Multi-Task Cross-Modal Reasoning with Physiological-Aware Fusion for Sexism Detection in Memes

EXIST at CLEF 2026

Changxin Sun[1], Yong Lu[2] and Leilei Kong[1,*]

[1]Foshan University, Foshan, 528000, China
[2]Guangzhou College of Technology and Business, School of AI Engineering, 510850, China

Abstract

To address the implicit dissemination of sexist content in Internet memes through humor, irony, and image-text incongruity, we propose MemeSense, a multi-task cross-modal semantic reasoning method for sexism detection in English memes. The proposed method focuses on three key challenges: cross-modal semantic understanding, communicative intent inference, and subjective uncertainty modeling. It integrates multimodal large language model reasoning, structured task constraints, and physiological response signals to jointly model image-text content, communicative intent, and variation in human perception. Results on the English Memes subset of the EXIST 2026 sexism detection shared task show that MemeSense achieves competitive performance. Under the Soft-Soft setting, it ranks second in Task 2.1, with an ICM-Soft Norm score of 0.6234. Under the Hard-Hard setting, it ranks third in both Task 2.1 and Task 2.2, with ICM-Hard Norm scores of 0.7537 and 0.6200, respectively. It also ranks first in Task 2.3, with an ICM-Hard Norm score of 0.5260. The results indicate that cross-modal semantic reasoning is the main source of system performance. Training-set ablations further show that physiological-signal fusion provides a small but consistent calibration gain for Task 2.1, while Tasks 2.2 and 2.3 are generated from the structured multimodal reasoning output without physiological fusion.

Keywords

sexism detection, multi-task cross-modal reasoning, memes, physiological signals

1. Introduction

Sexism generally refers to prejudice and unequal treatment based on gender differences. In both offline and online contexts, women are often subjected to explicit attacks, derogatory ridicule, or implicit insinuations. Although public awareness of gender equality has continued to increase, such biases remain deeply embedded in online interactions and are further amplified by the dissemination mechanisms of social media platforms [1]. Online sexism may undermine women's self-perception, social participation, and access to opportunities, while repeated exposure may lead younger audiences to normalize misogynistic expressions. Among various forms of online content, Internet memes can intensify this effect because discriminatory meanings are often expressed through humor, irony, and image-text incongruity. The core challenge is therefore to interpret implicit sexist meanings rather than simply detect isolated harmful words or visual elements.

Vision-language understanding and multimodal content analysis have therefore become important approaches for detecting, analyzing, and mitigating online sexism. Recent advances in visual instruction tuning have further improved the adaptability of models to image-text tasks [2]. Existing research has progressed from lexicon-based rules, statistical features, and conventional text classifiers to pretrained semantic understanding methods, enabling systems to better identify implicit bias, sarcasm, and context-dependent offensive content. In this paper, we participate in EXIST 2026 at CLEF and explore a method for sexism detection in memes, with particular attention to authorial intent and annotation disagreement.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
18779866976@163.com (C. Sun); luy_2019@126.com (Y. Lu); kongleilei@fosu.edu.cn (L. Kong)
0009-0007-9302-7905 (C. Sun); 0000-0002-4636-3507 (L. Kong)

This study focuses on sexism detection in Internet memes. The task first requires a system to determine whether a given meme contains sexist content targeting women. It then requires the system to identify the communicative intent of the content, namely whether it directly disseminates sexism or instead criticizes, exposes, or ironically presents sexist discourse. In addition, the system must determine the specific type of sexism expressed in the meme, such as gender-role inequality, stereotyping, objectification of women, or implications of sexual violence. The task includes both hard-label and soft-label evaluation settings. The former evaluates final categorical decisions, whereas the latter assesses whether the model can capture disagreements among annotators. Thus, the EXIST 2026 Memes task jointly involves cross-modal semantic judgment, multi-task output consistency, and subjective disagreement modeling [3].

To address these issues, MemeSense is designed around cross-modal semantic reasoning and structured multi-task output. It uses multimodal large language models to capture fine-grained signals such as irony, implicit bias, and authorial stance. To avoid semantic conflicts across subtasks, this paper also incorporates the task label space and output constraints into a unified modeling framework.

However, beyond cross-modal semantic understanding, another key challenge lies in the subjectivity of sexism judgments. Relying on a single predictive model or an absolute hard-label mechanism to determine whether a meme contains sexist content may introduce systematic bias, as such judgments are inherently subjective. Previous studies have shown that annotators' individual backgrounds, including gender, cultural context, and lived experience, can substantially shape their perception of the boundary between humor and offense [4, 5]. Therefore, the common practice of resolving disagreement through majority voting risks obscuring individual-level differences and weakening the representational richness provided by diverse annotator perspectives.

To address this challenge, we adopt the Learning with Disagreements (LWD) paradigm, treating annotator disagreement as a meaningful signal of subjective complexity rather than as noise to be removed [6]. Under this paradigm, the system avoids seeking a single absolute answer and instead uses soft labels to produce probability distributions that better reflect the diversity of human judgments. To further quantify and calibrate subjective uncertainty in disagreement-aware learning, we introduce physiological signals as implicit indicators of human cognitive and affective responses. Related studies on EEG-based emotion recognition also provide useful references for physiological signal analysis [7, 8]. By integrating these implicit physiological representations with cross-modal semantic features, the system can better model and calibrate perceptual-level uncertainty and approximate the diversity of human opinions.

The main contributions of this paper are as follows:

- We formulate sexism detection in memes as a problem involving cross-modal semantic understanding, communicative intent inference, and subjective uncertainty modeling, rather than as a conventional single-label classification task.
- We propose a structured reasoning method based on multimodal large language models, enabling the three subtasks to share image-text semantic analysis while maintaining output consistency.
- We introduce physiological signals, including eye tracking (ET), heart rate (HR), and electroencephalography (EEG), as auxiliary calibration information to support the modeling of human perceptual differences in high-disagreement and borderline samples.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the MemeSense methodology. Section 4 presents the experimental settings and results. Section 5 concludes the paper.

2. Related Work

In recent years, sexism detection in online images has emerged as an important research direction in multimodal harmful content detection. In relation to the proposed method, existing studies can be broadly grouped into three lines of research. The first concerns meme and multimodal harmful content

detection. The second involves multimodal large language models and prompt-based reasoning for complex semantic judgments. The third concerns physiological signal-assisted modeling, which uses viewer responses, such as eye tracking (ET), heart rate (HR), and electroencephalography (EEG), to provide supplementary information for model prediction.

2.1. Meme and Multimodal Harmful Content Detection

Prior work on meme and multimodal harmful content detection shows that relying solely on OCR (Optical Character Recognition) text or isolated visual features may lead to misclassification. Tasks such as MAMI and the Hateful Memes Challenge have shown that harmful image detection requires joint modeling of visual and textual information [9, 10]. Existing methods typically perform classification using visual encoders, text encoders, and cross-modal fusion modules. Among MAMI systems, TIB-VA adopts a multimodal architecture for misogynous meme detection [11], UPB enhances UNITER with image sentiment and graph convolutional modeling [12], and AMS-ADRN performs image-text joint modeling for multi-task misogyny identification [13]. Other systems explore different modeling strategies: Codec uses a multi-transformer framework [14], RubCSG relies on ensemble learning [15], and HateU focuses on multimodal misogyny classification [16]. However, most of these methods directly predict labels and lack explicit reasoning over authorial intent, ironic context, and complex social semantics.

2.2. Multimodal Large Language Models and Prompt-Based Image-Text Reasoning

To address complex semantic phenomena such as irony, metaphor, and communicative intent, recent studies have increasingly leveraged the image-text understanding and natural language reasoning capabilities of multimodal large language models. These models provide a new technical pathway for sexism detection in memes. Unlike conventional supervised classifiers, multimodal large language models can process both image and text inputs and perform semantic analysis, reasoning, and structured output generation through natural language prompts. LLaVA is relevant because it demonstrates the effectiveness of visual instruction tuning for multimodal reasoning [2]. More broadly, CLIP learns transferable visual representations from natural language supervision [17], BLIP-2 connects frozen image encoders with large language models [18], and InstructBLIP further adapts this paradigm through instruction tuning [19]. In misogynistic meme detection, chain-of-thought prompting provides a general reasoning mechanism [20], while M3Hop-CoT applies multi-hop reasoning to implicit and context-dependent meme semantics [21]. However, multimodal large language models may still produce unstable output formats, category drift, and overconfident probability estimates. Inspired by the separation between intermediate reasoning and subsequent action in ReAct-style prompting [22], this paper adopts a simplified two-stage prompting pipeline: one model call produces task-oriented semantic analysis, and a second call converts that analysis into schema-constrained outputs. Unlike the full ReAct setting, the proposed pipeline does not interleave reasoning with external-environment actions. Schema-based constraints are further incorporated to improve the stability of labels, probabilities, and hierarchical relations [23].

2.3. Physiological Signal-Assisted Modeling and Human Response Perception

In addition to image and text content, human viewing responses can provide auxiliary information for subjective content recognition. Eye tracking, heart rate, and electroencephalography signals are widely used in affective computing, human-computer interaction, and cognitive load analysis. Eye-tracking signals can reflect the distribution of visual attention, peripheral physiological signals such as heart rate are often associated with emotional arousal, and electroencephalography signals can capture aspects of cognitive processing and neural response. Related studies have shown that physiological signals provide complementary information at the level of bodily response, which differs from textual or behavioral evidence [7, 8]. For sexism detection in memes, physiological signals cannot replace image-text semantic models, because they do not directly indicate whether content contains sexist meanings. However, they

may reflect participants' responses to offensiveness, conflict, ambiguity, or cognitive load. EXIST 2026 introduces eye-tracking, heart rate, and electroencephalography data into the task of multimodal sexism characterization, and related human-centered multimodal fusion studies suggest that ET, HR, and EEG can serve as auxiliary signals beyond content analysis [24]. Therefore, this paper models physiological signals as an auxiliary probabilistic source and fuses them with multimodal large language model outputs at the decision level to calibrate predictions for borderline and high-disagreement samples.

3. Method

To address the challenges of cross-modal semantic understanding, task-level output consistency, and subjective uncertainty modeling in sexism detection in memes, this paper proposes MemeSense. MemeSense formulates sexism detection in memes as a multi-task cross-modal reasoning process, enabling sexism detection, communicative intent inference, and sexism type classification for each meme sample. It further incorporates physiological signals as auxiliary calibration information for subjective uncertainty, thereby integrating multimodal semantic reasoning, structured task constraints, and physiological signal-assisted fusion for complex semantic judgment in memes.

3.1. Task Definition

Given the i-th meme sample $x_i = (I_i, T_i, S_i)$, where I_i denotes the meme image, T_i denotes the text in the image, and $S_i = \{S_i^{ET}, S_i^{HR}, S_i^{EEG}\}$ denotes the physiological signals of eye tracking, heart rate, and electroencephalography, MemeSense produces structured predictions for three subtasks: sexism detection, communicative intent inference, and sexism type classification. The first determines whether the sample contains sexist content and estimates the corresponding probability; the second distinguishes direct sexist expression from critical, expository, or ironic presentation; and the third identifies the sexism categories involved in the sample. In this paper, the structured prediction results of the three tasks are denoted as

$$y_i = (y_i^{sex}, y_i^{intent}, y_i^{type}). \tag{1}$$

Here, y_i^{sex} denotes the result of sexism detection, taking the value YES or NO; y_i^{intent} denotes the result of communicative intent inference, taking the value DIRECT, JUDGEMENTAL, or NO; and y_i^{type} denotes the result of sexism type classification. Since the type classification task allows one sample to correspond to multiple sexism types, this paper represents its multi-label output in vector form as y_i^{type}.

3.2. Model Architecture

Under the above problem definition, MemeSense performs image-text semantic reasoning using a multimodal large language model, enforces multi-task output consistency through structured constraints, and incorporates physiological signals to calibrate probabilistic judgments in the sexism detection task. Overall, MemeSense consists of three main components: two-stage semantic analysis and structured output generation, schema-constrained hierarchical output, and physiological signal modeling with decision-level fusion. At the implementation level, LangGraph is used to organize the two model calls and validation steps, while Pydantic is used to define state objects and structured output schemas [23]. This design enables image-text semantic analysis, structured prediction, hierarchical consistency checking, and physiological signal fusion to be performed within a unified workflow state. The overall system architecture is shown in Figure 1.

3.2.1. Two-Stage Semantic Analysis and Structured Output Generation

The main difficulty in sexism detection in memes is that discriminatory meanings often emerge from the interplay among image content, in-image text, humorous context, and ironic expression. Directly requiring a multimodal large language model to generate final labels may lead to insufficient analysis of image-text relations, category drift, and inconsistent outputs across subtasks. Therefore, MemeSense

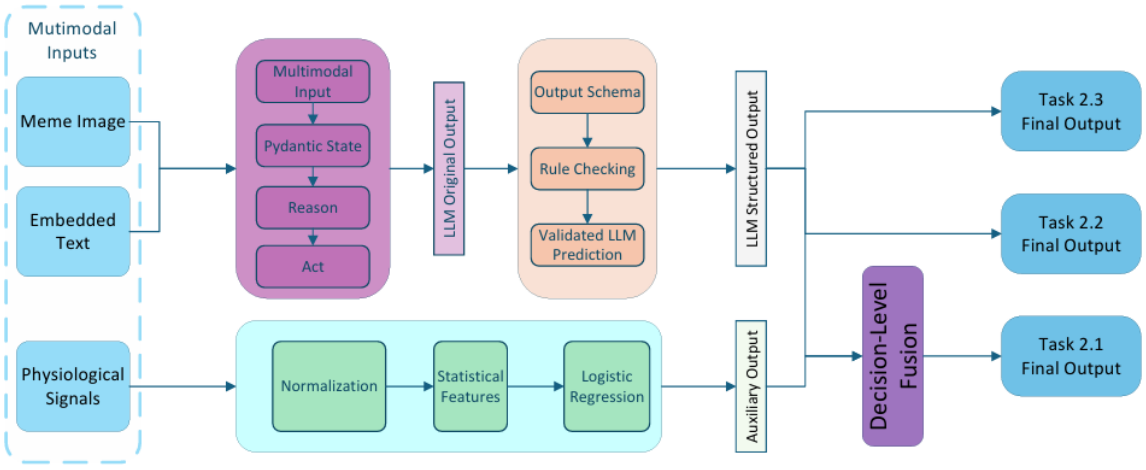

Figure 1: Architecture of MemeSense, including multimodal input processing, two-stage semantic analysis and structured output generation, schema-constrained prediction, physiological signal modeling, and decision-level fusion.

implements a simplified two-stage prompting pipeline using two prompt templates, as shown in Figures 2 and 3. This design is inspired by the conceptual separation of reasoning and acting in ReAct [22], but it is adapted here as two consecutive model calls rather than as an interactive process that alternates reasoning with external actions. Specifically, the semantic analysis stage corresponds to the first call to the multimodal large language model, taking the meme image I_i, the in-image text T_i, and the Reason Prompt as inputs. This prompt requires the model to analyze image content, textual meaning, image-text relations, sexism-related cues, and the distinction between direct expression and critical or ironic presentation, and to output a task-oriented semantic analysis result r_i. The structured output stage corresponds to a second call to the large language model, with r_i obtained from the previous stage, label definitions, and the Act Prompt as inputs. This prompt requires the model to generate structured JSON output according to predefined fields, including the sexism detection label and probability, the communicative intent inference result, the sexism type classification result, and the overall confidence score. Subsequently, the Pydantic schema validates the fields, label set, and probability ranges in the JSON output, ensuring that the output can be used by subsequent modules for hierarchical consistency checking, probability calibration, and decision-level fusion. In implementation, LangGraph organizes the two stages as consecutive nodes and stores r_i and the structured prediction result y_i in the workflow state, thereby completing the state transition from image-text semantic analysis to structured task output. Figures 2 and 3 show compact prompt sketches for readability; the complete prompts, label definitions, JSON schema, retry and repair rules, and a representative input-output example are provided in the supplementary prompt specification below.

3.2.2. Schema Constraints and Hierarchical Output

Semantic conflicts can easily occur in multi-task prediction. For example, a sample may be classified as non-sexist while still being assigned to a specific sexism type. To avoid such problems, MemeSense defines the label space, probability ranges, and cross-task relations as a structured output schema. Here, the schema refers to the output protocol designed in this paper for the three subtasks. Pydantic is a data validation tool based on Python type annotations. In this paper, it is used to parse and validate model output fields, label values, and probability ranges, rather than serving as part of the reasoning model. Specifically, the schema constrains the valid values of y_i^{sex}, y_i^{intent}, and y_i^{type}. In the post-processing stage, if $y_i^{sex} = \text{NO}$, the system sets both the communicative intent and sexism type to NO, assigns the probability of the non-sexist category to 1, and assigns the probabilities of the remaining categories to 0. If $y_i^{sex} = \text{YES}$, the system retains the model predictions and checks whether the label set and probability

Reason Prompt

Input:

Meme image I_i, OCR text T_i, task description

Instruction:

1. Understand the visual content, including people, objects, actions, facial expressions, and scene context.

2. Explain the OCR text and analyze how the text interacts with the image.

...

Output:

Task-oriented semantic analysis r_i

Figure 2: Reason Prompt template for analyzing image content, in-image text, image-text relations, sexism-related cues, and authorial intent.

Act Prompt

Input:

Semantic analysis r_i, label definitions, output schema

Instruction:

Generate a structured JSON output with the following fields:

- reasoning: brief summary of the analysis

- is_sexist: YES or NO

- intention: DIRECT, JUDGEMENTAL, or null

- categories: selected sexism categories

...

Output:

Structured prediction $\hat{y}_i$

Figure 3: Act Prompt template for generating structured task outputs, including labels, probabilities, confidence scores, and other structured fields.

ranges are valid. Through this constraint mechanism, the model-generated results are transformed into structured predictions that conform to the task definition, are parseable, and remain consistent across tasks.

Under the Soft-Soft setting, the probability outputs are generated from the structured prediction stage rather than from an explicit annotator-distribution calibration model. For Task 2.2, the model produces confidence scores for DIRECT, JUDGEMENTAL, and NO; these scores are clamped to $[0, 1]$ and renormalized so that the three probabilities sum to 1. If Task 2.1 predicts NO, the Task 2.2 distribution is overridden as NO=1. For Task 2.3, the model outputs independent marginal probabilities for NO and the five sexism categories. Because Task 2.3 is multi-label, these probabilities are not normalized to sum to 1; the schema only enforces valid ranges and hierarchical consistency, setting NO=1 and all positive

Table 1
Dataset overview.

Item	Description
Dataset	EXIST 2026 Memes
Subset used	English subset
Training set size	2,005 samples
Test set size	513 samples
Input content	Meme images and text in images
Physiological signals	ET, HR, EEG
Subtasks	Task 2.1, Task 2.2, Task 2.3
Main analysis settings	Task 2.1 Soft, Task 2.1 Hard, Task 2.2 Hard, Task 2.3 Hard
Training set labels	Gold labels are provided for ablation analysis, parameter selection, and validation of fusion strategies
Test set labels	Gold labels are not provided

category probabilities to 0 when Task 2.1 is NO.

3.2.3. Physiological Signal Modeling and Decision-Level Fusion

To combine image-text semantic judgments with participants' viewing responses, this paper adopts linear decision-level fusion to obtain the final probability of sexism detection:

$$P_{\text{final},i} = \alpha P_{\text{LLM},i} + (1 - \alpha) P_{\text{sensor},i}. \tag{2}$$

Here, $P_{\text{LLM},i}$ denotes the sexism probability predicted by the multimodal large language model based on the image I_i and text T_i, $P_{\text{sensor},i}$ denotes the probability predicted by the sensor model based on the physiological signal features z_i, and α controls the relative contribution of the two sources. For Task 2.1, $P_{\text{final},i}$ is used as the soft-label probability. Under the hard-label setting, it is converted into a discrete label using a threshold τ: if $P_{\text{final},i} \geq \tau$, the output is YES; otherwise, the output is NO. To estimate $P_{\text{sensor},i}$, the system standardizes the ET, HR, and EEG signals by participant and feature dimension, then extracts sample-level statistics and missing-value indicators to construct z_i. The sensor probability is estimated using logistic regression with L2 regularization.

It should be emphasized that physiological-signal fusion is used only for probability calibration in Task 2.1 sexism identification. Task 2.2 authorial intent detection and Task 2.3 sexism type classification are generated directly from multimodal semantic reasoning and structured post-processing, without additional recalibration using physiological signals.

4. Experiments

The experiments reported below use the official dataset split and the ICM-based evaluation protocol for hierarchical classification [25].

4.1. Dataset

This paper evaluates MemeSense on the English subset of the EXIST 2026 Memes dataset. The dataset is designed for sexism detection and characterization in social media memes and includes meme images, in-image text, task labels, and physiological signals, including eye tracking, heart rate, and electroencephalography [3, 26]. The training set contains 2,005 samples, while the test set contains 513 samples. Gold labels are provided for the training set and are used for parameter selection, ablation analysis, and validation of fusion strategies. In contrast, gold labels for the test set are not publicly released and are used only for official evaluation. An overview of the dataset is provided in Table 1.

4.2. Data Preprocessing

To ensure that images, text, and physiological signals can be processed by subsequent modules, this paper applies unified preprocessing to all modalities during the experimental stage. For each meme sample, the system formats the original image, in-image text, and task prompt as inputs to the multimodal large language model, and maps the labels of Task 2.1, Task 2.2, and Task 2.3 to their corresponding label spaces. For ET, HR, and EEG signals, the system standardizes the data by participant and feature dimension, and extracts sample-level statistical features and missing-value indicators to construct the sensor feature vector. After preprocessing, images and text are used for multimodal semantic reasoning, while physiological signal features are used for sensor-side probability modeling and subsequent decision-level fusion.

4.3. Evaluation Metrics

This paper follows the official evaluation settings of EXIST 2026, which mainly include the Hard-Hard and Soft-Soft settings. In the Hard-Hard setting, both system outputs and ground-truth annotations are discrete category labels. In the Soft-Soft setting, both system outputs and ground-truth annotations are category probability distributions, which are used to capture annotator disagreement. The evaluation metrics used in this paper are as follows:

- **ICM-Hard**: The primary evaluation metric under the Hard-Hard setting. It measures the information consistency between the system's discrete predictions and the gold hard labels. Compared with standard classification accuracy, ICM-Hard places greater emphasis on differences in label structure and information content.
- **ICM-Hard Norm**: The normalized version of ICM-Hard, which maps hard-label results from different systems onto a more comparable scale. This paper mainly uses this metric to analyze model performance under the Hard-Hard setting.
- **F1-score**: An auxiliary metric under the hard-label setting. It jointly considers precision and recall and reflects the model's classification performance in discrete category prediction. Compared with accuracy, F1-score is more informative when the class distribution is imbalanced.
- **ICM-Soft**: The primary evaluation metric under the Soft-Soft setting. It measures the information consistency between the predicted probability distribution and the gold soft-label distribution. This metric reflects whether the model can effectively characterize sample-level uncertainty and annotator disagreement.
- **ICM-Soft Norm**: The normalized version of ICM-Soft, facilitating comparison across different methods. This paper mainly uses this metric for model analysis and strategy selection under the Soft-Soft setting.
- **Cross Entropy**: An auxiliary metric under the soft-label setting. It measures the discrepancy between the predicted probability distribution and the gold soft-label distribution. A lower value indicates that the predicted distribution is closer to the ground-truth annotation distribution.

Overall, this paper uses ICM-Hard and ICM-Soft as the main evaluation metrics, with F1-score and Cross Entropy as supplementary metrics, to assess model performance in both discrete category prediction and probability distribution estimation [25].

4.4. Baselines

This paper uses the official non-informative baselines as experimental references. These baselines are intended to verify the output format and evaluation procedure, and to provide basic comparison points for system performance; they do not represent competitive methods for the task. The official baselines include two variants: Majority class and Minority class. The former assigns all samples to the majority class, whereas the latter assigns all samples to the minority class. Since neither baseline uses input information such as images, text, or physiological signals, both are non-informative systems. Under the

Hard-Hard setting, the Majority class and Minority class baselines directly output the corresponding hard labels. Under the Soft-Soft setting, they follow the same majority or minority class rule, assigning a probability of 1 to the selected class and 0 to all other classes. Thus, although these baselines are compatible with both hard and soft output formats, they are essentially deterministic class-prior predictors. This paper uses them as official references to evaluate the effectiveness of MemeSense relative to simple majority/minority rules.

4.5. Experimental Settings

4.5.1. Training Parameters

This paper performs model adaptation through prompt-based task formulation, schema-based output constraints, and post-processing calibration. The multimodal reasoning module uses a closed-source multimodal large language model, accessed through an OpenAI-compatible interface with the API configuration identifier `gpt-5.4`, to perform image-text semantic analysis and structured prediction. The inference temperature is set to 0 to reduce generation randomness and improve reproducibility. For the physiological signal model, logistic regression with L2 regularization is used to estimate the sensor-side probability $P_{\text{sensor},i}$. Under the soft setting, the sensor model is trained with 5-fold cross-validation, using candidate L2 regularization parameters of 0.001, 0.01, 0.1, and 1.0; $L_2 = 0.1$ is selected. Under the hard setting, 3-fold cross-validation is used, with candidate L2 regularization parameters of 0.01 and 0.1; $L_2 = 0.01$ is selected. The fusion stage adopts linear decision-level fusion. Under the soft setting, ICM-Soft Norm is used as the selection criterion, resulting in $\alpha = 0.85$. Under the hard setting, both the fusion weight and classification threshold are selected, resulting in $\alpha = 0.80$ and a threshold of 0.55. All parameters are determined through ablation experiments on the training set, ensuring that the fusion strategy is selected based on reproducible validation results rather than empirical adjustment using test-stage predictions.

4.5.2. Implementation Details

The experiments are conducted in a local environment equipped with a single NVIDIA GeForce RTX 5060 GPU. The multimodal reasoning module performs image-text understanding and structured prediction through a compatible API interface. All remaining experimental procedures, including ET/HR/EEG data preprocessing, feature aggregation, sensor model training, cross-validation, fusion parameter selection, and evaluation metric computation, are executed locally. The sensor-side model is implemented using logistic regression with L2 regularization, with statistical features of physiological signals and missing-value indicators as input. During the experiments, the system records structured prediction results, sensor-side probabilities, fused probabilities, and final outputs for subsequent ablation analysis, probability calibration, and error analysis.

4.6. Results

Before reporting the official test results, we first examine the contribution of physiological-signal fusion on the training set. The ablation results for Task 2.1 are shown in Table 2.

Table 2
Training-set ablation results for Task 2.1.

Setting	Hard Acc.	ICM-Hard Norm	ICM-Soft Norm
MLLM only	0.8122	0.7012	0.5791
Physiological only	0.6209	0.4041	0.2586
MLLM + physiological fusion	0.8169	0.7088	0.5813

In Table 2, MLLM only denotes image-text semantic reasoning without physiological fusion, Physiological only uses ET, HR, and EEG features, and MLLM + physiological fusion denotes the complete

Table 3
Results for Task 2.1 sexism identification in English memes under the Hard-Hard setting.

System	Rank	ICM-Hard	ICM-Hard Norm	F1 YES
EXIST2025-test_gold	0	0.9848	1.0000	1.0000
EXIST2025-test_majority-class	202	-0.4076	0.2931	0.6880
EXIST2025-test_minority-class	213	-0.6381	0.1761	0.0000
MRBLACK_1	3	0.4997	0.7537	0.8311

Table 4
Results for Task 2.1 sexism identification in English memes under the Soft-Soft setting.

System	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
EXIST2025-test_gold	0	3.0794	1.0000	0.5528
EXIST2025-test_majority-class	140	-2.2236	0.1390	4.4798
EXIST2025-test_minority-class	141	-3.1235	0.0000	5.4888
MRBLACK_1	2	0.7601	0.6234	0.8393

Table 5
Results for Task 2.2 authorial intent detection in English memes under the Hard-Hard setting.

System	Rank	ICM-Hard	ICM-Hard Norm	F1
EXIST2025-test_gold	0	1.4409	1.0000	1.0000
EXIST2025-test_majority-class	168	-1.0385	0.1396	0.1877
EXIST2025-test_minority-class	183	-2.0410	0.0000	0.0719
MRBLACK_1	3	0.3459	0.6200	0.5941

Table 6
Results for Task 2.2 authorial intent detection in English memes under the Soft-Soft setting.

System	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
EXIST2025-test_gold	0	4.5834	1.0000	0.9282
EXIST2025-test_majority-class	111	-4.6049	0.0000	5.4792
EXIST2025-test_minority-class	114	-18.1227	0.0000	7.8807
MRBLACK_1	40	-1.4848	0.3380	3.1077

Task 2.1 fusion system. The results show that multimodal semantic reasoning is the dominant source of performance, while physiological fusion provides a small but consistent improvement for Task 2.1. The official test results are then reported in Tables 3–8; this paper primarily focuses on the normalized ICM metrics, with F1-score and Cross Entropy used as supplementary indicators.

The system names beginning with `EXIST2025-test`, such as `EXIST2025-test_gold`, `EXIST2025-test_majority-class`, and `EXIST2025-test_minority-class`, are the original identifiers displayed by the official evaluation platform. They are preserved in the tables for consistency with the official output and do not indicate that the experiments were conducted on the EXIST 2025 task.

For Tasks 2.2 and 2.3, the gap between Hard-Hard and Soft-Soft evaluation is evident. The normalized score drops from 0.6200 to 0.3380 for Task 2.2, and from 0.5260 to 0.2099 for Task 2.3. This suggests that MemeSense is more reliable at selecting the most plausible label than at recovering the full annotator distribution. The gap is larger for Task 2.3 because it combines hierarchical and multi-label decisions, where overlapping sexism categories make probability calibration more difficult.

The official evaluation results show that MemeSense achieves competitive overall performance on the English subset. Tables 3–8 report the system results under the Hard-Hard and Soft-Soft settings for Task 2.1, Task 2.2, and Task 2.3. Overall, MemeSense outperforms the two non-informative baselines, Majority and Minority, in multiple settings, with particularly stable performance under Hard-Hard

Table 7
Results for Task 2.3 sexism type classification in English memes under the Hard-Hard setting.

System	Rank	ICM-Hard	ICM-Hard Norm	F1
EXIST2025-test_gold	0	2.3532	1.0000	1.0000
EXIST2025-test_majority-class	144	-2.0015	0.0747	0.0938
EXIST2025-test_minority-class	175	-3.3506	0.0000	0.2818
MRBLACK_1	1	0.1225	0.5260	0.5722

Table 8
Results for Task 2.3 sexism type classification in English memes under the Soft-Soft setting.

System	Rank	ICM-Soft	ICM-Soft Norm
EXIST2025-test_gold	0	9.2546	1.0000
EXIST2025-test_majority-class	76	-9.2546	0.0000
EXIST2025-test_minority-class	115	-53.0762	0.0000
MRBLACK_1	25	-5.3698	0.2099

evaluation. In Task 2.1 sexism detection, MemeSense ranks third under the Hard-Hard setting and second under the Soft-Soft setting, indicating that the system performs well in YES/NO classification and can also produce probability distributions that partially reflect annotator disagreement. For more fine-grained tasks, MemeSense also shows competitive performance. It ranks third under the Task 2.2 Hard-Hard setting, suggesting that it can distinguish authorial intent effectively. It ranks first under the Task 2.3 Hard-Hard setting, indicating that the two-stage semantic analysis and structured output pipeline, together with schema-based hierarchical constraints, contributes to more consistent sexism type prediction. Overall, MemeSense performs more consistently under hard-label settings, while the Task 2.1 soft-label results indicate its ability to capture subjective uncertainty in sexism detection to some extent.

5. Conclusion

This paper investigates the three subtasks of the EXIST 2026 Memes shared task: sexism detection, authorial intent detection, and sexism type classification. MemeSense achieves competitive performance on English meme data, ranking among the top three in multiple settings and first in the Task 2.3 Hard-Hard setting. These results show that cross-modal semantic reasoning is the main driver of system performance, especially for implicit and fine-grained meme interpretation. The pipeline, combining two-stage semantic analysis with schema-constrained output, also helps maintain consistency across sexism detection, intent inference, and type classification.

Physiological signals cannot independently replace the semantic model. In Task 2.1, however, they can serve as auxiliary calibration information and provide a lightweight complement for some borderline samples. Their contribution remains limited, so future work should develop more systematic physiological feature modeling and validate its effectiveness across tasks.

Acknowledgments

This work is supported by the National Social Science Foundation of China (No. 22BTQ101).

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5.4 for grammar and spelling checks. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for

the content of the publication.

References

[1] H. Kirk, W. Yin, B. Vidgen, P. Röttger, SemEval-2023 task 10: Explainable detection of online sexism, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2193–2210. URL: https://aclanthology.org/2023.semeval-1.305/. doi:10.18653/v1/2023.semeval-1.305.

[2] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, in: Advances in Neural Information Processing Systems, 2023. URL: https://openreview.net/forum?id=w0H2xGHlkw.

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

[4] A. M. Davani, M. Díaz, V. Prabhakaran, Dealing with disagreements: Looking beyond the majority vote in subjective annotations, Transactions of the Association for Computational Linguistics 10 (2022) 92–110. URL: https://doi.org/10.1162/tacl_a_00449. doi:10.1162/tacl_a_00449.

[5] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, Journal of Artificial Intelligence Research 72 (2021) 1385–1470. URL: https://doi.org/10.1613/jair.1.12752. doi:10.1613/jair.1.12752.

[6] E. Leonardelli, G. Abercrombie, D. Almanea, V. Basile, T. Fornaciari, B. Plank, V. Rieser, A. Uma, M. Poesio, SemEval-2023 task 11: Learning with disagreements (LeWiDi), in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2304–2318. URL: https://aclanthology.org/2023.semeval-1.314/. doi:10.18653/v1/2023.semeval-1.314.

[7] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, S. H. A. Chen, NeuroKit2: A python toolbox for neurophysiological signal processing, Behavior Research Methods 53 (2021) 1689–1696. URL: https://doi.org/10.3758/s13428-020-01516-y. doi:10.3758/s13428-020-01516-y.

[8] L. Stappen, A. Baird, L. Christ, L. Schumann, B. Sertolli, E.-M. Messner, E. Cambria, G. Zhao, B. W. Schuller, The MuSe 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress, 2021. URL: https://arxiv.org/abs/2104.07123. arXiv:2104.07123.

[9] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, SemEval-2022 task 5: Multimedia automatic misogyny identification, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 533–549. URL: https://aclanthology.org/2022.semeval-1.74/. doi:10.18653/v1/2022.semeval-1.74.

[10] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 2611–2624. URL: https://proceedings.neurips.cc/paper/2020/hash/1b84c4cee2b8b3d823b30e2d604b1878-Abstract.html.

[11] S. Hakimov, G. S. Cheema, R. Ewerth, TIB-VA at SemEval-2022 task 5: A multimodal architecture for the detection and classification of misogynous memes, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 756–760. URL: https://aclanthology.org/2022.semeval-1.105/. doi:10.18653/v1/2022.semeval-1.105.

[12] A. Paraschiv, M. Dascalu, D.-C. Cercel, UPB at SemEval-2022 task 5: Enhancing UNITER with image sentiment and graph convolutional networks for multimedia automatic misogyny identification, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 618–625. URL: https://aclanthology.org/2022.semeval-1.85/. doi:10.18653/v1/2022.semeval-1.85.

[13] D. Li, M. Yi, Y. He, AMS_ADRN at SemEval-2022 task 5: A suitable image-text multimodal joint modeling method for multi-task misogyny identification, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 711–717. URL: https://aclanthology.org/2022.semeval-1.97/. doi:10.18653/v1/2022.semeval-1.97.

[14] A. Mahran, C. Alessandro Borella, K. Perifanos, Codec at SemEval-2022 task 5: Multi-modal multi-transformer misogynous meme classification framework, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 679–688. URL: https://aclanthology.org/2022.semeval-1.93/. doi:10.18653/v1/2022.semeval-1.93.

[15] W. Yu, B. Boenninghoff, J. Röhrig, D. Kolossa, RubCSG at SemEval-2022 task 5: Ensemble learning for identifying misogynous MEMEs, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 626–635. URL: https://aclanthology.org/2022.semeval-1.86/. doi:10.18653/v1/2022.semeval-1.86.

[16] A. Arango, J. Perez-Martin, A. Labrada, HateU at SemEval-2022 task 5: Multimedia automatic misogyny identification, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 581–584. URL: https://aclanthology.org/2022.semeval-1.80/. doi:10.18653/v1/2022.semeval-1.80.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.

[18] J. Li, D. Li, S. Savarese, S. Hoi, BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), Proceedings of the 40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 19730–19742. URL: https://proceedings.mlr.press/v202/li23q.html.

[19] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, S. Hoi, InstructBLIP: Towards general-purpose vision-language models with instruction tuning, 2023. URL: https://arxiv.org/abs/2305.06500. arXiv:2305.06500.

[20] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Advances in Neural Information Processing Systems, volume 35, 2022, pp. 24824–24837. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

[21] G. Kumari, K. Jain, A. Ekbal, M3Hop-CoT: Misogynous meme identification with multimodal multi-hop chain-of-thought, 2024. URL: https://arxiv.org/abs/2410.09220. arXiv:2410.09220.

[22] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, ReAct: Synergizing reasoning and acting in language models, in: Proceedings of the International Conference on Learning Representations, 2023. URL: https://openreview.net/forum?id=WE_vluYUL-X.

[23] M. X. Liu, F. Liu, A. J. Fiannaca, T. Koo, L. Dixon, M. Terry, C. J. Cai, We need structured output: Towards user-centered constraints on large language model output, 2024. URL: https://arxiv.org/abs/2404.07362. arXiv:2404.07362.
[24] I. Arcos, P. Rosso, E. Gomis-Vicent, Human-centered multimodal fusion for sexism detection in memes with eye-tracking, heart rate, and EEG signals, 2026. URL: https://arxiv.org/abs/2602.23862. arXiv:2602.23862.
[25] E. Amigó, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5809–5819. URL: https://aclanthology.org/2022.acl-long.399/. doi:10.18653/v1/2022.acl-long.399.
[26] L. Plaza, J. Carrillo-de-Albornoz, I. Arcos, M. Aloy-Mayo, P. Rosso, E. García-Arias, D. Spina, Overview of EXIST 2026: Physiological data for multimodal sexism characterization in social media (Extended Overview), in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.

A. Prompt, Output, and Submission Specification

This appendix provides the complete prompting, validation, and submission-format specification used by MemeSense. The prompt sketches in Figures 2 and 3 are abbreviated for readability; the full prompts and post-processing rules are given here to support reproducibility.

The JSON object produced by the Act Prompt is an internal intermediate representation. It is used for schema validation, hierarchical consistency checking, and probability calibration. Before evaluation, this internal representation is converted into the official PyEvALL submission format for each subtask and output type.

A.1. Label Definitions

- **Task 2.1, sexism identification**: `YES` indicates that the meme contains sexist expressions or behaviors, including direct sexism, a description of a sexist situation, or criticism of sexist behavior. `NO` indicates that no sexist expression or behavior is identified.
- **Task 2.2, source intention**: `DIRECT` indicates that the meme itself expresses or encourages sexism. `JUDGEMENTAL` indicates that the meme presents sexist situations or behaviors in order to criticize, expose, or condemn them. `NO` is used only when Task 2.1 is `NO`.
- **Task 2.3, sexism categorization**: the official output label set is `NO`, `IDEOLOGICAL-INEQUALITY`, `STEREOTYPING-DOMINANCE`, `OBJECTIFICATION`, `SEXUAL-VIOLENCE`, and `MISOGYNY-NON-SEXUAL-VIOLENCE`. The five positive categories are multi-label; `NO` is used when no sexism category is predicted.

A.2. Full Reason Prompt

```
System:
You are an expert annotator for the EXIST 2026 English meme sexism characterization task.
Your goal is to analyze the meme before final labels are assigned.
Do not output final labels in this stage. Do not output JSON.

User:
Input:
- Meme image: <image>
- OCR text: <ocr_text>
- Task context: The sample will later be classified for sexism identification, source intention,
    and sexism categorization.
```

```
Instructions:
1. Describe the visible people, objects, actions, setting, facial expressions, gestures, and
     relevant visual symbols.
2. Explain the OCR text literally and pragmatically.
3. Analyze how the image and text interact. State whether the text changes the meaning of the
     image or whether the image changes the meaning of the text.
4. Identify possible sexism-related cues, including gender-role assumptions, stereotypes,
     objectification, sexual harassment or violence, misogyny, humiliation, and claims about
     inequality.
5. Determine whether the meme appears to directly express sexism, criticize or expose sexism, or
      contain no sexist content.
6. Discuss ambiguity, irony, humor, and context. Mention if multiple interpretations are
     plausible.
7. Do not invent external context that is not visible in the image or OCR text.
8. Produce a concise task-oriented semantic analysis for the subsequent structured prediction
     stage.

Output format:
VISUAL_CONTENT:
TEXT_MEANING:
IMAGE_TEXT_RELATION:
SEXISM_CUES:
AUTHORIAL_INTENT_ANALYSIS:
AMBIGUITY_AND_UNCERTAINTY:
FINAL_ANALYSIS_SUMMARY:
```

A.3. Full Act Prompt

```
System:
Convert a semantic analysis of an English meme into a valid JSON object for the EXIST 2026 Memes
      task.
Return JSON only. Do not include Markdown, explanations outside JSON, or code fences.

User:
Input:
- Semantic analysis r_i: <analysis_from_first_call>
- OCR text: <ocr_text>
- Label definitions:
  Task 2.1: YES, NO.
  Task 2.2: DIRECT, JUDGEMENTAL, NO. Use NO only if Task 2.1 is NO.
  Task 2.3: NO, IDEOLOGICAL-INEQUALITY, STEREOTYPING-DOMINANCE, OBJECTIFICATION, SEXUAL-VIOLENCE
       , MISOGYNY-NON-SEXUAL-VIOLENCE.

Probability instructions:
1. For Task 2.1, output a probability distribution over YES and NO. The two probabilities must
     be in [0, 1] and sum to 1.
2. For Task 2.2, output a probability distribution over DIRECT, JUDGEMENTAL, and NO. The three
     probabilities must be in [0, 1] and sum to 1.
3. For Task 2.3, output independent marginal probabilities for all six official labels,
     including NO, because the task is multi-label. Each value must be in [0, 1]; these values do
      not need to sum to 1.
4. If Task 2.1 is NO, set Task 2.2 label to NO, set Task 2.2 probability for NO to 1, set Task
     2.3 labels to ["NO"], set Task 2.3 probability for NO to 1, and set all positive Task 2.3
     category probabilities to 0.
5. If Task 2.1 is YES, select DIRECT or JUDGEMENTAL for Task 2.2 and select one or more positive
      Task 2.3 categories when supported by the analysis.
```

```
6. Use calibrated confidence values. Avoid 0 or 1 unless the evidence is explicit and
    unambiguous.

Return exactly this JSON structure:
{
  "reasoning": "brief evidence-based justification",
  "task_2_1": {
    "label": "YES|NO",
    "probabilities": {"YES": 0.0, "NO": 1.0}
  },
  "task_2_2": {
    "label": "DIRECT|JUDGEMENTAL|NO",
    "probabilities": {"DIRECT": 0.0, "JUDGEMENTAL": 0.0, "NO": 1.0}
  },
  "task_2_3": {
    "labels": ["NO"],
    "probabilities": {
      "NO": 1.0,
      "IDEOLOGICAL-INEQUALITY": 0.0,
      "STEREOTYPING-DOMINANCE": 0.0,
      "OBJECTIFICATION": 0.0,
      "SEXUAL-VIOLENCE": 0.0,
      "MISOGYNY-NON-SEXUAL-VIOLENCE": 0.0
    }
  },
  "confidence": 0.0
}
```

A.4. JSON Schema and Validation

```
{
  "required": ["reasoning", "task_2_1", "task_2_2", "task_2_3", "confidence"],
  "reasoning": string(minLength=1),
  "task_2_1": {
    "required": ["label", "probabilities"],
    "label": enum("YES", "NO"),
    "probabilities": {
      "required": ["YES", "NO"],
      "YES": number(min=0, max=1),
      "NO": number(min=0, max=1)
    }
  },
  "task_2_2": {
    "required": ["label", "probabilities"],
    "label": enum("DIRECT", "JUDGEMENTAL", "NO"),
    "probabilities": {
      "required": ["DIRECT", "JUDGEMENTAL", "NO"],
      "DIRECT": number(min=0, max=1),
      "JUDGEMENTAL": number(min=0, max=1),
      "NO": number(min=0, max=1)
    }
  },
  "task_2_3": {
    "required": ["labels", "probabilities"],
    "labels": unique array of enum(
      "NO",
      "IDEOLOGICAL-INEQUALITY",
      "STEREOTYPING-DOMINANCE",
```

```
      "OBJECTIFICATION",
      "SEXUAL-VIOLENCE",
      "MISOGYNY-NON-SEXUAL-VIOLENCE"
    ),
    "probabilities": {
      "required": [
        "NO",
        "IDEOLOGICAL-INEQUALITY",
        "STEREOTYPING-DOMINANCE",
        "OBJECTIFICATION",
        "SEXUAL-VIOLENCE",
        "MISOGYNY-NON-SEXUAL-VIOLENCE"
      ],
      "NO": number(min=0, max=1),
      "IDEOLOGICAL-INEQUALITY": number(min=0, max=1),
      "STEREOTYPING-DOMINANCE": number(min=0, max=1),
      "OBJECTIFICATION": number(min=0, max=1),
      "SEXUAL-VIOLENCE": number(min=0, max=1),
      "MISOGYNY-NON-SEXUAL-VIOLENCE": number(min=0, max=1)
    }
  },
  "confidence": number(min=0, max=1)
}
```

Pydantic validation checks required fields, allowed labels, numeric ranges, Task 2.1 and Task 2.2 normalization, and hierarchical consistency. If `task_2_1.label` is `NO`, Task 2.2 and Task 2.3 are forced to `NO`, with `NO=1` and all positive Task 2.3 probabilities set to 0. If `task_2_1.label` is `YES`, Task 2.2 must be `DIRECT` or `JUDGEMENTAL`. Task 2.3 hard labels are selected from positive categories with marginal probability at least 0.5; if none is selected, the hard label is `NO`. Task 2.3 soft output preserves all six marginal probabilities and is not normalized to sum to 1.

A.5. Official Submission Mapping

The internal object above is not submitted directly. For official evaluation, MemeSense exports one PyEvALL JSON file per subtask and output type. Each submitted item contains `id`, `test_case`, and `value`; following the EXIST 2026 lab guidelines, `test_case` is kept as `EXIST2025` for compatibility with the evaluation platform. For Tasks 2.1 and 2.2, `value` is either a hard label or a probability dictionary over the official labels. For Task 2.3, `value` is a list of one or more hard labels under the hard setting, and a probability dictionary over the six official labels under the soft setting.

A.6. Retry, Repair, and Failure Handling

```
1. Parse the model response and extract the first complete JSON object.
2. Validate the object using the Pydantic schema.
3. If JSON parsing or validation fails, call the structured-output model again with:
   - the original semantic analysis,
   - the invalid output,
   - the Pydantic error message,
   - an instruction to return corrected JSON only.
4. Repeat step 3 for at most two retries.
5. If the output is still invalid, apply deterministic repair:
   - normalize label spelling and capitalization;
   - clamp all probabilities to [0, 1];
   - renormalize Task 2.1 and Task 2.2 probabilities to sum to 1;
   - derive each hard label from the largest valid probability;
   - enforce the hierarchy: if Task 2.1 is NO, set Task 2.2 to NO and set Task 2.3 to ["NO"];
```

```
   - for Task 2.3, keep positive categories with probability >= 0.5; if none are retained, use
       ["NO"].
6. If deterministic repair still fails, mark the sample as invalid and use a conservative
    fallback:
   Task 2.1 = NO, Task 2.2 = NO, Task 2.3 = ["NO"], confidence = 0.0.
```

A.7. Representative Input-Output Example

```
Input example:
OCR text: "Women belong in the kitchen."
Image summary: A man laughs while pointing toward a woman near a kitchen counter.

Semantic analysis excerpt:
The text assigns women to a domestic role and the image reinforces, rather than criticizes, the
    stereotype. The likely intention is direct sexist expression.

Validated structured output:
{
  "reasoning": "Direct gender-role stereotype without signs of criticism.",
  "task_2_1": {"label": "YES", "probabilities": {"YES": 0.91, "NO": 0.09}},
  "task_2_2": {"label": "DIRECT",
    "probabilities": {"DIRECT": 0.86, "JUDGEMENTAL": 0.10, "NO": 0.04}},
  "task_2_3": {
    "labels": ["STEREOTYPING-DOMINANCE", "IDEOLOGICAL-INEQUALITY"],
    "probabilities": {
      "NO": 0.05, "IDEOLOGICAL-INEQUALITY": 0.55,
      "STEREOTYPING-DOMINANCE": 0.88, "OBJECTIFICATION": 0.08,
      "SEXUAL-VIOLENCE": 0.02, "MISOGYNY-NON-SEXUAL-VIOLENCE": 0.12
    }
  },
  "confidence": 0.86
}
```