CYUT at EXIST 2026: GPT-Enhanced Multilingual Ensemble for Sexism Detection in Memes

Notebook for the EXIST Lab at CLEF 2026

Shih-Hung Wu*, Yu-Feng Huang

Chaoyang University of Technology, Taichung, Taiwan (R.O.C)

Abstract

In this paper, we participated in the CLEF 2026 EXIST shared task and describe the submission methodology of the CYUT-Giantsshoulders_1 system for the EXIST 2026 Memes track, with a primary focus on Task 2.1, Sexism Identification, and Task 2.2, Source Intention Detection. Detecting sexism in memes is highly challenging because sexist meanings are often conveyed through the interaction of visual content, embedded text, cultural context, sarcastic expressions, and implicit stereotypes. Inspired by ArcosGPT at EXIST 2025, this study explores a GPT-based semantic augmentation approach to enhance multimodal meme understanding.

Our system first employs GPT-4o to extract OCR text, contextual descriptions, and analytical reasoning from meme images and tweet text. These generated features are then concatenated with the original text to construct a richer textual representation. To improve the robustness of the model on both English and Spanish memes, our final approach adopts a language-aware weighted ensemble strategy. For English samples, we combine RoBERTa-large and XLM-RoBERTa-large; for Spanish samples, we use XLM-RoBERTa-large and BERTIN Spanish RoBERTa. Each model type is trained with multiple random seeds, and their probability outputs are first averaged before applying language-specific weighted fusion.

On the official leaderboard, CYUT-Giantsshoulders_1 achieved an ICM-Hard Norm score of 0.6527 and an F1 score of 0.7590 under the Hard-Hard evaluation over all instances setting for Task 2.1, ranking 11th overall. For Task 2.2, our system obtained an ICM-Hard Norm score of 0.5231 and an F1 score of 0.5041, ranking 10th overall. These results indicate that GPT-enhanced textual features, combined with a language-aware ensemble strategy, are effective for sexism identification in memes and provide competitive performance in the source intention classification task.

Keywords

Sexism Characterization, GPT-4o, Large Language Models (LLMs), Multilingual Transformers, Ensemble Learning

1. Introduction

Sexism on social media is often expressed not only through explicit offensive language, but also through implicit stereotypes, sarcasm, humor, and references that depend on cultural context. The EXIST shared task aims to promote research on the automatic detection of sexism in social networks, covering a wide range of phenomena from explicit misogynistic discourse to more subtle forms of implicit sexist behavior [1]. Compared with purely textual posts, memes introduce additional challenges because their meanings are often constructed through the interaction among visual content, embedded text, and social context. When only the textual content is considered, a meme may appear neutral; however, when the image and text are interpreted together, it may convey sexist assumptions or stereotypes.

This paper describes the submission methodology of the CYUT-Giantsshoulders_1 system for the EXIST 2026 Memes track. This study focuses on Task 2.1, Sexism Identification, and Task 2.2, Source Intention Detection. The objective of Task 2.1 is to determine whether a given meme contains sexist content, with the output labels being YES and NO. Task 2.2 further identifies the source intention of the sexist content. In our system, Task 2.2 is handled using a hierarchical prediction strategy: if a meme is predicted as non-sexist in Task 2.1, its corresponding Task 2.2 output is directly assigned as NO; otherwise, the system further predicts whether the sexist content belongs to the DIRECT or

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

*Corresponding author.

shwu@cyut.edu.tw (S. Wu); s11227615@gm.cyut.tw (Y. Huang)

0000-0002-1769-0613 (S. Wu)

JUDGEMENTAL category. This design reflects the dependency between sexism identification and source intention classification.

This study is primarily inspired by ArcosGPT at EXIST 2025, which explored the use of generative visual and semantic descriptions for sexism detection in memes. ArcosGPT demonstrated that enriching OCR-based meme representations with image captions and GPT-4o-generated descriptions can improve macro-F1 performance compared with text-only baseline models[2]. Following this direction, this study employs GPT-4o as a multimodal semantic feature extractor. For each meme, GPT-4o receives both the meme image and the original tweet text, and generates structured information such as OCR text, contextual descriptions, and analytical reasoning. These generative features are concatenated with the original tweet to form a richer textual representation, which is then used as input to downstream classifiers.

To improve the robustness of the model on both English and Spanish memes, this study proposes a language-aware weighted ensemble approach. Since the dataset contains multilingual content, a single encoder may not be the optimal choice across all language settings. Therefore, our final system combines multilingual and language-specific Transformer encoders. For English samples, this study combines RoBERTa-large and XLM-RoBERTa-large. RoBERTa is a robustly optimized variant of BERT that improves the pre-training procedure and has been widely used as a strong English text encoder [3]. XLM-RoBERTa is a multilingual masked language model trained on large-scale multilingual CommonCrawl data, making it suitable for cross-lingual transfer tasks involving languages such as English and Spanish [4]. For Spanish samples, this study combines XLM-RoBERTa-large with BERTIN Spanish RoBERTa, the latter being a language-specific encoder designed for Spanish [5].

The final ensemble is performed at the probability level. Each model type includes three checkpoints trained with different random seeds, and their output probabilities are first averaged to reduce instability caused by seed variation. Language-specific weights are then applied. For English samples, RoBERTa-large and XLM-RoBERTa-large are combined with weights of 0.7 and 0.3, respectively. For Spanish samples, XLM-RoBERTa-large and BERTIN Spanish RoBERTa are combined with weights of 0.8 and 0.2, respectively. The final probabilities are used to generate predictions for Task 2.1 and Task 2.2, and hierarchical correction is applied to ensure consistency between the two tasks.

The main contributions of this paper are threefold. First, this study extends the ArcosGPT-style GPT-based semantic augmentation strategy to the EXIST 2026 Memes setting by converting meme images and tweet text into OCR, contextual descriptions, and reasoning-based textual features. Second, this study proposes a language-aware ensemble approach that combines multilingual and language-specific encoders for English and Spanish memes. Third, this study demonstrates that seed averaging and language-specific weighted fusion can provide competitive official results in both sexism identification and source intention detection.

On the official EXIST 2026 leaderboard, CYUT-Giantsshoulders_1 achieved an ICM-Hard Norm score of 0.6527 and an F1 score of 0.7590 under the Hard-Hard evaluation over all instances setting for Task 2.1, ranking 11th overall. For Task 2.2, our system obtained an ICM-Hard Norm score of 0.5231 and an F1 score of 0.5041, ranking 10th overall. Compared with ArcosGPT_1 at EXIST 2025, our system achieved a comparable F1 score on Task 2.1 and a slightly higher ICM-Hard Norm score on Task 2.2. These results suggest that GPT-enhanced textual features combined with language-aware weighted ensembling constitute a competitive research direction for sexism detection in memes.

2. Related Work

2.1. Sexism Detection and Learning with Disagreement in EXIST

The EXIST shared task primarily focuses on sexism identification and characterization in social media, covering tweets, memes, and other multimodal content. One important characteristic of recent EXIST tasks is the adoption of a learning with disagreement setting, in which systems are encouraged to model annotator disagreement rather than relying solely on aggregated hard labels[6].

2.2. GPT-enhanced Multimodal Meme Analysis

The most closely related prior work to this study is ArcosGPT at EXIST 2025, which investigated a GPT-4o-based semantic augmentation approach and applied it to sexism detection in memes [2]. Their results showed that incorporating generative image descriptions and GPT-4o-generated explanations into textual representations significantly improved macro-F1 performance, further demonstrating the importance of transforming the visual context of memes into textual semantic features.

2.3. Multilingual Transformer Encoders for Text Classification

In our language-aware ensemble approach, RoBERTa-large is used to process English samples, XLM-RoBERTa-large serves as a multilingual encoder for both English and Spanish samples, and BERTIN Spanish RoBERTa is employed as a language-specific encoder for Spanish samples.

• **RoBERTa-large.** RoBERTa is a robustly optimized variant of BERT that improves BERT's performance through more extensive pre-training strategies and larger training data. We use RoBERTa-large for English samples because it has been shown to be a strong English text encoder. The RoBERTa paper argues that BERT was undertrained and demonstrates that stronger performance can be achieved through larger batch sizes, longer training, and improved training configurations [3].

• **XLM-RoBERTa-large.** XLM-R is a large-scale multilingual masked language model trained on data from 100 languages and more than 2 TB of CommonCrawl corpora, making it suitable for English–Spanish cross-lingual settings. The XLM-R paper shows that large-scale multilingual pre-training can improve performance across a wide range of cross-lingual transfer tasks [4].

• **BERTIN Spanish RoBERTa.** BERTIN is a RoBERTa-based model trained specifically for Spanish. According to its Hugging Face model card, it is a RoBERTa-base model trained from scratch on the Spanish portion of mC4. The MarIA Spanish language models paper also describes BERTIN as a Spanish RoBERTa-base model with 12 layers, 12 attention heads, and a hidden size of 768 [5].

2.4. Vision-language Models and Multimodal Fusion

In addition to textual augmentation, several EXIST systems have explored direct multimodal fusion approaches. UMUTeam combined Transformer-based text encoders, visual representations, and soft-label learning, showing that jointly modeling both modalities and annotator disagreement is beneficial for sexism detection in memes. The CLIP paper proposed learning transferable visual models from natural language supervision by pre-training on large-scale image–text pairs, enabling transfer to a wide range of downstream vision tasks. UMUTeam at EXIST 2025 employed multimodal Transformer architectures and soft-label learning for sexism detection, making it the closest setting to approaches based on XLM-R combined with CLIP or ViT representations and soft labels [7].

2.5. Generative LLMs and Parameter-efficient Fine-tuning

LoRA substantially reduces the number of trainable parameters required for downstream tasks by freezing the pre-trained weights and inserting trainable low-rank matrices, while QLoRA further enables efficient fine-tuning by combining LoRA with a 4-bit quantized pre-trained model [8]. GrootWatch and I2C-UHU-Altair both employed Qwen- or Qwen-VL-related methods at EXIST 2025 to process meme images or generate descriptions before performing classification [9]. In particular, GrootWatch used a large-scale vision-language model to process meme images; however, such approaches typically require substantial computational resources. Therefore, this study adopts an alternative route that is comparatively lightweight and more feasible [10].

3. Dataset and Task Description

3.1. Dataset

This study uses the official EXIST 2026 Memes dataset for the experiments. The original training data consist of 4,044 samples with a bilingual distribution, including 2,034 Spanish samples and 2,010 English samples. The official test set contains 1,053 meme samples, including 513 English samples and 540 Spanish samples. Taken together with the training data, these statistics indicate that the EXIST Memes dataset is generally balanced between English and Spanish, while still preserving slight variations in language distribution across different data splits.

Table 1
Language distribution of the original dataset.

Dataset	Total	English	Spanish
Training Dataset	4044	2010	2034
Test Dataset	1053	513	540

3.2. Task Definitions

This study primarily focuses on and evaluates the following two core tasks:
• Task 2.1 (Sexism Identification): This is a binary classification task that aims to automatically determine whether an input meme contains sexist content, with the output label being either YES or NO.
• Task 2.2 (Source Intention): This is an intention classification task that further analyzes the author's communicative intention for memes identified as sexist. The categories include non-sexist content (NO), direct sexism (DIRECT), and judgmental sexism (JUDGEMENTAL).
• Auxiliary Task 2.3 (Sexism Categorization): This is a multi-label classification task designed to identify specific types of sexism, such as SEXUAL-VIOLENCE and MISOGYNY. In this study, this task is mainly used as an auxiliary training objective within the multi-task learning framework, enabling the model to implicitly learn more fine-grained semantic representations of sexist content.

4. Methodology

For the EXIST 2026 Memes task, this study designs and evaluates several machine learning and deep learning architectures. The following sections provide a detailed description of the five system approaches proposed in this study.

4.1. Baseline

In the baseline experiments, this study constructs a basic multimodal dual-stream architecture.

• Textual branch: The raw OCR text extracted from each meme is used as input and encoded using a pre-trained XLM-RoBERTa-base model to extract high-level semantic features.
• Visual branch: The meme image is fed into a pre-trained ResNet-50 network to extract spatial visual features.
• Feature fusion and training: The model concatenates the textual feature vector and the visual feature vector, which are then passed into a multimodal fusion layer to produce a joint representation. This study adopts a multi-task learning (MTL) framework, in which three independent classification heads are attached after the fusion layer to jointly train Task 2.1 (binary sexism classification), Task 2.2

(three-class intention classification), and Task 2.3 (multi-label sexism categorization). During training, the gold labels in the training set are used as supervision signals. Cross-entropy loss is used for Task 2.1 and Task 2.2, while binary cross-entropy with logits loss is used for Task 2.3. The final optimization objective for backpropagation is defined as the sum of the losses from the three tasks.

4.2. Method 1: GPT-4o Feature Augmentation

This approach is primarily based on the large language model (LLM)-based semantic feature augmentation strategy proposed by ArcosGPT at EXIST 2025 [2]. That study indicated that incorporating high-level semantic descriptions of images into OCR text can effectively compensate for the limitations of conventional visual models in understanding complex meme contexts, such as sarcasm and metaphor.

• GPT-4o semantic extraction: This study adapts this concept to the EXIST 2026 task by providing both the meme image and the original text as input to GPT-4o, using it as a multimodal semantic feature extractor. The model is instructed to generate structured textual features in JSON format, including contextual descriptions (Context_Description) and analytical reasoning (Analytical_Reasoning).
• Feature concatenation and classification: Specifically, this approach concatenates the original text with the contextual descriptions and analytical reasoning generated by GPT-4o, forming a substantially enriched composite textual input. This input is then fed into the larger XLM-RoBERTa-large classification model, which is also trained under a multi-task learning framework to predict the three tasks jointly.

4.3. Method 2: GPT-4o/RAG + XLM-R + CLIP/ViT Fusion

Compared with Method 1, which relies only on textualized features, Method 2 further constructs a disagreement-aware multimodal multi-task learning framework that jointly considers textual semantics, raw visual embeddings, and label uncertainty.

Step 1: GPT-4o + RAG feature generation. To enable the LLM to interpret the sociocultural context behind memes more accurately, this study incorporates the concept of retrieval-augmented generation (RAG) into the prompt design by introducing background knowledge query cues for common meme templates, such as “kitchen,” “stay-at-home,” and “cat-lady.” After integrating these contextual cues, GPT-4o generates five structured features for each meme: image text `OCR_Text`, context descriptions enriched with background knowledge `Context_Description`, image–text relational reasoning `Analytical_Reasoning`, implicit bias analysis `Implicit_Bias`, and author intention analysis `Author_Intent`.

Step 2: XLM-RoBERTa-large textual encoding. On the textual input side, this method deeply integrates multiple types of features. During validation and inference, the input text consists of the original text (RAW_TEXT), OCR text, GPT-generated contextual descriptions, and GPT-generated reasoning content. This long-form textual input is then fed into XLM-RoBERTa-large to obtain deep textual vector representations.

Step 3: CLIP-ViT visual encoding. On the visual input side, this study uses CLIP-ViT, specifically openai/clip-vit-base-patch32, as the main visual encoder. The model directly processes meme images and extracts dense visual feature representations, while the implementation framework also preserves the flexibility to switch to a standard ViT-base model.

Step 4: Soft-label multi-task training. This method concatenates the feature vectors from the textual and visual branches and passes them into a fusion layer and task-specific classification heads. To align with the core setting of learning with disagreement in the EXIST shared task, Method 2 discards hard

targets and directly uses the official soft labels as training objectives. Task 2.1 and Task 2.2 use soft cross-entropy loss, implemented as KL divergence, to fit the probability distributions derived from multiple annotators, while Task 2.3 uses BCEWithLogitsLoss.

Step 5: Hierarchical inference correction. During prediction, this study implements the `apply_hierarchy()` rule for logical correction to ensure the validity of multi-task outputs. If Task 2.1 is predicted as NO, indicating non-sexist content, then the outputs of Task 2.2 and Task 2.3 are forcibly corrected to NO. If Task 2.1 is predicted as YES, indicating sexist content, the model then uses the actual predicted probabilities from the downstream Task 2.2 intention classification and Task 2.3 category prediction heads.

4.4. Method 3: Qwen2.5-7B LoRA Generative Classification

Method 3 follows the approach of the I2C team. Considering the strong capabilities of recent large language models (LLMs) in contextual understanding and instruction following, this method reformulates sexism detection as a generative classification task [11].

• Preprocessing and feature extraction: This method does not call GPT-4o in real time during inference. Instead, during the data preprocessing stage, the original tweet text, OCR text, and previously generated structured analytical features from GPT-4o, including contextual descriptions, author intention, implicit bias, and analytical reasoning, are integrated into a complete prompt.
• Model fine-tuning and inference: This study uses Qwen2.5-7B-Instruct as the core generative foundation model and applies parameter-efficient fine-tuning (PEFT) through low-rank adaptation techniques, namely LoRA/QLoRA. During training, the official soft labels are converted into hard targets by selecting the class with the highest probability, and the Qwen2.5 model is trained to directly generate standardized JSON-formatted text containing the prediction results for task2.1, task2.2, and task2.3 based on the input prompt. During inference, after the model outputs the JSON labels, the aforementioned hierarchical rules are also applied for logical parsing and correction to ensure output validity.

4.5. Method 4: XLM-R + CLIP Cross-Attention Fusion

Method 4 is inspired by the approach of UMUTeam. To further model the cross-modal interaction between textual tokens and visual patches in memes, Method 4 replaces simple late fusion with a cross-attention fusion architecture. This method is primarily inspired by and adapted from both UMUTeam and ArcosGPT [2][7].
• Cross-modal encoding: The textual input consists of the original tweet, OCR text, and GPT-generated semantic features, including contextual descriptions, author intention, implicit bias, and analytical reasoning. This input is encoded by XLM-RoBERTa-large into token-level representations. On the visual side, CLIP-ViT, specifically openai/clip-vit-base-patch32, is used to extract image patch-level visual embeddings.
• Cross-attention mechanism: During the fusion stage, this study projects the CLIP visual patch embeddings into the same hidden dimension as XLM-R. A cross-attention layer is then constructed using nn.MultiheadAttention, where the token embeddings from XLM-RoBERTa-large are used as the Query (Q), and the image patch embeddings from CLIP are used as the Key (K) and Value (V). This enables the textual representations to actively attend to relevant image regions during the fusion process.
• Fusion and soft-label training: The model extracts the textual [CLS] representation after the cross-attention operation and concatenates it with the original textual [CLS] representation. The concatenated representation is then passed through a fusion multi-layer perceptron (Fusion MLP) to

produce a joint multimodal representation, which is finally fed into the three task-specific classification heads. This method also adopts soft-label learning. For Task 2.1 and Task 2.2, KL divergence loss (KLDivLoss) is used to learn the annotator distribution, and temperature scaling is applied to the soft labels during training to control the smoothness of the target distribution. Hierarchical correction is also applied during inference.

4.6. Method 5: Language-aware Weighted Ensemble

This method, namely the *Language-aware Seed-averaged Weighted Ensemble*, serves as the final submitted system in this study. Its core motivation is to combine the complementary strengths of multilingual general-purpose models and language-specific models while reducing prediction instability on a relatively small meme dataset through multi-seed ensembling, as illustrated in Figure 1.

Input features: This method continues to use GPT-4o-generated semantic features as input augmentation by concatenating the original tweet text, OCR text, contextual descriptions, and analytical reasoning generated by GPT-4o.

Language-aware decision making and weighted fusion: This method dynamically adopts different text encoder groups and weight configurations for English (EN) and Spanish (ES) samples.

- **English samples (EN):** RoBERTa-large and XLM-RoBERTa-large are combined using weights of 0.7 and 0.3, respectively.
- **Spanish samples (ES):** XLM-RoBERTa-large and BERTIN Spanish RoBERTa, a language-specific model designed for Spanish, are combined using weights of 0.8 and 0.2, respectively.

Seed averaging: To maximize generalization ability, each model family contains three checkpoints trained using different random seeds. During inference, the system first computes the arithmetic mean of the softmax probabilities produced by the three checkpoints within the same model family. The averaged probabilities are then combined using the language-specific weighted fusion strategy described above. For Task 2.1 and Task 2.2, the final class label is selected according to the highest weighted probability. For the auxiliary Task 2.3, category-specific thresholds are applied to generate multi-label predictions.

The GPT-4o feature construction strategy adopted in this study is inspired by the ArcosGPT system proposed for the EXIST 2025 shared task [2]. ArcosGPT demonstrated that GPT-generated semantic descriptions can effectively enhance sexism detection in memes by transforming multimodal information into textual representations suitable for downstream classification models.
In their approach, GPT-4o is mainly employed to generate a high-level description of the meme, identify whether the content involves sexism, infer the possible intention of the author, and provide a general explanation supporting the analysis. These generated descriptions serve as semantic augmentations that complement the original textual content.

Although our work follows the same general principle of leveraging GPT-4o as a semantic feature extractor, the proposed prompt design differs substantially from ArcosGPT in both output structure and task orientation. Rather than requesting a free-form textual description, GPT-4o is instructed to return a structured JSON object containing multiple task-specific semantic fields. The generated features include `OCR_Text`, `Context_Description`, `Sexism_Identification`, `Author_Intent`, `Analytical_Reasoning`, `Implicit_Bias`, and `Sentiment`.

Although GPT-4o generates fields such as `Sexism_Identification` and `Author_Intent`, these outputs are used solely as semantic contextual features rather than as ground-truth labels or direct task predictions. They are concatenated with the other generated features to construct an enriched textual representation. Final predictions for Task 2.1 and Task 2.2 are produced exclusively by the downstream fine-tuned Transformer classifiers trained on the official training data.

This structured representation provides a more consistent feature space across samples and facilitates direct integration with Transformer-based classification models.

Specifically, `OCR_Text` captures textual content embedded within the meme image, which often contains crucial information related to humor, sarcasm, stereotypes, or discriminatory messages. `Context_Description` summarizes the overall meme scenario by jointly considering both the image and the accompanying tweet text.
`Sexism_Identification` provides an initial semantic assessment of whether sexist content is present, while `Author_Intent` is designed to support Task 2.2 by explicitly modeling the communicative intention behind the meme. Furthermore, `Analytical_Reasoning` and `Implicit_Bias` aim to capture deeper semantic cues, including implicit stereotypes, cultural references, multimodal interactions, and hidden forms of sexism that may not be directly observable from surface-level text alone. Finally, `Sentiment` characterizes the overall tone of the meme, such as humor, sarcasm, anger, or neutrality.

After feature generation, selected fields are concatenated with the original tweet text to construct an enriched textual representation. In the final submitted system, the primary features used for downstream classification include the original tweet text, `OCR_Text`, `Context_Description`, and `Analytical_Reasoning`. These features are concatenated using special separators and subsequently encoded by multilingual Transformer models. Through this design, multimodal information is converted into structured semantic features that can be effectively exploited by text-based classifiers without requiring direct end-to-end multimodal training.

Therefore, compared with ArcosGPT, the proposed approach can be viewed as a task-oriented extension of GPT-based semantic augmentation. While ArcosGPT mainly relies on general semantic descriptions, our method explicitly decomposes GPT outputs into multiple semantically meaningful fields aligned with the objectives of Task 2.1 and Task 2.2. This design improves feature consistency, interpretability, and task relevance before the information is processed by downstream classification models. To provide an overview of the proposed system, Figure 1 presents the complete processing pipeline adopted in this study.

The framework consists of three major stages: GPT-4o semantic feature construction, language-aware weighted ensemble classification, and task-specific prediction generation.

To improve the consistency and task relevance of GPT-generated semantic features, a structured prompt template was designed. The prompt instructs GPT-4o to jointly analyze the meme image and the original tweet text and return a JSON object containing seven task-oriented semantic fields. The complete prompt template used in this study is provided in Appendix A.

5. Experimental Setup

5.1. Dataset and Splitting

This study uses the EXIST 2026 Memes dataset as the core experimental dataset, which contains meme images and their corresponding OCR text. The original training data consist of 4,044 samples, including 2,034 Spanish (ES) samples and 2,010 English (EN) samples. To conduct local model validation and hyperparameter tuning, this study independently splits the original training data into a training set and a development set at a ratio of 7:3, resulting in 2,832 training samples and 1,212 development samples. The development set maintains a balanced distribution, with 606 English samples and 606 Spanish samples.

5.2. Implementation Details

All experiments in this study were implemented using the PyTorch framework. For the discriminative classification models (Method 1, Method 2, Method 4, and Method 5), pre-trained Transformer encoders were obtained from the Hugging Face model repository. Specifically, XLM-RoBERTa-large was adopted

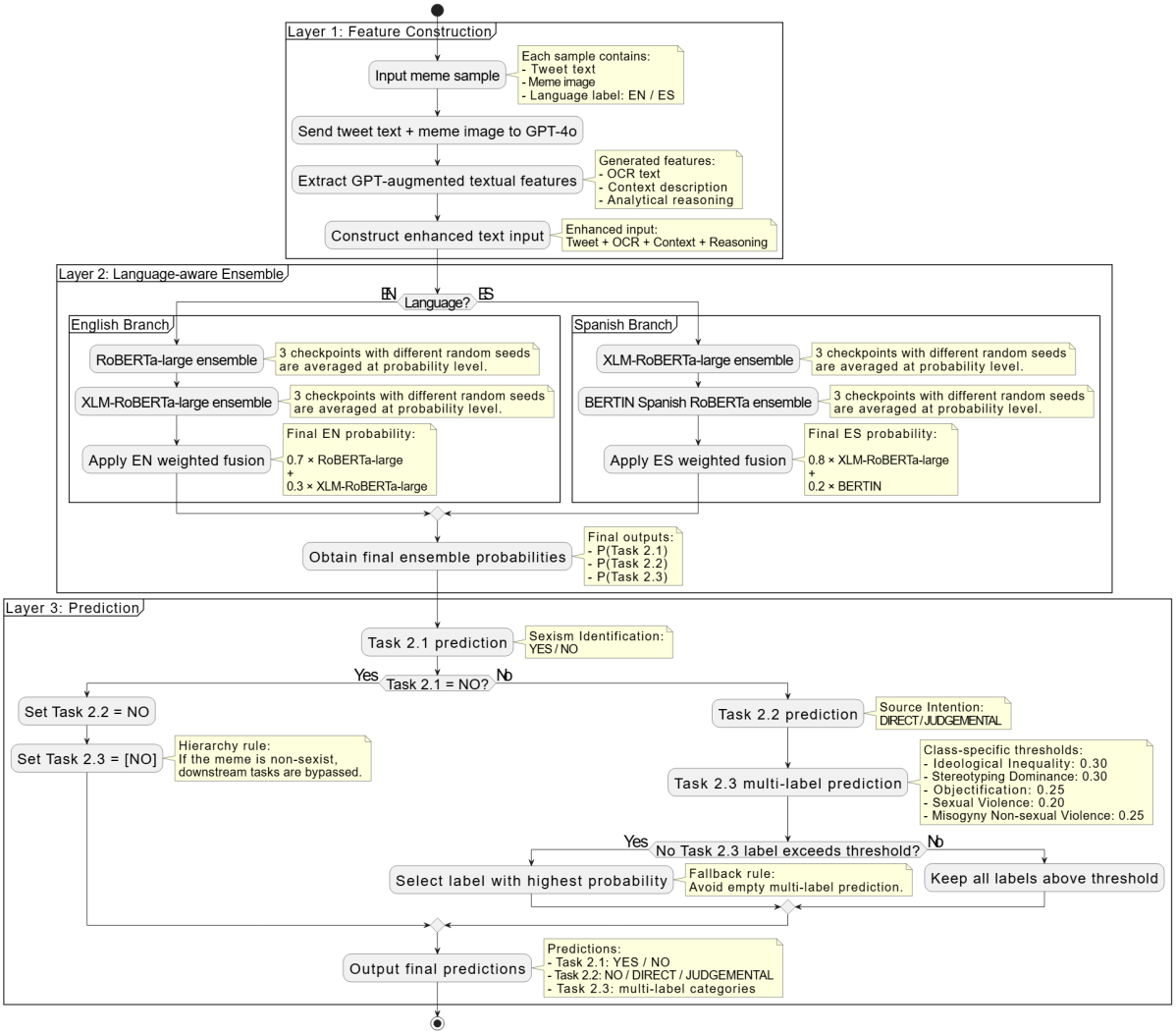

Figure 1: Overview of the proposed Language-aware Seed-averaged Weighted Ensemble framework.

Table 2
This table shows the distribution of the dataset after splitting the training data.

Dataset	Total	English	Spanish	Ratio
Train set	2832	1404	1428	70%
Development set	1212	606	606	30%
Total	4044	2010	2034	100%

as the primary multilingual encoder, RoBERTa-large was used as an English-specific encoder, and BERTIN Spanish RoBERTa was employed as a Spanish-specific encoder. For multimodal experiments, CLIP-ViT was utilized as the visual feature extractor.

Before feature construction, the original tweet text and meme images provided in the official EXIST 2026 dataset were used directly without additional text preprocessing or normalization. GPT-4o jointly analyzed the meme image and the original tweet text to generate structured semantic features, including `OCR_Text`, `Context_Description`, and `Analytical_Reasoning`. The `OCR_Text` field was extracted directly from the meme image by GPT-4o. These generated features were concatenated with the original tweet text to construct the final textual input for downstream classification. The resulting sequence was tokenized using the corresponding Transformer tokenizer with a maximum sequence length of 512 tokens. Sequences exceeding this limit were truncated.

Unless otherwise specified, all discriminative models were trained using the AdamW optimizer with an initial learning rate of 2×10^{-5}. The batch size was set to either 16 or 32 depending on the available GPU memory.

For Method 3, which adopts a generative classification paradigm, Qwen2.5-7B-Instruct was fine-tuned using QLoRA under 4-bit quantization. The LoRA rank (r) was set to 8 and the scaling factor (α) was set to 16. This configuration significantly reduces GPU memory consumption while maintaining competitive fine-tuning performance, enabling efficient adaptation of a 7B-parameter large language model under limited computational resources.

For the final submitted system (Method 5), a language-aware weighted ensemble strategy was employed. RoBERTa-large and XLM-RoBERTa-large were combined for English samples, while XLM-RoBERTa-large and BERTIN Spanish RoBERTa were combined for Spanish samples. Each model family was trained using three different random seeds, and the resulting probability distributions were averaged prior to weighted fusion. This seed-averaging strategy was adopted to improve robustness and reduce variance caused by random initialization during training.

6. Experiment Result

6.1. Internal Development Results

To evaluate the effectiveness of each method, this study first conducts a comprehensive evaluation on the locally split development set. The Macro-F1 scores of each method on Task 2.1, Sexism Identification, and Task 2.2, Source Intention Classification, are shown in Table 3.

Table 3
Performance comparison of different methods on the local development set.

Method	Task 2.1 Macro-F1	Task 2.2 Macro-F1	Task 2.3 Macro-F1
Baseline: XLM-R-base + ResNet-50 MTL	0.5096	0.3535	0.0033
Method 1: GPT-4o Feature Augmentation + XLM-R-large	0.7800	0.4900	–
Method 2: GPT-4o/RAG + CLIP Late Fusion	0.7589	0.5584	0.3615
Method 3: Qwen2.5-7B LoRA Generative Classification	0.6002	0.3657	0.3925
Method 4: XLM-R-large + CLIP Cross-Attention Fusion	0.7571	0.4828	0.3129
Method 5: Language-aware Weighted Ensemble	**0.7783**	**0.6067**	**0.6004**

Based on the experimental results in Table 3, this study derives the following key findings:

1. The critical role of GPT-based semantic feature augmentation: All methods that incorporate GPT-4o-derived features (Methods 1, 2, 4, and 5) significantly outperform the baseline model, which does not use such features. This confirms that, for meme content that is highly context-dependent and contains implicit bias, high-level contextual descriptions and reasoning generated by large language models can provide crucial decision-making cues.

2. The improvement of multimodal fusion for intention classification (Task 2.2): Compared with Method 1, which uses only textual features, Method 2, which incorporates CLIP visual embeddings and soft-label training, substantially improves the Macro-F1 score for Task 2.2 from 0.49 to 0.5584. This indicates that visual features from images provide complementary information for identifying fine-grained author intentions, such as DIRECT or JUDGEMENTAL.

3. Limitations of complex fusion mechanisms: Method 4, which adopts a cross-attention mechanism, does not outperform Method 2, which uses simple late fusion. This suggests that, under conditions of limited data size and severe class imbalance, overly complex cross-modal interaction layers may instead make the model more prone to overfitting.

4. Superior performance of language-aware ensembling: Method 5 achieves the best performance across all tasks by distinguishing between English and Spanish samples and incorporating multi-seed averaging (Task 2.1 F1 = 0.7783; Task 2.2 F1 = 0.6067). Therefore, this study adopts this architecture as

the final decision system submitted to the official evaluation.

6.2. Official Leaderboard Results

This study submitted the final prediction results to the official EXIST 2026 evaluation platform under the team name CYUTGiantsshoulders_1. The official ranking was conducted using the Hard-Hard evaluation setting over all instances, with ICMHard and ICMHard Norm as the core evaluation metrics. The final results and rankings of this study on the official test set are shown in Table 4.

In the official final ranking, our system demonstrated highly competitive performance. It ranked 11th worldwide in the binary sexism identification task (Task 2.1) and achieved 10th place worldwide in the more challenging source intention classification task (Task 2.2).

Table 4
Final results and rankings of CYUTGiantsshoulders_1 on the official EXIST 2026 test set.

Task	Scope	Ranking	ICM-Hard	ICM-Hard Norm	Macro-F1 / F1
Task 2.1 (Sexism Identification)	ALL	11th	0.3003	0.6527	0.7590
Task 2.2 (Source Intention)	ALL	10th	0.0664	0.5231	0.5041

6.3. Comparison with ArcosGPT_1

To further validate the academic value of the proposed system, we compare the official evaluation results with ArcosGPT_1, the core inspiration for this study, across different years [2]. Since EXIST 2025 and EXIST 2026 used the same meme test set, this comparison provides highly informative and meaningful reference value. The comparison results are shown in Tables 5 and 6.

Table 5
Comparison of official Task 2.1 performance between the proposed system and ArcosGPT_1.

Task	Scope	System	Year	Rank	ICM-Hard	ICM-Hard Norm	F1 (YES)
2.1	ALL	ArcosGPT_1	2025	3	0.3200	0.6627	0.7571
2.1	ALL	CYUT-Giantsshoulders_1	2026	11	0.3003	0.6527	**0.7590**
2.1	ES	ArcosGPT_1	2025	3	0.3284	0.6673	**0.7608**
2.1	ES	CYUT-Giantsshoulders_1	2026	12	0.2705	0.6378	0.7483
2.1	EN	ArcosGPT_1	2025	3	0.3109	0.6578	0.7530
2.1	EN	CYUT-Giantsshoulders_1	2026	13	**0.3294**	**0.6673**	**0.7713**

Table 6
Comparison of official Task 2.2 performance between the proposed system and ArcosGPT_1.

Task	Scope	System	Year	Rank	ICM-Hard	ICM-Hard Norm	Macro-F1
2.2	ALL	ArcosGPT_1	2025	3	0.0597	0.5208	**0.5109**
2.2	ALL	CYUT-Giantsshoulders_1	2026	10	**0.0664**	**0.5231**	0.5041
2.2	ES	ArcosGPT_1	2025	3	**0.0344**	**0.5120**	**0.4889**
2.2	ES	CYUT-Giantsshoulders_1	2026	10	0.0176	0.5061	0.4370
2.2	EN	ArcosGPT_1	2025	3	0.0843	0.5292	0.5326
2.2	EN	CYUT-Giantsshoulders_1	2026	10	**0.1150**	**0.5399**	**0.5493**

Based on the detailed comparison in Tables 5 and 6, the following conclusions can be drawn:

• Analysis of Task 2.1

In the ALL evaluation setting for Task 2.1, CYUTGiantsshoulders_1 achieved an F1 score for the YES class of 0.7590, slightly higher than the 0.7571 obtained by ArcosGPT_1. However, its ICM-Hard Norm score was 0.6527, slightly lower than ArcosGPT_1's 0.6627. This result indicates that the proposed method achieves performance comparable to ArcosGPT_1 in identifying the sexist YES class, although its overall information contrast-based evaluation score remains slightly lower. The language-specific results show that the proposed method outperforms ArcosGPT_1 on English samples, but still exhibits a performance gap on Spanish samples.

• Analysis of Task 2.2

In the ALL evaluation setting for Task 2.2, CYUTGiantsshoulders_1 achieved an ICM-Hard Norm score of 0.5231, slightly higher than ArcosGPT_1's 0.5208. However, its Macro-F1 score was 0.5041, slightly lower than ArcosGPT_1's 0.5109. The language-specific results show that the proposed method achieves better performance on English samples, with both ICM-Hard Norm and Macro-F1 outperforming ArcosGPT_1. However, on Spanish samples, particularly in the Source Intention Detection task, its performance is clearly lower than that of ArcosGPT_1. This result suggests that the language-aware weighted ensemble strategy is effective for English memes, while the contextual meanings, sarcasm, and author intentions in Spanish memes remain important directions for future improvement.

Overall, CYUTGiantsshoulders_1 outperforms ArcosGPT_1 on both Task 2.1 and Task 2.2 for English samples, but still lags behind on Spanish samples. The overall ALL results are close to those of ArcosGPT_1, with the ICM-Hard Norm score for Task 2.2 being slightly higher than that of ArcosGPT_1.

7. Discussion

This section provides an in-depth analysis of the asymmetric patterns and errors observed in the experiments.

7.1. Challenges in Fine-grained Intention Classification (Task 2.2) and Minority Classes

The development set and official evaluation results show that, although the methods perform strongly and consistently on Task 2.1, Sexism Identification, the downstream tasks, namely Task 2.2 and Task 2.3, remain highly challenging.

• Performance bottleneck of the JUDGEMENTAL class: In Task 2.2, the model demonstrates a stronger ability to capture DIRECT intentions, whereas its recognition performance for the JUDGEMENTAL class is considerably lower. This is mainly because evaluative or judgmental sexist expressions are often embedded in highly sarcastic or metaphorical contexts. Even textual derivative features generated by a powerful model such as GPT-4o may sometimes fail to fully interpret the implicit bias arising from image–text complementarity.
• Long-tail effects in minority classes: In the auxiliary Task 2.3, the model even obtains an F1-score of 0 for certain highly challenging categories, such as SEXUAL-VIOLENCE and MISOGYNY-NON-SEXUAL-VIOLENCE. This indicates that the model still lacks sufficient recognition ability for minority classes and highly context-dependent content. This issue is mainly attributable to the extreme scarcity of samples for these categories in the dataset, the higher degree of annotator disagreement, and the fact

that such content is often expressed through highly implicit or indirect visual and textual symbols.

7.2. Reflections on Complex Multimodal Fusion Mechanisms

Method 4 introduces a cross-attention fusion mechanism, originally aiming to enable textual tokens to actively attend to image patches and thereby achieve more fine-grained cross-modal alignment. However, the results show that this more complex interaction architecture does not outperform the simpler late fusion approach used in Method 2 across all metrics. This provides an important academic implication: in a shared-task setting with limited data size and highly imbalanced label distributions, overly designed attention-based interaction layers may instead amplify noise and even lead to overfitting in boundary decisions for rare labels, such as JUDGEMENTAL. In contrast, the language-aware weighted ensemble strategy adopted in Method 5 demonstrates stronger generalization and stability under uncertain labels by enhancing the semantic specialization of text encoders and combining it with probability averaging across multiple random seeds.

7.3. Reasons for Not Replicating More Resource-intensive Experimental Methods

One limitation of this study is that we did not directly fine-tune large-scale vision-language models, such as Qwen2.5-VL-32B. Previous work, such as GrootWatch at EXIST 2025, has shown that fine-tuned vision-language models can achieve strong performance in sexism detection in memes. However, this approach requires substantial GPU memory and training resources. GrootWatch conducted its experiments using high-memory GPUs such as NVIDIA A40 and A100, whereas our available experimental environment was relatively limited. Due to these computational resource constraints, directly fine-tuning large-scale vision-language models was not feasible under our experimental setting [10].

To address this limitation, this study adopts a more resource-efficient alternative. Instead of fine-tuning a large-scale vision-language model in an end-to-end manner, we use GPT-4o as a multimodal semantic feature extractor to convert meme images into textual features, such as OCR text, contextual descriptions, and analytical reasoning. These features are then processed by Transformer-based text encoders and combined through a language-aware weighted ensemble approach. This design enables us to incorporate multimodal information while keeping the trainable models within a manageable computational scale.

8. Conclusion & Future Work

8.1. Conclusion

In the EXIST 2026 Memes Track, this study presents CYUTGiantsshoulders_1, a multimodal system based on large-model semantic augmentation and language-aware ensembling. The experimental results demonstrate that extending ArcosGPTstyle LLM-based textual reasoning features, combined with RAG-based contextual prompting, can substantially improve the baseline performance of sexism detection in memes. The proposed language-aware weighted ensemble method achieved strong results on the official test set, ranking 11th worldwide in Task 2.1 and 10th worldwide in Task 2.2. Overall, the proposed method achieves performance comparable to ArcosGPT_1 and obtains better results on English samples in both Task 2.1 and Task 2.2; however, there remains considerable room for improvement on Spanish samples.

8.2. Future Work

1. Future research may further introduce larger open-source multimodal large models for parameter-efficient fine-tuning when computational resources permit, such as Qwen2.5VL, LLaVA, InternVL, or other vision-language models with image understanding capabilities. Due to the limitations of available GPU memory and training cost, this study was unable to directly perform full fine-tuning on large-scale multimodal models.
Therefore, the final system adopts GPT-4o-based semantic feature extraction and text encoder ensembling to indirectly leverage image and contextual information. With more sufficient hardware resources, future work may explore parameter-efficient fine-tuning methods such as LoRA or QLoRA, allowing the model to directly learn cross-modal relationships among meme images, overlaid text, and tweet context.
This direction is expected to address the current limitation that fine-grained visual cues may be lost when using GPT-generated textual features or CLIP-ViT visual encoders for indirect fusion.

2. For annotator disagreement and extremely imbalanced minority classes, future work may design more robust loss functions or contrastive learning mechanisms to further overcome the bottlenecks in fine-grained intention classification and sexism category prediction.
Another promising direction is to explore data augmentation strategies based on controlled generation, followed by human or model-assisted filtering and soft-label annotation.
Such strategies may help mitigate the tendency of models to favor majority classes while overlooking minority categories.

Acknowledgments

This study was supported by the National Science and Technology Council under Grant NSTC 115-2221-E-324-014.

References

[1] L. Plaza, J. C. de Albornoz, I. Arcos, M. Aloy-Mayo, P. Rosso, E. García-Arias, D. Spina, Overview of exist 2026: Physiological data for multimodal sexism characterization in social media, in: In Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer-Verlag, Berlin, Heidelberg, 2026. URL: https://doi.org/10.1007/978-3-032-04354-2_16. doi:10.1007/978-3-032-04354-2_16.

[2] I. Árcos, Arcosgpt at exist 2025: Identifying sexism in memes with multimodal deep learning: Fusing text and visual cues, in: Conference and Labs of the Evaluation Forum, 2025. URL: https://api.semanticscholar.org/CorpusID:282248302.

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019). URL: https://arxiv.org/abs/1907.11692. doi:10.48550/arXiv.1907.11692.

[4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[5] J. D. la Rosa y Eduardo G. Ponferrada y Manu Romero y Paulo Villegas y Pablo González de Prado Salas y María Grandury, Bertin: Efficient pre-training of a spanish language model using perplexity sampling, Procesamiento del Lenguaje Natural 68 (2022) 13–23. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403.

[6] A. F. M. de Paula, G. Rizzi, E. Fersini, D. Spina, Ai-upv at exist 2023 – sexism characterization

using large language models under the learning with disagreements regime, 2023. URL: https://arxiv.org/abs/2307.03385. arXiv:2307.03385.
[7] R. Pan, T. Bernal Beltrán, J. A. García Díaz, R. Valencia-García, Umuteam at exist 2025: multimodal transformer architectures and soft-label learning for sexism detection, Working Notes of CLEF (2025).
[8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.
[9] Qwen, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, Z. Qiu, Qwen2.5 technical report, 2025. URL: https://arxiv.org/abs/2412.15115. doi:10.48550/arXiv.2412.15115.
[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, Lecture Notes in Computer Science (2025).
[11] M. Guerrero-García, F. Carrillo García, J. Mata, V. Pachón-Álvarez, I2c-uhu-altair at exist2025: multimodal sexism detection and classification using advanced vision-language models blip2 and qwen, large language models, and learning with disagreement frameworks, Working Notes of CLEF (2025).

A. Appendix

A.1. Prompt design

GPT-4o Prompt

You are an assistant specialized in analyzing social media memes, with expertise in identifying sexism and social bias.
The original tweet text is as follows:
`{tweet_text}`
Please jointly analyze the tweet text and the provided image, and output the following features in JSON format using English key names:
1. `OCR_Text`: Extract all textual content appearing in the image.
2. `Context_Description`: Describe the meme context and identify whether it involves sexist themes. Focus on social and gender-related meanings while avoiding irrelevant details.
3. `Sexism_Identification`: Identify whether the meme involves sexist content.
4. `Author_Intent`: Explain the author's intention, such as humor, criticism, normalization of bias, or direct derogation.
5. `Analytical_Reasoning`: Provide the reasoning behind the analysis, especially the deeper social meaning produced by the combination of image and text.
6. `Implicit_Bias`: Determine whether the combination of image and text involves gender stereotypes, objectification of women, or violent metaphors.
7. `Sentiment`: Identify the overall sentiment, such as sarcasm, anger, humor, or neutrality.
Please ensure that the output is a valid JSON object without Markdown code block markers.