data also includes physiological signals collected while subjects observed the memes. These signals comprise eye-tracking, heart-rate and EEG measurements. The official test set contains 1053 memes.

2.1. Subtasks

Task 2.1 is binary sexism detection. The output space is {NO, YES}. In the hard setting, the system returns one of these labels. In the soft setting, it returns a probability distribution over the two classes.

Task 2.2 is source-intention classification. The classes are NO, DIRECT and JUDGEMENTAL. The distinction is structurally important: NO represents absence of sexism, while DIRECT and JUDGEMEN-

TAL are different forms of sexist intention. This is why the final models do not treat the three labels as a completely flat softmax problem, but use a hierarchical formulation.

Task 2.3 is multi-label sexism categorization. The five official categories are IDEOLOGICAL-INEQUALITY, STEREOTYPING-DOMINANCE, OBJECTIFICATION, SEXUAL-VIOLENCE and MISOGYNY-NON-SEXUAL-VIOLENCE. A meme can belong to more than one category, so the model uses independent sigmoid outputs and category-wise thresholds rather than a single softmax.

2.2. Modalities

Table 1 summarizes the modalities considered during the project and how they were used in the final submission.

Table 1: Input modalities considered in the project and final usage.

Modality	Dimension	Final status	Rationale
OCR text	variable	kept	Main text signal; contains the explicit meme text.
Raw image / ViT	768	removed	Explored as a visual branch. ViT (Vision Transformer) splits the meme image into patches and encodes them with self-attention into a single 768-dimensional embedding; this proved weaker than the natural-language description produced by Gemini.
Gemini text	variable	kept	Provides visual description, sexism analysis, reasoning, intention, irony and category cues.
Gemini probabilities	task-dependent	selected runs	Used as auxiliary features or as direct probability blends.
Eye-tracking	24 per subject	removed	Did not provide consistent gains relative to complexity.
Heart rate	4 per subject	removed	Not sufficiently informative in validation.
EEG	80 per subject	kept	Small complementary physiological signal, pooled with attention.
Ekman emotions	7	kept	Compact affective representation computed from OCR.

3. Learning from Annotator Disagreement

The competition data contains six annotations per meme. This is not only a labeling detail, but an important modeling signal. Sexism perception is subjective: different annotators may disagree because of age, gender, cultural background, personal experience or interpretation of irony. Reducing the six annotations to a single majority label would discard exactly the disagreement that the task is designed to capture.

For Task 2.1, the soft target is defined as

$$y_i = \frac{\#\text{YES votes for meme } i}{\#\text{valid annotations for meme } i}.$$

A meme with six YES votes has target 1.0. A meme with five YES and one NO has target 0.833. A meme with a 3–3 split has target 0.5. This allows the model to learn that ambiguous memes should not necessarily receive extreme probabilities.

For Task 2.2, the target is the empirical distribution over NO, DIRECT and JUDGEMENTAL. If a meme receives two NO votes, three DIRECT votes and one JUDGEMENTAL vote, the target is $[0.333, 0.500, 0.167]$. For Task 2.3, each category has its own target equal to the fraction of annotators that selected that category.

The models are trained with binary cross-entropy with logits using continuous targets. For the binary task, the loss is

$$\mathcal{L} = -y \log(\sigma(z)) - (1 - y) \log(1 - \sigma(z)),$$

where z is the model logit and $y \in [0, 1]$. The same principle is extended to multi-class distributions and multi-label categories. For hard submissions, probabilities are converted to labels by thresholds optimized on validation data.

This decision also influenced the interpretation of the official metrics. The hard metrics evaluate the final discrete decision, while the soft metrics evaluate the probability distribution. Therefore, the best hard model is not necessarily the best soft model. This is one of the reasons why diverse runs were submitted for each subtask.

4. Exploratory Analysis of Disagreement

Before finalizing the modeling strategy, an exploratory analysis was run to verify whether annotator disagreement was meaningful. The training data provides demographic information for annotators, including age, gender, ethnicity, study level and country. The Task 2.1 annotations were reshaped into a long table with one row per meme–annotator pair, and the binary decision was modeled with a mixed-effects logistic regression.

The model included demographic fixed effects and crossed random intercepts for memes and annotators:

$$\text{logit}\left(P(y_{ij} = 1)\right) = \beta_0 + x_j^T \beta + u_i + v_j,$$

where u_i is the meme-level random effect and v_j is the annotator-level random effect. The analysis showed that demographic factors such as age, ethnicity, gender and country were statistically associated with the probability of marking a meme as sexist. It also showed substantial residual variation at both the meme and annotator level.

Demographics were not used as model input in the submitted systems. However, the analysis justified the use of soft labels. It confirmed that disagreement is not random noise to be discarded, but part of the phenomenon. This was especially important for a task involving social interpretation and sensitive content.

5. Preprocessing and Precomputation

5.1. OCR Cleaning

Minimal OCR cleaning was used. URLs were removed, hashtag symbols were stripped while preserving the word, and repeated whitespace was collapsed. Capitalization, accents, emojis, emoticons and OCR errors were deliberately kept. In meme classification, these elements may carry information: capitalization can signal emphasis, emojis can signal irony or affect, and hashtags can reveal stance. Since XLM-RoBERTa is robust to noisy multilingual text, aggressive normalization could remove useful signals.

5.2. Gemini Precomputation

Gemini was called offline once per meme. The goal was not simply to ask Gemini for a final label, but to obtain a structured semantic representation of the meme. Depending on the subtask, the cached JSON contained visual descriptions, sexism analysis, reasoning, intention reasoning, irony summaries, category hints and zero-shot probabilities.

The role of Gemini in the pipeline was not planned from the start. Our initial design followed the conventional multimodal recipe of encoding the meme image with a Vision Transformer, concatenating that visual vector with the OCR representation and the physiological signals, and passing the fusion to a classifier. This is what models M1 and M2 in Table 6 correspond to, and they reached only AUC

0.737–0.741 on Task 2.1, which was below both the reference literature and simple LLM zero-shot baselines. During the multi-view exploration described in Section 6, one specific view stood out clearly from the rest: the one that replaced the ViT branch with a Gemini-generated textual description of the meme reached AUC 0.880 on its own. Averaging it with the weaker views only diluted its advantage, so at that point we stopped treating Gemini as one component among several and made it the central semantic mediator of the pipeline.

Once that decision was taken, the way we called Gemini followed a simple rule: one prompt, one call per meme, all three subtasks at once. Instead of designing separate prompts for binary sexism, intention and category detection, we consolidated everything into a single instruction that asks Gemini to describe the meme, analyze whether it is sexist, expose its step-by-step reasoning, judge the source intention with confidence and irony cues, and produce independent probability estimates for each of the five official categories. The full text of that prompt is reproduced in Appendix A. Working with a single multi-task prompt had three practical consequences that shaped the rest of the system. The precomputation cost is amortized across every downstream model, since the same cached JSON is reused by all Task 2.1, 2.2 and 2.3 variants and exploring dozens of configurations adds no additional API cost. The reasoning and the description come out of the same call as the class-level probabilities, so both views of the meme stay mutually consistent and the supervised model sees a semantic version (the free-text explanation) and a probabilistic version (the zero-shot class scores) that were produced together. And this modularity is what makes it possible to reuse Gemini in three different roles at the fusion layer, as enriched text, as auxiliary numerical features and as a probability source for blending, without any additional prompting effort.

Gemini was used in three different ways. First, it generated enriched text that was concatenated with the OCR and passed to XLM-RoBERTa or Longformer. Second, in Tasks 2.2 and 2.3, it produced numerical features such as class probabilities, confidence scores and irony flags. Third, in selected runs, its zero-shot probability was directly blended with the trained model probability.

The blend is a simple linear combination between the supervised model output and Gemini's zero-shot probability, taken from the same offline JSON cached per meme. The mixing weight is fixed on validation (not learned) and its concrete form depends on the subtask. For Task 2.1, the final probability is

$$P_{final} = 0.6\, P_{model} + 0.4\, P_{Gemini},$$

giving slightly more weight to the supervised model while letting Gemini's better-calibrated prior soften the fine-tuned model's overconfidence. For Task 2.2, only the two sexist-intention components (DIRECT and JUDGEMENTAL) are averaged between the trained model and Gemini, and $P(NO) = 1 - P(sexist)$ is reconstructed afterwards so the distribution stays normalized. No submitted Task 2.3 run uses a probability blend. That the blend is not a mere heuristic is confirmed by the official leaderboard: the Task 2.1 soft blend ranked first overall out of 144 systems (ICM-Soft 0.7206, Table 11.1), and the Task 2.2 hard blend also ranked first out of 186 systems (Table 11.2).

This design made it possible to test three hypotheses: Gemini as a semantic text generator, Gemini as an auxiliary feature provider, and Gemini as a calibrating probability source. The final results show that all three roles can be useful, but not equally in every subtask. All Gemini usage in the submitted runs is zero-shot; extending the prompting regime to few-shot, so that Gemini receives a small set of annotated meme exemplars before scoring each new meme, is a natural direction for future work (Section 18).

5.3. Text Encoding

All final models use a multilingual transformer encoder. The standard backbone is XLM-RoBERTa-base [5]. For long enriched inputs, especially in Task 2.2 and in some Task 2.3 variants, a multilingual Longformer is used [6]. XLM-RoBERTa-base is limited to shorter sequences, while Longformer allows a maximum length of 1100 tokens in these runs.

Masked mean pooling [14] is used instead of relying only on the first token representation. Given

token embeddings $h_1, \ldots, h_T$ and attention mask m_t, the sentence representation is

$$h_{text} = \frac{\sum_{t=1}^{T} m_t h_t}{\sum_{t=1}^{T} m_t}.$$

This was more stable in early experiments than using the raw transformer first-token output, especially during the frozen-backbone warm-up stage.

5.4. Ekman Emotion Features

Seven Ekman-style emotion probabilities [15] are computed from the OCR text: anger, disgust, fear, joy, neutral, sadness and surprise. Since the memes are bilingual, language-specific emotion classifiers were used and their outputs mapped to the same seven-dimensional space. These emotion features are already probabilities and are concatenated directly with the transformer representation. They are not the main source of improvement, but they provide a cheap affective signal.

5.5. Physiological Signals

The released training data includes physiological measurements from subjects who viewed the memes. Eye-tracking, heart-rate and EEG features were explored. Eye-tracking contains features such as pupil diameter, blink statistics, fixation information, saccades and reaction time. Heart-rate contains mean, standard deviation, minimum and maximum heart rate. EEG contains 80 features derived from frequency-band power over multiple channels.

In the final GEMF models, EEG was kept and eye-tracking and heart-rate were removed. The decision was empirical: ET and HR increased complexity but did not provide consistent gains in validation, whereas EEG gave a small but stable complementary signal. Because the number of subjects varies across memes, EEG is pooled with a Set Attention Pooling module inspired by Set Transformer [16]. The pooling module maps a variable-size set of subject-level vectors to a fixed 256-dimensional representation.

6. From Baseline Multimodal Models to GEMF

The final system was not the first architecture tested. The starting point was a conventional multimodal late-fusion model. This baseline used OCR encoded with XLM-RoBERTa, image embeddings from a frozen ViT, physiological features and Ekman emotions. All components were concatenated and passed to an MLP. This model established the basic pipeline but did not achieve the desired performance.

Variants with set-pooling over subjects and different combinations of modalities were then tested. These experiments were useful because they showed that adding more modalities did not automatically improve the system. The most important observation was that the view using Gemini-generated descriptions was much stronger than the views using raw image embeddings. The original multi-view ensemble diluted the strong Gemini-based view with weaker views, so GEMF was promoted from one view inside an ensemble to the main model family.

Table 2 summarizes the internal development pattern for Task 2.1. The exact values correspond to validation experiments and are not official leaderboard scores, but they explain the design decisions.

Table 2: Internal development pattern for Task 2.1. Values are validation results used for architectural decisions.

Model family	Main inputs	AUC	ICM	ICM-Soft
M1 baseline	OCR + ViT + sensors + Ekman	0.737	-0.040	negative
M2 set-pooling	M1 + attention over subjects	0.741	-0.011	negative
GEMF	OCR + Gemini text + EEG + Ekman	0.880	0.323	0.457
GEMF max512+R	GEMF + longer input + reasoning	0.893	0.411	–
GEMF + Gemini blend	Model + Gemini probability	0.889	–	0.596

The conclusion from these experiments was clear: the largest gain came from Gemini-based semantic enrichment. Physiological signals and emotion features were kept as complementary signals, but the main modeling shift was the move from raw visual encoding to LLM-generated textual interpretation.

7. GEMF Architecture

The general GEMF pipeline is shown in Figure 1. A meme is decomposed into its transcribed text (OCR) and its image. Gemini interprets the image offline and produces both enriched text and numerical features. The enriched text is concatenated with the OCR and encoded by a multilingual transformer with masked mean pooling, the variable-size EEG subject set is summarized with set attention pooling, and Ekman emotion probabilities are computed from the OCR. The three resulting vectors $h_{text} \in \mathbb{R}^{768}$, $h_{EEG} \in \mathbb{R}^{256}$ and $h_{emo} \in \mathbb{R}^{7}$ are concatenated and passed to a shared trunk and to the task-specific heads.

For Task 2.1, the final feature vector is

$$h = [h_{text}; h_{EEG}; h_{emo}] \in \mathbb{R}^{1031}.$$

The classifier is an MLP followed by a sigmoid:

$$P(YES) = \sigma(f_{MLP}(h)).$$

For Task 2.2, seven additional Gemini numerical features are concatenated, obtaining a 1038-dimensional vector. For Task 2.3, six category-related Gemini features are concatenated, obtaining a 1037-dimensional vector. These vectors are passed to task-specific heads described below.

8. Training and Inference Protocol

The standard fine-tuning pipeline follows a two-stage process. In the first stage, the transformer backbone is frozen and only the classification head is trained. This stabilizes the head before updating the language encoder. In the second stage, the transformer is unfrozen and trained with smaller learning rates than the head. Early stopping on validation metrics was used, and thresholds were selected on validation predictions for hard submissions.

The training objective always follows the probabilistic targets described above. In the hard setting, the model is not retrained with hard labels; instead, hard labels are obtained from probabilities. This is important because it allows the same underlying model to be evaluated under both hard and soft perspectives.

For Task 2.2 and Task 2.3, calibration or thresholding strategies were also used. Task 2.2 soft runs include Platt scaling [17] in two variants. Task 2.3 uses category-wise thresholds because rare categories require lower thresholds than frequent categories. WeightedRandomSampler was tested to improve minority classes in hard multi-label prediction, but it was avoided in the best soft Task 2.3 run because it reduced calibration.

9. Task-Specific Systems

9.1. Task 2.1: Binary Sexism Detection

Task 2.1 outputs $P(YES)$. The hard output is obtained by thresholding this probability, while the soft output is the distribution $[1 - P(YES), P(YES)]$.

The hard runs test the impact of reasoning and long-context encoding. `task2_1_hard_Ordantis_1` uses XLM-RoBERTa-base with maximum length 512 and includes OCR, Gemini description, sexism analysis and reasoning. `task2_1_hard_Ordantis_2` uses Longformer with maximum length 1100 and excludes reasoning. `task2_1_hard_Ordantis_3` uses XLM-RoBERTa-base with maximum length 512 and excludes reasoning. All three runs use EEG, Ekman and a binary MLP head.

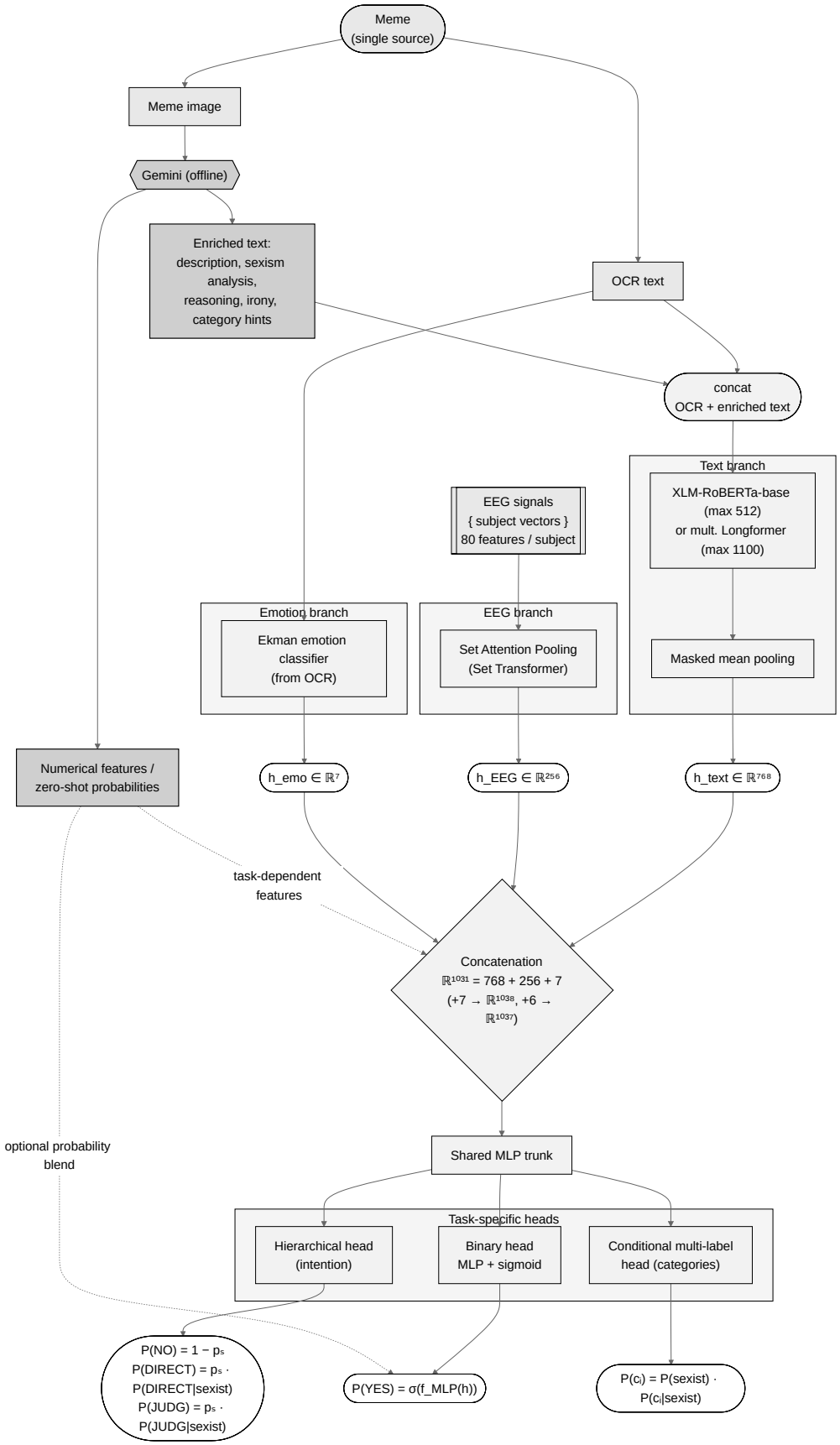

Figure 1: Overview of the GEMF architecture. A meme is split into OCR text and image; Gemini (offline) turns the image into enriched text and numerical features. The enriched text, EEG and Ekman emotion branches are pooled, concatenated and passed to a shared trunk and task-specific heads. Dashed arrows denote Gemini's optional roles: task-dependent numerical features added at fusion and a direct probability blend at the output.

The soft runs test the role of calibration. `task2_1_soft_Ordantis_1` blends the model probability with Gemini's zero-shot probability with the fixed ratio described in Section 5.2:

$$P_{final} = 0.6\, P_{model} + 0.4\, P_{Gemini}.$$

The other two soft runs are Longformer probabilistic outputs without and with Gemini reasoning respectively. In these two variants, Gemini is used only as enriched input text, not as a final probability blend.

9.2. Task 2.2: Source Intention Classification

Task 2.2 uses a hierarchical head. First, the model predicts

$$p_s = P(sexist).$$

Then it predicts the conditional distribution over the two sexist intentions:

$$P(DIRECT \mid sexist), \quad P(JUDGEMENTAL \mid sexist).$$

The final distribution is reconstructed as

$$P(NO) = 1 - p_s,$$

$$P(DIRECT) = p_s \cdot P(DIRECT \mid sexist),$$

$$P(JUDGEMENTAL) = p_s \cdot P(JUDGEMENTAL \mid sexist).$$

This head reflects the semantics of the labels: NO is the absence of sexism, not a third type of sexist intention.

The input text for Task 2.2 is richer than for Task 2.1. It includes OCR, visual description, intention reasoning, irony summary, sexism analysis and general reasoning. In addition, seven Gemini numerical features are concatenated: Gemini's probability of sexism, confidence, the three intention probabilities, an irony flag and irony confidence.

`task2_2_hard_Ordantis_1` uses Longformer with maximum length 1100 and the hierarchical head. `task2_2_hard_Ordantis_2` uses the same architecture with XLM-RoBERTa-base and maximum length 512. `task2_2_hard_Ordantis_3` is a model-Gemini blend. In this blend, DIRECT and JUDGEMENTAL probabilities are averaged between the trained model and Gemini, and $P(NO)$ is reconstructed afterwards.

The soft runs include one raw blend and two calibrated model-only variants. `task2_2_soft_Ordantis_1` averages the full model distribution with Gemini's raw distribution. `task2_2_soft_Ordantis_2` applies Platt scaling to the Longformer hierarchical model, while `task2_2_soft_Ordantis_3` applies Platt scaling to the XLM-R hierarchical model.

9.3. Task 2.3: Multi-Label Sexism Categorization

Task 2.3 is modeled with a conditional multi-label head. The model first predicts $P(sexist)$ and then predicts one conditional probability per category. For each category c_i,

$$P(c_i) = P(sexist) \cdot P(c_i \mid sexist).$$

The motivation is that category labels should be suppressed when the meme is likely non-sexist.

Gemini is used as text and as category-related features. The textual fields include description, sexism analysis, category reasoning and, in one run, additional reasoning. The numerical features are $P(sexist)$ and the five category probabilities produced by Gemini. Unlike Tasks 2.1 and 2.2, no submitted Task 2.3 run uses a final probability blend with Gemini.

`task2_3_hard_Ordantis_1` uses XLM-RoBERTa-base with maximum length 512, the conditional head and WeightedRandomSampler. `task2_3_hard_Ordantis_2` removes the sampler. `task2_3_hard_Ordantis_3` adds Gemini reasoning and also uses the sampler. The soft runs test calibration-oriented variants: `task2_3_soft_Ordantis_1` uses Longformer without sampler, `task2_3_soft_Ordantis_2` uses XLM-R without sampler, and `task2_3_soft_Ordantis_3` uses Longformer with sampler.

10. Submitted Runs

Table 3 summarizes the 18 submitted systems. XLM-R denotes XLM-RoBERTa-base and LF denotes the multilingual Longformer.

Table 3: Summary of submitted Ordantis runs.

Task	Type	Run	Backbone	Post-processing	Main configuration
2.1	hard	1	XLM-R	threshold	Gemini description, sexism analysis and reasoning; EEG and Ekman
2.1	hard	2	LF	threshold	Gemini description and sexism analysis; longer context
2.1	hard	3	XLM-R	threshold	Same as run 2 but with XLM-R and max length 512
2.1	soft	1	blend	average	GEMF probability blended with Gemini zero-shot probability
2.1	soft	2	LF	none	Longformer probabilistic output without reasoning
2.1	soft	3	LF	none	Longformer probabilistic output with reasoning
2.2	hard	1	LF	hierarchical thresholds	Text and numerical Gemini features; hierarchical intention head
2.2	hard	2	XLM-R	hierarchical thresholds	Same head as run 1 with shorter context
2.2	hard	3	blend	average + reconstruction	Model-Gemini blend for DIRECT and JUDGEMENTAL
2.2	soft	1	blend	average	Raw blend of model and Gemini distributions
2.2	soft	2	LF	Platt scaling	Longformer hierarchical model with calibration
2.2	soft	3	XLM-R	Platt scaling	XLM-R hierarchical model with calibration
2.3	hard	1	XLM-R	category thresholds	Conditional multi-label head with WeightedRandomSampler
2.3	hard	2	XLM-R	category thresholds	Same architecture without sampler
2.3	hard	3	XLM-R	category thresholds	Sampler plus additional Gemini reasoning
2.3	soft	1	LF	none	Conditional multi-label head without sampler
2.3	soft	2	XLM-R	none	XLM-R conditional model without sampler
2.3	soft	3	LF	none	Longformer conditional model with sampler

11. Official Results

This section reports the official all-instances results for the submitted Ordantis runs. Rankings correspond to the global leaderboard in each evaluation setting.

11.1. Task 2.1 Results

The strongest hard run was `Ordantis_2`, ranking third out of 217 systems with ICM-Hard 0.4079 and F1 YES 0.8006. The strongest soft run was `Ordantis_1`, ranking first out of 144 systems with ICM-Soft 0.7206 and ICM-Soft Norm 0.6158.

Table 4: Official all-instances Task 2.1 results for Ordantis (217 systems in Hard-Hard, 144 in Soft-Soft).

Setting	Run	Rank	ICM	ICM Norm	Extra
Hard	Ordantis_2	3	0.4079	0.7074	F1 YES 0.8006
Hard	Ordantis_1	4	0.4005	0.7037	F1 YES 0.8003
Hard	Ordantis_3	6	0.3943	0.7005	F1 YES 0.7969
Soft	Ordantis_1	1	0.7206	0.6158	CE 0.8155
Soft	Ordantis_3	2	0.5175	0.5832	CE 0.8227
Soft	Ordantis_2	3	0.5099	0.5820	CE 0.8146

The main observation is that Task 2.1 benefited strongly from Gemini. The hard runs are very close to each other, suggesting that both XLM-R and Longformer were competitive for the binary decision. In contrast, the best soft result came clearly from probability blending with Gemini, confirming the value of Gemini as a calibration source.

11.2. Task 2.2 Results

The strongest hard run was `Ordantis_3`, ranking first out of 186 systems with ICM-Hard 0.3709 and F1 0.6158. The strongest soft run was `Ordantis_1`, also ranking first out of 117 systems, with ICM-Soft 0.0114 and ICM-Soft Norm 0.5012.

Table 5: Official all-instances Task 2.2 results for Ordantis (186 systems in Hard-Hard, 117 in Soft-Soft).

Setting	Run	Rank	ICM	ICM Norm	Extra
Hard	Ordantis_3	1	0.3709	0.6289	F1 0.6158
Hard	Ordantis_2	5	0.1790	0.5622	F1 0.5453
Hard	Ordantis_1	11	0.0659	0.5229	F1 0.5368
Soft	Ordantis_1	1	0.0114	0.5012	CE 1.3334
Soft	Ordantis_2	2	-0.5650	0.4399	CE 1.5248
Soft	Ordantis_3	3	-0.5685	0.4395	CE 1.4874

These results are particularly informative. Longformer was designed to reduce truncation and was strong in validation, but the official hard winner among these runs was the model-Gemini blend. This suggests that Gemini's zero-shot intention probabilities generalized well and provided complementary information. The soft results show an even stronger pattern: the raw blend was clearly better calibrated than the Platt-scaled model-only variants.

11.3. Task 2.3 Results

Task 2.3 exhibited a marked asymmetry between the two evaluation settings. The strongest hard run was `Ordantis_1`, ranking 132 out of 187 systems with F1 0.379, whereas the same run in the soft setting ranked 10 out of 118 systems (ICM-Soft -4.6874). Because both settings evaluate the *same* underlying probabilistic output, this asymmetry indicates that the multi-label model itself is competitive at ranking probabilities, but that the discrete binarization step does not transfer to the test distribution. A detailed diagnosis is given in Section 14.

Table 6: Official all-instances Task 2.3 results for Ordantis (187 systems in Hard-Hard, 118 in Soft-Soft).

Setting	Run	Rank	ICM	ICM Norm	Extra
Hard	Ordantis_1	132	-1.9465	0.0962	F1 0.3790
Hard	Ordantis_2	136	-1.9631	0.0927	F1 0.3786
Hard	Ordantis_3	143	-2.0800	0.0685	F1 0.3731
Soft	Ordantis_1	10	-4.6874	0.2516	–
Soft	Ordantis_2	26	-5.7804	0.1937	–
Soft	Ordantis_3	38	-6.3589	0.1630	–

The drop in hard Task 2.3 indicates that the category thresholds and rare-class treatment did not generalize sufficiently. The sampler helped expose the model to minority categories, but it also risked producing less calibrated probabilities. This trade-off is visible in the soft results, where the no-sampler Longformer variant was the strongest and the sampler-based Longformer variant fell to rank 38. The evidence is consistent with a post-processing problem rather than a modeling one: the underlying probabilistic model ranks well (soft rank 10/118) but its binarization via category thresholds does not survive the validation-to-test distribution shift (Section 14).

12. Decision Analysis

Table 7 summarizes the main technical decisions, the alternatives considered and the reason for the final choice.

Table 7: Main modeling decisions and rationale.

Decision	Rationale
Soft labels instead of majority voting	Preserves annotator disagreement and makes training consistent with soft evaluation.
Gemini-generated text instead of raw ViT as the main visual signal	Gemini descriptions and analyses captured visual, pragmatic and cultural meaning more effectively than generic image embeddings.
Masked mean pooling instead of first-token pooling	Produced a more stable sentence representation during warm-up and fine-tuning.
Longformer variants for long enriched inputs	Reduced truncation when descriptions, reasoning, intention and irony fields produced long sequences.
EEG kept; eye-tracking and heart rate removed	EEG added a small complementary signal, while ET and HR did not justify their additional complexity.
Hierarchical head for Task 2.2	Matches the label structure: NO is absence of sexism, whereas DIRECT and JUDGEMENTAL are sexist intentions.
Conditional multi-label head for Task 2.3	Suppresses category probabilities when the meme is likely non-sexist.
Blends and Platt scaling for calibration	Soft metrics reward calibrated probabilities, not only correct class decisions.
WeightedRandomSampler for hard Task 2.3	Increased exposure to rare categories, with a known trade-off in probability calibration.

Several decisions were empirical rather than purely theoretical. For example, Longformer was not universally better than XLM-R, and WeightedRandomSampler was not universally better than standard sampling. The final submissions therefore intentionally cover different hypotheses. This diversity was important because the official leaderboard showed that different designs generalized best for different subtasks.

13. Calibration Analysis

The official ICM-Soft metric rewards calibration only in aggregate. To assess calibration directly, we report Expected Calibration Error (ECE) with 10 equal-width bins, Maximum Calibration Error (MCE) and Brier score, computed on the validation split ($n = 598$) and comparing raw probabilities against post-hoc calibration. For Task 2.3, all figures correspond to the submitted run `Ordantis_1` (Table 11.3). Table 13 summarizes the main results.

Table 8: Calibration on the validation split ($n = 598$). ECE with 10 equal-width bins; per-class Platt scaling is fitted and evaluated on the same split (see caveat below).

Subtask (run)	ECE raw	ECE calibrated	Calibration method
2.1 binary (blend vs. raw)	0.108	**0.075**	Gemini blend (temperature worsens ECE-hard)
2.2 hierarchical (Ordantis_2/3)	0.131	**0.037**	per-class Platt (-72%)
2.3 conditional (Ordantis_1)	0.078	**0.033**	per-class Platt (-58%)

The main pattern in the table is that our supervised models are already reasonably calibrated in the raw setting (ECE between 0.08 and 0.13), which is what one would expect from training on soft targets instead of hard majority labels: the model learns to produce probabilities close to the empirical annotator distribution rather than saturating at 0 or 1. Per-class Platt scaling still helps in Tasks 2.2 and 2.3, cutting ECE by roughly 60 to 70 percent, but the same technique does not help in Task 2.1: temperature scaling pushes probabilities towards 0.5 and slightly worsens ECE-hard, while the model-Gemini blend does bring it down (from 0.108 to 0.075). The reason is visible in Gemini's own probabilities, which have a low global ECE but a very high MCE (0.55). In other words, Gemini is often confidently wrong in localized regions of the probability space, and that is exactly why blending $0.6\,P_{model} + 0.4\,P_{Gemini}$ improves Task 2.1 calibration whereas relying only on Gemini would not: the supervised model corrects Gemini's confident mistakes while borrowing its overall smoother distribution.

One caveat is that both Platt and temperature parameters are fitted on the same validation split where ECE is measured. The reported improvements should therefore be read as optimistic upper bounds on the calibration gains achievable in production, not as unbiased estimates.

14. Error Analysis for Task 2.3

Task 2.3 is the weakest subtask in the hard evaluation, but its behavior is more informative than the raw rank suggests. This section adds the per-category, co-occurrence and generalization analysis that Section 11.3 only sketches at a summary level.

Hard-soft asymmetry. The most striking pattern is the gap between the two evaluation settings on the same model output. In the soft setting, our best submitted run ranked 10 out of 118 systems, which places it in the 91st percentile. In the hard setting, that same run fell to rank 132 out of 187, with F1-macro dropping from 0.715 on validation to 0.379 on the official test. Since soft and hard evaluate identical probabilities, this asymmetry cannot come from the representation or the features: the underlying probability ranking is competitive at world level, but the binarization step does not survive the move from validation to test. The most likely cause is overfitting of the category thresholds and the WeightedRandomSampler to the validation split. The thresholds (t_{sex}, t_{cat}) were tuned by grid search of the official ICM on the 598 validation memes, and with five categories and a sufficiently fine grid, that search can easily land on combinations that maximize the metric on validation without capturing patterns that generalize. This is consistent with the fact that the same problem does not show up in Tasks 2.1 and 2.2, whose post-processing involves only one or two thresholds respectively (both of them ranked first in their soft evaluation).

Per-category performance. Table 14 reports precision, recall and F1 per category on validation for the submitted run `Ordantis_1`, using the official gold and the evaluation decoder that reproduces Table 11.3.

Table 9: Per-category performance of run `Ordantis_1` on validation ($n = 598$).

Category	Gold	Predicted	Precision	Recall	F1
STEREOTYPING-DOMINANCE	372	487	0.682	0.893	**0.773**
OBJECTIFICATION	324	377	0.695	0.809	0.748
IDEOLOGICAL-INEQUALITY	346	419	0.666	0.806	0.729
SEXUAL-VIOLENCE	183	201	0.672	0.738	0.703
MISOGYNY-NON-SEXUAL-VIOLENCE	210	335	0.505	0.805	**0.620**
F1-macro	–	–	–	–	0.715

The first pattern in this table is that every category has high recall (between 0.74 and 0.89) but noticeably lower precision. That is the direct signature of thresholds tuned aggressively on validation to recover minority classes: the model over-predicts, which helps the metric being tuned but does not travel well to test. The second pattern is that the weakest category is not the rarest one. MISOGYNY-NON-SEXUAL-VIOLENCE has the lowest F1 (0.620), but its failure is one of precision (0.505), not scarcity: it is predicted 335 times against only 210 gold instances. The actual rarest category, SEXUAL-VIOLENCE (183 gold), achieves a higher F1 (0.703). So MISOGYNY-NSV suffers from boundary confusion with other, more frequent categories, not from lack of training signal.

Category co-occurrence in gold. The origin of that boundary confusion becomes clear when the gold co-occurrence structure is examined (Table 14).

Table 10: Most frequent gold category co-occurrences on the validation split.

Category pair	Co-occurrences
STEREOTYPING-DOMINANCE + IDEOLOGICAL-INEQUALITY	224
STEREOTYPING-DOMINANCE + OBJECTIFICATION	215
STEREOTYPING-DOMINANCE + MISOGYNY-NSV	156
OBJECTIFICATION + IDEOLOGICAL-INEQUALITY	153
OBJECTIFICATION + SEXUAL-VIOLENCE	141
OBJECTIFICATION + MISOGYNY-NSV	130
IDEOLOGICAL-INEQUALITY + MISOGYNY-NSV	120
STEREOTYPING-DOMINANCE + SEXUAL-VIOLENCE	107
SEXUAL-VIOLENCE + MISOGYNY-NSV	83
IDEOLOGICAL-INEQUALITY + SEXUAL-VIOLENCE	65

STEREOTYPING-DOMINANCE acts as a hub category. It shows up in the three most frequent pairs and in four out of the top five, and MISOGYNY-NSV co-occurs with it more often than with any other category (156 pairs). The two rarely appear in isolation from each other in the training distribution, so the model struggles to draw a clean boundary between them. Our conditional multi-label head uses independent per-category sigmoids and cannot encode this dependency structure. That is a structural limitation of the head, not something a better threshold could fix, and it motivates the graph-based and label-dependency directions discussed in Section 18.

Summary. The Task 2.3 failure in the hard setting decomposes into two distinct effects. The first is validation-driven over-prediction from thresholds that do not transfer to test (visible as the systematic recall vs. precision gap in Table 14). The second is unmodeled label dependencies concentrated on a hub category (Table 14). The underlying probabilistic model itself is competitive, which is what the soft rank 10 out of 118 demonstrates.

15. Computational Cost and Modality Ablation

Computational cost. All experiments run on a single NVIDIA RTX 4500 Ada (25.3 GB) in BF16 precision. The dominant one-off cost is the offline Gemini precomputation of enriched text, features and zero-shot probabilities for all 5037 memes, which takes roughly 73.7 minutes of wall-clock time across the full corpus and is computed once and then reused across every run. This precomputation cost approximately $47 in API usage, fully covered by Google's welcome credit, so it did not translate into any actual expense for the authors. The probability-blend runs add no training cost of their own, since they only recombine already-computed probabilities. Each final model trains in two phases, a frozen-backbone head warm-up followed by joint fine-tuning, for a total of 8 to 13 epochs per model with roughly 615k to 620k trainable head parameters on top of the multilingual encoder backbone. We do not report total GPU-hours, FLOPs or energy for the final models: per-epoch timing was not instrumented for those runs, so any such figure would be an extrapolation, and we prefer to stick to what was directly measured.

Modality ablation. To quantify the contribution of the physiological and affective branches, we ablate them on the validation split, keeping only the Gemini-enriched text and re-evaluating each subtask with the same head and thresholds. Removing EEG and Ekman together costs approximately 2–3 points across all three subtasks, as summarized in Table 15. This confirms that the bulk of performance comes from the Gemini-enriched semantic text (which alone accounts for the AUC jump from 0.741 to 0.880 reported in Table 6), while EEG and Ekman provide a small but consistent complementary signal. Test-time ablations are not reported because the official test labels are held by the organizers.

Table 11: Modality ablation on the validation split ($n = 598$). Removing EEG and Ekman keeps the text-only branch of GEMF.

Subtask	Metric	GEMF full	Text only (no EEG, no Ekman)
2.1	AUC	0.893	0.872
2.2	F1-macro	0.602	0.575
2.3	F1-macro	0.715	0.688

16. Discussion

16.1. What Worked

The most successful component was Gemini-based semantic enrichment. This was especially clear in Task 2.1 and Task 2.2. Memes often require interpretation beyond literal OCR. Gemini provided descriptions of the visual scene, inferred relations between image and text, detected ironic or judgmental cues and produced initial probability estimates. Once this information was converted into text, XLM-RoBERTa and Longformer could exploit it effectively.

The second important success was probability blending. In Task 2.1 soft and Task 2.2 soft, the best runs were blends between the trained model and Gemini. In Task 2.2 hard, the official best run was also a blend. This does not mean that Gemini should replace the trained model. Instead, it suggests that Gemini provides a complementary prior: the supervised model learns from the competition data, while Gemini contributes broad visual-semantic knowledge and often smoother probability estimates.

The third success was the use of task-specific heads. The hierarchical head in Task 2.2 reflects the label structure and makes the model interpretable: first detect sexism, then classify the type of sexist intention. The conditional head in Task 2.3 follows the same principle for multi-label categories.

16.2. What Did Not Work as Expected

The original multimodal idea of using all available signals did not produce the strongest models. ViT image embeddings, eye-tracking and heart-rate features did not provide a gain comparable to Gemini enrichment. This does not mean that visual or physiological signals are useless, but in this implementation their effect was much smaller than the effect of converting the meme into semantic text.

Task 2.3 was the main failure point in the hard evaluation. Internal validation suggested that the conditional multi-label model and sampler were promising, and the soft submission actually placed us in the top ten of the leaderboard (rank 10 out of 118 systems); the drop is concentrated in the hard setting, where the same run fell to rank 132 out of 187. This gap between soft and hard on the same underlying probabilities points at the post-processing step rather than at the model itself, and it can be traced to a combination of class imbalance, sensitivity to thresholds tuned on the validation split, rare-category distribution shift and imperfect modeling of label dependencies. A detailed diagnosis is given in Section 14. Multi-label sexism categorization is not just five independent binary tasks: categories co-occur and some are semantically close, and the current head does not explicitly model these dependencies.

Platt scaling also had mixed effects. It improved some model-only probability outputs but did not always improve over raw model-Gemini blends. This shows that calibration is not only a post-processing issue; it also depends on the quality and diversity of the probability sources being combined.

16.3. Potential of LLMs in Multimodal Meme Understanding

One of the main conclusions of this work is that LLMs are highly valuable in multimodal meme classification, but their best role is not necessarily to act as standalone black-box classifiers. Their greatest contribution in this system was as semantic mediators. Gemini transformed visual content, OCR, irony and implicit context into structured natural language that could be processed by supervised encoders.

This is important because memes contain meanings that are difficult to represent with standard pixel embeddings. A visual encoder may detect objects, faces or layout, but it may not explain why a juxtaposition is sexist, why a phrase is ironic or why an image reinforces a stereotype. Gemini-generated text provided this missing intermediate representation.

At the same time, relying only on Gemini would be risky. LLM outputs can be biased, overconfident, inconsistent or miscalibrated. The supervised model remains necessary because it learns the competition-specific annotation distribution, task definitions and evaluation constraints. The best approach in these experiments was not to choose between Gemini and supervised learning, but to combine them in controlled ways: as text, as features or as probability blends.

Therefore, the broader lesson is that LLMs can act as powerful multimodal interpreters, especially when paired with supervised models trained on task-specific disagreement-aware targets. This hybrid strategy is particularly promising for sensitive social-media tasks where meaning depends on context, implicit stereotypes and subjective interpretation.

16.4. Why Soft and Hard Results Differ

The official results confirm that hard and soft evaluation reward different behavior. Hard metrics reward a good final decision boundary. Soft metrics reward a probability distribution close to the annotator distribution. A model can be strong in hard classification while being poorly calibrated. Conversely, a softer probability blend may be better in soft evaluation even if it is not the most decisive classifier.

This explains why the best hard and soft runs are sometimes different. It also validates the decision to submit multiple diverse systems rather than only variants of a single model. In a setting with disagreement and uncertainty, the evaluation target determines what "best" means.

17. Limitations

Dependence on a single external interpreter. A central limitation of the system is that it relies on a single external multimodal model, Gemini, as its main source of visual and semantic enrichment. There are two sides to this dependence. The Gemini branch is precisely what made the system competitive in the first place: replacing raw ViT embeddings with Gemini-generated text raised validation AUC from 0.741 to 0.880 on Task 2.1 (Table 6), which is the largest single design gain in the entire pipeline. At the same time, Gemini never decides on its own in any of the submitted runs. In the probability-blended runs the supervised model keeps the larger weight (0.6 against 0.4), and the calibration analysis in Section 13 shows that Gemini's confidently wrong regions (its MCE of 0.55) are corrected rather than inherited by the pipeline. As a concrete example, Gemini assigns SEXUAL-VIOLENCE a very low mean probability of 0.089, and yet the supervised model does not collapse to that estimate. Even so, the single-model bottleneck remains a genuine risk, especially if Gemini's biases or blind spots are systematic for specific visual styles, languages or categories. Two natural ways to mitigate it are ensembling several vision-language interpreters and, more radically, distilling Gemini's knowledge into a Gemini-free student so that no external LLM call is required at inference. Both are discussed in Section 18.

Robustness of hard multi-label prediction. Task 2.3 shows that the current design is not robust enough for rare-category prediction under strict binarization. As Section 14 makes explicit, the underlying probabilistic model ranks well (soft rank $10/118$), but its per-category thresholds and sampler do

not transfer from validation to test, and the head does not model label dependencies concentrated on a hub category (STEREOTYPING-DOMINANCE). This is a structural issue that goes beyond the choice of threshold.

Modeling of annotator disagreement. The system uses annotator disagreement only through empirical soft targets and the exploratory mixed-effects analysis of Section 4; annotator groups are not part of the prediction process itself. Because that analysis found age, gender, ethnicity and country to be significantly associated with sexism perception, an explicit next step would be to condition predictions on annotator groups rather than only on aggregated soft labels.

Cross-dataset generalization. The GEMF architecture has been validated only on EXIST 2026. Being built on a general-purpose multimodal LLM interpreter and a multilingual text encoder, it is in principle dataset-agnostic, but transfer to other meme or hate-speech benchmarks, alternative label taxonomies and languages beyond English and Spanish remains empirically untested.

Verification of LLM-generated reasoning. The model treats Gemini's descriptions, sexism analyses and reasoning fields as input features but does not verify their factual or social correctness. In sensitive domains such as sexism detection, LLM-generated explanations should be treated as auxiliary signals rather than as ground truth.

Computational overhead. The final pipeline is computationally more complex than a plain text classifier because it requires offline LLM precomputation. Although a one-off cost, this precomputation is what makes every run in the paper reproducible only when API access to Gemini (or an equivalent multimodal LLM) is available.

18. Future Work

The analysis of the submitted runs suggests several concrete lines of future work, most of them directly connected to specific findings reported earlier in the paper.

Explicit modeling of label dependencies. Task 2.3 requires better multi-label modeling. The conditional head used here treats the five categories as independent per-category sigmoids and cannot exploit the strong co-occurrence structure of the gold data (Section 14). Label-dependency layers, classifier chains and graph-based category relations [18] are natural candidates, particularly because they would address the boundary confusion between MISOGYNY-NON-SEXUAL-VIOLENCE and its hub co-occurrent STEREOTYPING-DOMINANCE.

Distilling Gemini out of inference. Gemini is currently required at inference time for every meme, which is both the main runtime cost and the single point of failure discussed in Section 17. A promising line of work is to treat Gemini's outputs (descriptions, analyses, reasoning fields and zero-shot probabilities) as *privileged information* that is only available during training [19], and then distil that information into a compact student model that at inference only sees the OCR text and the meme image [20]. If this works well, the deployed system would make no external API call at test time and would run end-to-end on the local encoder alone, which would eliminate the external LLM dependency and its precomputation cost while preserving most of the semantic enrichment gain reported in this paper.

Few-shot prompting. All Gemini usage in the submitted runs is zero-shot. Extending this to few-shot prompting, where Gemini receives a small set of annotated meme exemplars per subtask before scoring each new meme, is a natural extension consistent with recent evidence that few-shot in-context examples can substantially improve subjective classification and calibration.

Losses aligned with the ICM-Soft metric. The system is trained with binary cross-entropy on soft targets and its hard decisions are obtained through post-hoc thresholds. Training directly with a differentiable surrogate of ICM-Soft, or with a loss that jointly optimizes the probabilistic and thresholded outputs, would reduce the sensitivity to threshold choices that produced the validation-to-test gap in Task 2.3.

Threshold robustness. Independently of the loss, the search space for category thresholds should be reduced (for example a single shared threshold across categories, or cross-validated thresholds within the training split) to make the tuned decision boundary less specific to the validation partition.

Predictions conditioned on annotator groups. The exploratory analysis of Section 4 showed that age, gender, ethnicity and country are significantly associated with sexism perception. Predicting distributions conditioned on annotator groups, or using mixture models that represent different perspectives, would model disagreement explicitly rather than only using it as a training signal.

Structured prompting and evidence selection. A complementary line of work is to integrate Gemini through more structured prompting, with separate prompts for visual description, OCR–image alignment, ironic intent and category evidence, and to extract compact evidence spans or structured attributes for the classifier instead of the full free-text reasoning. This would reduce noise and improve interpretability of the final decisions.

19. Conclusion

This paper presented the Ordantis system for EXIST 2026 Task 2. The system combines learning with disagreement, Gemini-based semantic enrichment, multilingual transformer encoders, EEG representations and Ekman emotion features. It uses task-specific heads for binary sexism detection, hierarchical intention classification and conditional multi-label categorization.

The official results show that the approach is strong for Task 2.1 and Task 2.2, especially in the soft setting. The best soft runs ranked first in both tasks, and the best Task 2.2 hard run also ranked first. The results also show that Gemini probability blending is a powerful strategy for calibration and generalization. However, Task 2.3 remained difficult, indicating that multi-label rare-category prediction requires more robust modeling than the current conditional head and sampler strategy.

The broader conclusion is that LLMs have strong potential for multimodal sexism characterization in memes. Their value is not limited to zero-shot classification; they can also translate visual, cultural and pragmatic information into structured language that supervised models can exploit. Nevertheless, LLMs should be integrated carefully, validated against task-specific data and combined with calibration-aware supervised learning, especially in sensitive domains involving subjective judgments.

Final Note on PyEvALL Evaluation

The official leaderboard values reported in this paper were taken from the final EXIST 2026 result files, which were computed with the PyEvALL evaluation framework. This is relevant because PyEvALL is the reference implementation used for the official ICM and ICM-Soft scores, and these scores may differ from auxiliary validation metrics such as F1, AUC or local threshold-search results.

For this reason, throughout the paper a distinction is made between validation behavior, which guided model design and run selection during development, and official test behavior, which is the final criterion for ranking the submitted systems. In particular, some runs that looked strongest on validation were not always the best according to the official PyEvALL-based scores. This explains the importance of submitting diverse variants: XLM-RoBERTa and Longformer backbones, models with and without reasoning, models with and without sampler, calibrated models, and probability blends with Gemini.

Consequently, the final interpretation of the system should be based on the official PyEvALL scores reported in the results tables, while the development metrics should be understood as diagnostic evidence used to motivate the architecture and the submitted runs.

Acknowledgments

The authors thank the organizers of EXIST 2026 for providing the dataset and evaluation framework. This work was developed by the Ordantis team as part of the EXIST 2026 shared task.

Declaration on Generative AI

During the preparation of this work, the authors used Claude Sonnet 4.6 (Anthropic) and Claude Opus 4.7 (Anthropic) as writing and coding assistants. Claude Sonnet 4.6 was used in order to: grammar and spelling check, paraphrase and reword passages, and translation of passages from Spanish to English. Claude Opus 4.7 was used in order to: generate and debug specific Python code fragments for data preprocessing, model training, and result analysis, always from precise and detailed specifications provided by the authors describing exactly the desired behaviour and inputs/outputs. After using these tools, the authors reviewed and edited all content and code as needed and take full responsibility for the publication's content.

A. Gemini Prompt

For reproducibility, this appendix reproduces the exact prompt used to query Gemini flash 3.1 in the offline precomputation stage described in Section 5.2. The prompt is sent multimodally, together with the meme image, in a single API call per meme, and covers the three subtasks at once. Gemini's response is constrained to a JSON object (`response_mime_type="application/json"`), and the four safety filters are set to `BLOCK_NONE` so the model is allowed to analyze sexist content without being blocked. The placeholder `{ocr_text}` is replaced at runtime with the meme's OCR text.

```
You are an expert annotator analyzing a meme for the EXIST 2026 shared task
at CLEF, which focuses on automatic detection of sexism in social media memes.

You will analyze this meme according to THREE complementary tasks:

TASK 1 - Binary sexism detection
A meme is SEXIST if it expresses, describes, perpetuates or criticizes sexist
behavior, stereotypes or discrimination against women.

TASK 2 - Author intention (only relevant if sexist)
- DIRECT: the author endorses or perpetuates sexism. Sexist content presented
  as message.
- JUDGEMENTAL: the author criticizes or denounces sexism. Uses irony, sarcasm,
  or contrast to expose sexist behavior. Look for contradictions between image
  and text, mocking tone, hashtags used ironically, showing absurdity of
  sexist beliefs.

TASK 3 - Categories of sexism (multi-label, can have several or none)
- IDEOLOGICAL-INEQUALITY: denies inequality, discredits feminism, claims men
  are oppressed.
- STEREOTYPING-DOMINANCE: women in submissive roles, "women's place is
  kitchen", weak/emotional women, men as dominant.
- OBJECTIFICATION: women reduced to body parts, focus on physical attributes
  only, depersonalization.
- SEXUAL-VIOLENCE: sexual harassment, assault, rape culture, unwanted sexual
  advances.
- MISOGYNY-NON-SEXUAL-VIOLENCE: hatred toward women, non-sexual aggression,
```

```
 gendered insults.

OCR text extracted from the meme: "{ocr_text}"

INSTRUCTIONS:
1. Examine image and text carefully.
2. Pay attention to context, irony, sarcasm - especially contradictions
  between visual and textual elements.
3. A meme can have multiple categories simultaneously (Task 3).
4. If the meme is NOT sexist, set task2_2.intention to "NO" and
  task2_3.categories_present to [].

Return ONLY a valid JSON object:
{
 "description": "<brief literal description of image and text>",
 "sexism_analysis": "<analysis of why or why not sexist>",
 "reasoning": "<step-by-step reasoning>",
 "task2_1": {
  "sexist_probability": <float 0.0-1.0>,
  "confidence": <float 0.0-1.0>
 },
 "task2_2": {
  "intention": "<NO | DIRECT | JUDGEMENTAL>",
  "intention_probabilities": {
   "NO": <float>, "DIRECT": <float>, "JUDGEMENTAL": <float>
  },
  "intention_reasoning": "<why this intention>",
  "irony_detected": <true | false>,
  "irony_confidence": <float 0.0-1.0>
 },
 "task2_3": {
  "categories_present": [<list of categories that apply, can be empty>],
  "category_probabilities": {
   "IDEOLOGICAL-INEQUALITY": <float>,
   "STEREOTYPING-DOMINANCE": <float>,
   "OBJECTIFICATION": <float>,
   "SEXUAL-VIOLENCE": <float>,
   "MISOGYNY-NON-SEXUAL-VIOLENCE": <float>
  },
  "category_reasoning": "<justification for each category>"
 }
}
intention_probabilities must sum to 1.0. Category probabilities are
independent (each in 0.0-1.0).
```

The response fields are mapped to the pipeline as follows. The `description`, `sexism_analysis` and `reasoning` strings are concatenated with the OCR text and fed to the multilingual transformer encoder as the enriched text input. The numerical estimates in `task2_1`, `task2_2` and `task2_3` provide the auxiliary Gemini features (seven for Task 2.2, six for Task 2.3) that are concatenated at the fusion layer, and, in the blended runs, the zero-shot probabilities that participate in the linear combinations described in Section 5.2.

References

[1] Roberto Labadie-Tamayo, Berta Chulvi, and Paolo Rosso. Everybody hurts, sometimes. Overview of HUrtful HUmour at IberLEF 2023: Detection of humour spreading prejudice in Twitter. In *Procesamiento del Lenguaje Natural (SEPLN)*, volume 71, pages 383–395, 2023.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pages 5998–6008, 2017.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, 2019.

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[5] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020.

[6] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

[10] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 2611–2624, 2020.

[11] Iván Arcos, Paolo Rosso, and Elena Gomis-Vicent. Human-centered multimodal fusion for sexism detection in memes with Eye-Tracking, Heart Rate, and EEG signals. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2026)*, 2026.

[12] Laura Plaza, Jorge Carrillo-de Albornoz, Iván Arcos, María Aloy-Mayo, Paolo Rosso, Enrique García-Arias, and Damiano Spina. Overview of EXIST 2026: Physiological data for multimodal sexism characterization in social media. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026)*, 2026.

[13] Laura Plaza, Jorge Carrillo-de Albornoz, Iván Arcos, María Aloy-Mayo, Paolo Rosso, Enrique García-Arias, and Damiano Spina. Overview of EXIST 2026: Physiological data for multimodal sexism characterization in social media (extended overview). In Eva Sánchez Salido, Alberto Barrón-Cedeño, Alba García Seco de Herrera, Sean MacAvaney, and Julia Maria Struß, editors, *CLEF 2026 Working Notes*, 2026.

[14] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992, 2019.

[15] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3–4):169–200, 1992.

[16] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3744–3753, 2019.

[17] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

[18] Paolo Italiani, David Gimeno-Gómez, Luca Ragazzi, Gianluca Moro, and Paolo Rosso. MemeWeaver: Inter-meme graph reasoning for sexism and misogyny detection. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2120–2134, 2026.

[19] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009.

[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.