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

Notebook for the EXIST Lab at CLEF 2025

Manuel Guerrero-García*, Fernando Carrillo-García, Jacinto Mata-Vázquez and Victoria Pachón-Álvarez

I2C Research Group, University of Huelva, Spain

Abstract

In this paper, we present the contributions of the I2C-UHU-Altair team to the EXIST2025 Lab at CLEF 2025, addressing the identification and classification of sexism in multimodal online content, particularly memes. Our system leverages recent advances in large language models (LLMs) and vision-language models (VLMs) to process both textual and visual information in a unified manner. We tackle three subtasks: binary classification of memes as sexist or non-sexist, classification of the author's intent behind the meme, and multi-label categorization of sexist content. To enhance model robustness, we adopt the Learning with Disagreement framework, allowing the system to benefit from divergent annotations that reflect the inherent ambiguity and subjectivity in sociolinguistic tasks. We detail our multimodal architecture, preprocessing pipeline, and fine-tuning strategy. Our system demonstrated competitive performance in the shared task, achieving notable positions across all subtasks. Specifically, we ranked 5th in Subtask 2.1 (Soft-Soft), 3rd in Subtask 2.2 (Soft-Soft), and 3rd and 6th in Subtask 2.3 (Soft-Soft and Hard-Hard, respectively). Our findings highlight the potential of multimodal learning for detecting nuanced expressions of sexism in online environments and open avenues for future research in social media moderation and fairness-aware NLP.

Keywords

Sexism in memes, Multimodal learning, Vision-Language Models, Learning with disagreement, Natural language processing

1. Introduction

In the EXIST2025 Lab at CLEF 2025, the I2C-UHU-Altair team extended its previous efforts in the detection and analysis of sexist content on tweets, focusing this time on multimodal data, specifically memes. Unlike traditional text-only classification, memes integrate visual and textual components, often using humor, irony, or ambiguity, making the detection of harmful content more challenging. This year's tasks include: binary classification of memes as sexist or non-sexist, classification of authorial intent (direct or judgmental), and multi-label categorization of sexist content into five nuanced categories.

To address these tasks, we leverage recent advancements in large language models (LLMs) and vision-language models (VLMs), integrating them into a unified pipeline capable of handling multimodal signals. Recognizing the inherent subjectivity in interpreting memes, especially regarding intention and categorization, we adopt the Learning with Disagreement framework [1], which enables the model to learn from the disagreement between annotators rather than being constrained by a single reference label.

Our contributions include: a multimodal classification architecture tailored for memes, the application of Learning with Disagreement in a multimodal context, and a detailed evaluation of model performance

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

†These authors contributed equally.

manuel.guerrero790@alu.uhu.es (M. Guerrero-García); fernando.carrillo051@alu.uhu.es (F. Carrillo-García); mata@uhu.es (J. Mata-Vázquez); vpachon@dti.uhu.es (V. Pachón-Álvarez)

across tasks. The remainder of this paper is structured as follows: we present related works, describe the dataset and annotation process, detail our methodology, report on experimental results from the shared task, and outline future research directions.

2. Related Works

In the realm of detecting sexism in online content, particularly in social media, various approaches have been developed to tackle both explicit and implicit forms of harmful discourse. The EXIST 2025 shared task [2, 3] introduced a multimodal and multi-perspective benchmark, emphasizing the role of ambiguity and disagreement in labeling sexist content across text, memes, and videos. The task builds on previous editions, but incorporates a more diverse set of modalities and adopts the "Learning with Disagreement" paradigm to model subjectivity in annotations.

Binary classification methods have traditionally formed the basis of sexism detection. Early efforts applied machine learning techniques to detect hate speech on Twitter using lexical and contextual features. Similarly, Waseem and Hovy [4] highlighted the importance of incorporating demographic and sociolinguistic features in the detection of hate and sexist language.

Recent advances rely heavily on transformer-based models, which have significantly improved performance in tasks involving subtle linguistic cues and contextual interpretation. The integration of multimodal data—particularly in tasks like meme analysis—necessitates models that can process both visual and textual information effectively. This has led to the exploration of Vision-Language Models (VLMs) in tandem with traditional NLP approaches.

Furthermore, the Learning with Disagreement paradigm [1] is gaining traction as a robust method for handling subjectivity in annotation, especially in sociolinguistic tasks like sexism detection. Rather than forcing annotator consensus, it leverages the distribution of opinions to improve generalization and realism.

Together, these strands of research point toward the need for models that are not only technically robust but also sensitive to the nuanced and multifaceted nature of online sexism, particularly as it appears in multimodal and ambiguous formats.

3. Tasks and Dataset Description

This section describes the tasks in which the I2C-UHU team participated, along with the corresponding datasets provided by the organizers of the EXIST 2025 lab [2, 3].

3.1. Subtask 2.1: Sexism Identification in Memes

This subtask consists of a binary classification task where the objective is to determine whether a meme is sexist. A meme is labeled as *"YES"* if it contains sexist content, describes a sexist situation, or criticizes a sexist behavior. Conversely, memes are labeled *"NO"* if they do not contain any prejudice, stereotyping, or discrimination against women. Importantly, sexism can manifest in various forms—whether seemingly friendly, humorous, offensive, or violent. Thus, subtle sexism is treated with equal importance as explicit expressions.

The annotation guidelines draw on the Oxford English Dictionary definition of sexism: "prejudice, stereotyping or discrimination, typically against women, on the basis of sex." These definitions informed the annotators' decisions, ensuring comprehensive coverage of both overt and covert sexist messages.

3.2. Subtask 2.2: Source Intention in Memes

In this multi-class classification task, the focus shifts to the author's intention behind the sexist content in a meme. Only memes identified as sexist in Subtask 2.1 are considered for this task. Due to the nature of memes, which rarely report events, the *REPORTED* category is excluded. Thus, each meme is classified into one of two categories:

- **DIRECT:** The meme explicitly conveys a sexist message. For example, memes reinforcing traditional gender roles or mocking feminist movements fall into this category.
- **JUDGEMENTAL:** The meme condemns or criticizes sexist behaviors, often employing satire or irony to challenge discriminatory norms.

Understanding author intent is essential for interpreting the function and potential impact of memes in online discourse.

3.3. Subtask 2.3: Sexism Categorization in Memes

This is a multi-label classification task, where each sexist meme is categorized according to one or more of the following sexism types:

- **IDEOLOGICAL AND INEQUALITY:** Memes that deny gender inequality, discredit feminist movements, or portray men as victims of sexism.
- **STEREOTYPING AND DOMINANCE:** Memes that reinforce traditional gender roles or suggest male superiority.
- **OBJECTIFICATION:** Memes that reduce women to physical traits or aesthetic ideals, often neglecting their dignity or personhood.
- **SEXUAL VIOLENCE:** Memes that contain sexual harassment, coercion, or suggestions of sexual assault.
- **MISOGYNY AND NON-SEXUAL VIOLENCE:** Memes that express hatred or promote violence toward women outside the sexual domain.

This categorization facilitates a deeper semantic understanding of the nature and scope of online sexist discourse, providing valuable granularity for model training and evaluation.

3.4. Dataset Description

The dataset used in this work consists of over 5,000 instances in JSON format, each representing a meme annotated by multiple individuals. Annotations include both demographic information of the annotators and various levels of interpretation of the content. The memes are available in two languages—Spanish and English—with a balanced distribution across both. The dataset is split into two partitions: *training* (4,044 instances) and *test* (1,053 instances).

Each JSON object includes the following key fields:

- `id_EXIST`: Unique identifier for each meme.
- `lang`: Language of the content (`en` or `es`).
- `text`: Automatically extracted text from the meme.
- `meme` and `path_memes`: Filename and file path of the meme image.
- `number_annotators`: Number of annotators who labeled the instance.
- `annotators`: Unique identifiers of the annotators.
- Annotators' demographic information:
  - `gender_annotators`: Gender ("F" for female, "M" for male).
  - `age_annotators`: Age group ("18–22", "23–45", "46+"),
  - `ethnicity_annotators`: Self-declared ethnic group.
  - `study_level_annotators`: Highest level of education achieved.
  - `country_annotators`: Country of residence.
- `labels_task2_1`: Binarized sexism labels by each annotator.
- `labels_task2_2`: Labels related to the author's intent.
- `labels_task2_3`: Multiclass categories describing the type of sexism.
- `split`: Dataset partition (`TRAIN` or `TEST`).

3.5. Annotation Analysis and Relevant Statistics

The primary goal of the annotation analysis was to uncover consistent patterns and critical characteristics of the dataset that influence the task of multimodal sexism detection. Understanding these patterns helps to better tailor modeling approaches and evaluation strategies, especially given the nuanced and subjective nature of the task.

The dataset consists of a total of 4,044 memes, each annotated by multiple individuals for various sexism-related categories and author intent. Notably, a large proportion of the memes (approximately 73%) are labeled with multiple sexism types simultaneously, indicating that sexist content in memes is often multifaceted rather than fitting neatly into a single category. On average, each meme in this subset carries about 2.3 distinct sexism labels, with `STEREOTYPING-DOMINANCE` emerging as the most frequently occurring type across the dataset.

Regarding the classification of author intent, the label `DIRECT` is the most common, suggesting that many memes express sexism in an explicit manner. However, the annotation process revealed substantial disagreement among annotators: over 76% of memes showed differences in their binary sexism classification (sexist vs. non-sexist). This high level of annotator divergence underscores the inherent subjectivity in interpreting sexist content, especially in a multimodal context involving images and text.

The distribution of votes among annotators further confirms variability in perceptions of sexism. Some annotators tend to label more instances as sexist, while others are more conservative, highlighting the importance of leveraging *soft labels* or probabilistic approaches that capture this uncertainty instead of relying solely on hard labels.

Overall, this analysis brings to light the complexity of the task, where multiple sexism categories coexist and annotator perspectives vary significantly. These insights are critical for developing robust models that can effectively learn from ambiguous and subjective annotations, emphasizing the need to incorporate disagreement and uncertainty during training, evaluation, and validation phases.

3.6. Data Cleaning and Normalization

Data preprocessing was a critical step to ensure the quality and consistency of both textual and visual inputs. The following procedures were applied:

Text

- **Lowercasing**: All text was converted to lowercase to avoid ambiguities caused by case differences between tokens.
- **Removal of special characters**: Non-alphabetic symbols and irrelevant punctuation were removed, while retaining basic punctuation marks (such as question or exclamation marks) when informative.
- **Whitespace normalization and removal of empty texts**: Extra spaces were normalized and empty text instances were discarded to maintain data integrity.

Images

- **Resizing and formatting**: All images were resized to a standard resolution and converted to RGB format to ensure uniformity.
- **Existence verification**: Each instance was checked to confirm the presence of a valid image file, eliminating entries with missing or invalid paths.
- **Preprocessing for multimodal models**: Images were normalized by scaling pixel values to the [0,1] range and subjected to visual integrity checks.

These cleaning and normalization steps guarantee coherent, noise-free inputs for both text-based and multimodal architectures.

3.7. Data Splitting

The final dataset was divided into two subsets:

- **Training (80%)**: 3,235 instances.
- **Validation (20%)**: 809 instances.

For Subtask 2.3, labels are provided as *soft* distributions over six categories (`NO`, `IDEOLOGICAL-INEQUALITY`, `STEREOTYPING-DOMINANCE`, `OBJECTIFICATION`, `SEXUAL-VIOLENCE`, `MISOGYNY-NON-SEXUAL-VIOLENCE`), with probabilities summing to one. To ensure representative splits, the **dominant class**—defined as the category with the highest probability per instance—was used as the basis for stratified splitting.

The same stratification strategy was analogously applied to Subtasks 2.1 (binary classification) and 2.2 (author intent), according to their respective dominant labels.

4. Methodology

This section describes the different processing pipelines developed for multimodal meme classification.

4.1. Proposed Architectures

To address the complexity of detecting subtle and contextual expressions of sexism in memes, we designed two modular pipelines. Each pipeline incorporates a distinct combination of multimodal understanding and classification components, optimized to balance performance and interpretability.

4.1.1. Pipeline A: BLIP-2 with XGBoost

This pipeline corresponds to one of the official runs submitted to the challenge. It is based on an architecture that combines advanced computer vision and natural language processing techniques. Specifically, it employs the **BLIP-2**[5] model to extract multimodal embeddings from both the image and textual content of each meme. These embeddings are then fed into a **XGBoost**[6] model to perform supervised classification based on the extracted vector representations.

Figure 1 provides an overview of the multimodal architecture used in this pipeline.

This architecture leverages BLIP-2's ability to capture complex semantic relationships between images and text, integrating them into a unified latent space. After processing and normalization, a robust supervised classifier is trained to predict soft labels, enabling interpretable probability distributions over the possible classes. The components of this architecture are described in detail below.

BLIP-2 Architecture and Multimodal Embedding Extraction BLIP-2 is a modular vision-language model that integrates a Vision Transformer (ViT)[7], a Q-Former encoder, and a decoder-based language model (e.g., OPT, T5):

- **ViT Backbone:** Encodes images into sequences of visual tokens.
- **Q-Former:** A cross-attention encoder with N learnable query tokens that extract semantic information from visual tokens.
- **Language Model (LLM):** Consumes Q-Former outputs for generative or comprehension tasks.

The Q-Former aligns vision and language modalities by projecting the visual token set $V = \{v_1, \ldots, v_T\}$ into a shared semantic space via attention over query tokens $\{q_1, \ldots, q_N\}$.

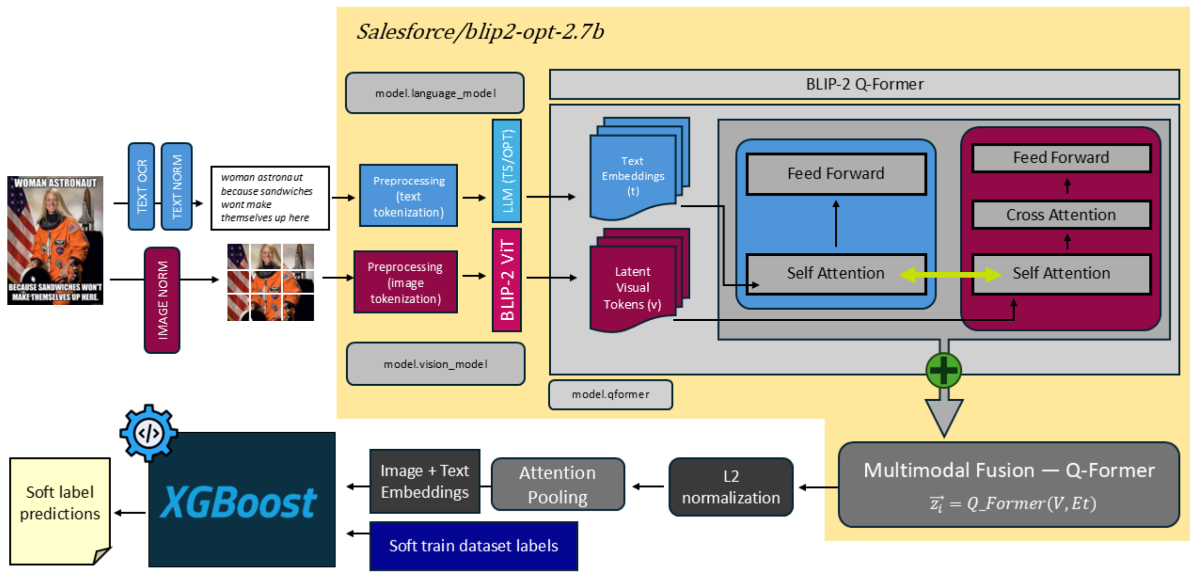

Figure 1: Multimodal architecture for meme analysis based on BLIP-2 for embedding extraction and XGBoost for supervised classification.

Embedding Extraction and Preprocessing Each meme is processed with BLIP-2 to obtain separate embeddings for image and text:

- **Image Embeddings:** Generated via the ViT and Q-Former pipeline, then padded and L2-normalized:

$$\hat{x}_i = \frac{x_i}{\|x_i\|_2}$$

- **Text Embeddings:** Produced by the LLM, also padded and L2-normalized.

Padding ensures uniform batch size, while L2 normalization maintains angular similarity and stabilizes attention mechanisms.

Attention Pooling and Multimodal Fusion To obtain fixed-size vectors, attention pooling is applied over each token sequence:

$$\alpha_t = \frac{e^{e_t}}{\sum_{k=1}^{T} e^{e_k}}, \quad \vec{x} = \sum_{t=1}^{T} \alpha_t e_t$$

The final multimodal representation is formed by concatenating the pooled text and image embeddings:

$$\vec{x}_i = [\vec{t}_i \| \vec{v}_i]$$

Classification with XGBoost The multimodal vectors $\vec{x}_i$ are used to train an XGBoost model with the following setup:

- **Objective:** Multilabel regression using squared error loss (`reg:squarederror`).
- **Parameters:**
  - Number of trees: 300
  - Max depth: 10
  - Learning rate: 0.05

    - L2 regularization: $\lambda = 1$
    - GPU acceleration: `tree_method='gpu_hist'` (NVIDIA L4)
- **Input:** Feature matrix $\mathbf{X} \in \mathbb{R}^{N\times d}$, label matrix $\mathbf{y} \in \mathbb{R}^{N\times C}$
- **Output:** Predicted class distributions:

$$\hat{\mathbf{y}}_i = \text{softmax}\left(\text{XGBoost}(\vec{x}_i)\right)$$

Prediction and Evaluation: During inference, the same embedding and pooling pipeline is applied to the test set. The trained XGBoost model produces raw outputs normalized via:

$$\hat{y}_i = \frac{y_i}{\sum_j y_j} \tag{1}$$

The final predicted class is given by:

$$\text{Predicted class} = \arg\max_j \hat{y}_{ij} \tag{2}$$

The model is evaluated using:

- **Accuracy:** Percentage of correctly classified instances.
- **Macro F1-score:** Average of per-class F1-scores.
- **Macro Precision:** Average of per-class precision scores.

Discussion The BLIP-2 architecture allows the model to effectively capture complex semantic interactions between image and text, outperforming unimodal or naïve fusion approaches. The Q-Former serves as a critical bridge between modalities. XGBoost provides an efficient and robust GPU-compatible classifier that benefits from the high-quality embeddings. The use of soft probabilistic labels improves interpretability and supports applications where confidence estimation is important.

Conclusion Pipeline A proposes a scalable and robust architecture for meme analysis, combining BLIP-2 for multimodal embedding extraction and XGBoost for classification. This fusion of vision-language embeddings, attention pooling, normalization, and supervised learning enables accurate modeling of the intricate semantics in memes. Furthermore, soft label prediction facilitates nuanced interpretation, which is essential for uncertainty-aware classification systems.

4.2. Pipeline B: Multimodal Understanding with Qwen-VL 2.5 and Class Decision via a Mistral-based Model

Following the implementation of Pipeline A utilizing other multimodal models such as BLIP-2, we introduce Pipeline B, which leverages Qwen-VL 2.5 [8] to enhance semantic and contextual accuracy in meme description. This pipeline exploits the optimized architecture of Qwen-VL 2.5, specifically designed for tasks requiring tight visual-linguistic integration, with an emphasis on fine-grained alignment between text and image modalities.

Once meme descriptions are generated by Qwen-VL 2.5, these textual outputs serve as inputs for a subsequent classification stage. This stage employs large language models (LLMs) from the Mistral family[9], including *OpenHermes-2.5*, *Nous-Hermes-2-DPO*[10], and *MythoMax-L2*, among others. These models are all built upon the `Mistral-7B` architecture and have undergone instruction fine-tuning.

To enable few-shot training for each model using a limited number of representative class examples, instances from the training dataset are selected based on a custom curation algorithm. The training set, provided by the EXIST2025 organization and annotated by multiple evaluators, follows a *learning with disagreement* paradigm, where labels may reflect varying perspectives.

The selection algorithm identifies subsets of examples per class exhibiting the highest inter-annotator agreement, thereby serving as clear prototype references. This approach maximizes the quality of prompts provided during few-shot inference, ensuring that the instructions are semantically aligned with the underlying problem distribution. Detailed methodology for selection and evaluation of these instances is described in later sections.

Figure 2 depicts the overall architecture of Pipeline B, illustrating the integration of the multimodal Qwen-VL 2.5 model for meme description generation, followed by a textual classification phase utilizing fine-tuned Mistral family models.

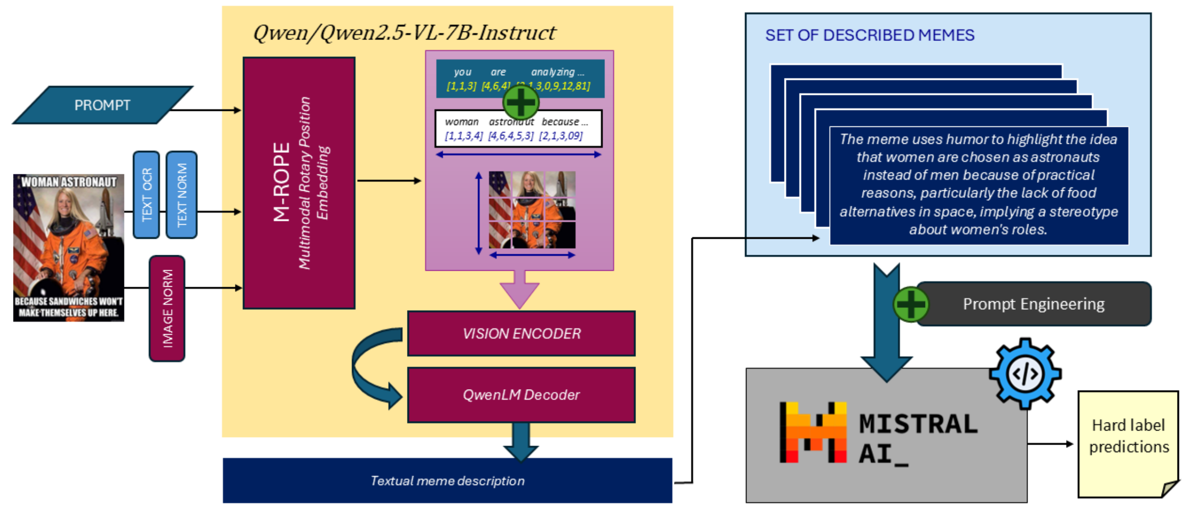

Figure 2: Overview of Pipeline B: Qwen-VL 2.5 is used for multimodal understanding and description generation, which are then classified using Mistral models trained in a few-shot manner with carefully curated examples.

4.2.1. Multimodal Semantic Description of Memes with Qwen-VL 2.5

General Architecture of Qwen-VL 2.5 Qwen-VL 2.5 is a multimodal extension of the Qwen 2.5 model, maintaining an enhanced GPT-style Transformer decoder adapted to process both textual and visual inputs. Its design incorporates:

- **Grouped Query Attention (GQA)**: Improves the efficiency of KV cache usage, particularly relevant for long-context generation.
- **SwiGLU Activation Function**[11]: Provides greater non-linear modeling capacity.
- **Rotary Positional Embeddings (RoPE)**[12]: Enables generalization to very long contexts (up to 128K input tokens).
- **RMSNorm and Bias in QKV Projections**: Enhances training stability.
- **BBPE Tokenization**: Expanded vocabulary of 151,643 tokens plus 22 control tokens for multimodal tasks.

On the visual side, it integrates a Vision Transformer (ViT-G/14) encoder, pretrained on large multimodal corpora, which allows maintaining the native input resolution (without destructive cropping) and supports both images and videos.

Rationale for Using Qwen-VL 2.5 Qwen-VL 2.5 was chosen in this project due to several technical advantages:

- **Direct Multimodal Input Capability**: Unlike previous models requiring input adaptation (e.g., CLIP + GPT-2 combinations), Qwen-VL 2.5 natively integrates image, text, and video within a unified semantic space.
- **Fine-Grained Fusion via Cross-Attention**: Cross-attention layers between image and text modalities improve the visual grounding of generated language.

- **Native Resolution Handling**: Qwen-VL accepts visual inputs at full resolution, avoiding loss of key visual information crucial in memes.
- **Enhanced Causal and Contextual Modeling**: Inherits substantial improvements from Qwen 2.5 in instruction following, long generation, and structured reasoning.

Prompt Engineering Prompt engineering is a key technique in designing applications based on language models, involving careful formulation of instructions to guide model behavior and obtain coherent, relevant, and useful outputs.

In this case, the prompt was designed to request a description of the meme focusing on its social or gender implications, avoiding unnecessary visual details. The goal is to maximize the relevance of the description for sexist content detection, preventing neutral or overly generic interpretations.

Example Prompt Used

```
You are analyzing a meme for sexism classification.
Given the image and its embedded text, describe briefly and concisely
the main situation or message conveyed. Focus on the social or
gender implications. Avoid unnecessary visual details.
Limit to one or two sentences.
Describe this meme:
```

This prompt is embedded within a multimodal input message containing both the textual prompt and the meme image encoded in base64. The model employed, **Qwen/Qwen2.5-VL-7B-Instruct**, is an instructive visual language model capable of jointly processing both input types.

Technical Implementation The implementation was performed in a Google Colab environment with access to a dedicated GPU (typically **NVIDIA T4** with 16 GB VRAM or **NVIDIA A100** with 40 GB).

To run Qwen-VL 2.5 efficiently, a GPU with at least 12–16 GB memory is required, as the full model consumes between **11 and 14 GB VRAM** during inference, depending on backend settings and batch size.

Generated Description Examples Some examples of descriptions generated for various memes are:

- **Input**: Image of a man pointing to a kitchen with the text "where you should be".
  Generated Description: "The meme suggests that a woman's place is in the kitchen, reinforcing traditional gender roles."
- **Input**: Image of a job interview where the interviewer asks: "Are you planning to have children?".
  Generated Description: "The meme implies that women face discrimination in job interviews due to potential motherhood."
- **Input**: Image of a woman driving poorly with the text "that's why they shouldn't have a license".
  Generated Description: "The meme portrays women as bad drivers, perpetuating a sexist stereotype."

These examples demonstrate that the model captures not only literal content but also the social implications, which is essential for classification.

Justification of the Approach Using descriptions generated by multimodal models abstracts meme content into a semantically useful representation for classification tasks. This reduces complexity compared to directly processing images and embedded text and improves system interpretability.

Additionally, prompt engineering ensures the model focuses on social and gender aspects, avoiding drift toward neutral or visually oriented descriptions.

This approach has proven effective in generating coherent, relevant, and specific descriptions, constituting a valuable tool for automated sexist discourse analysis in digital content.

4.2.2. Textual Classification Using Mistral-based Models: Architecture, Rationale, and Few-Shot Technique

After automatic generation of semantic meme descriptions via Qwen-VL 2.5, these descriptions are fed as textual input into a supervised classification process. This stage employs instruction-tuned language models based on **Mistral-7B**, a Transformer decoder architecture optimized for efficient inference and contextual reasoning.

Mistral-7B Architecture and Variants **Mistral-7B** is a 7-billion parameter autoregressive decoder-only model incorporating several key optimizations:

- **Sliding Window Attention**: Allows handling of long contexts while maintaining computational efficiency.
- **Grouped Query Attention (GQA)**: Inherited from models such as Qwen, this technique reduces memory costs during inference without sacrificing attention capability.
- **SwiGLU Activations and RoPE**: SwiGLU activations combined with rotary positional embeddings improve reasoning and generalization in long sequences.

Several fine-tuned variants built upon the Mistral-7B base are employed in this project, including:

- **OpenHermes-2.5**: Targeted at general reasoning tasks with instruction following.
- **Nous-Hermes-2-DPO**: Fine-tuned via Direct Preference Optimization to align responses with human preferences.
- **MythoMax-L2** and **Bagel-7B**: Specialized in semantic understanding and natural language instruction following.

Rationale for Use in the Pipeline The use of these models is justified by several technical and functional criteria:

1. **High Performance in Contextual Semantic Classification**: These models have demonstrated advanced capabilities in interpreting, categorizing, and reasoning about text descriptions with social or subjective nuances, as found in memes.
2. **Few-Shot Learning Capability**: Instruction tuning enables generalization from a limited number of labeled examples without explicit retraining.
3. **Compatibility with Efficient Quantization (GGUF)**: Quantized weights (e.g., `Q4_K_M`, `Q6_K`) allow local execution on GPU or CPU, facilitating experimentation and reproducibility.

Few-Shot Prompting and Classification Technique Classification is performed via instructive few-shot prompting. The model receives a general instruction along with 3 to 5 manually annotated examples as context. These examples are selected from the training set via a curation algorithm maximizing class representativeness.

This algorithm operates on a dataset annotated by multiple human evaluators under the *Learning With Disagreement* framework. Here, a single instance may have multiple labels. The selected examples are those with highest inter-annotator agreement, ensuring that presented examples are prototypical and semantically clear, thereby improving inference accuracy. This setup allows the model to induce labeling logic directly from representative examples, removing the need for explicit retraining and maintaining high flexibility across multiple tasks (e.g., sexism, offensive humor, symbolic violence).

4.3. Data Preparation Using Learning with Disagreement

To enhance the reliability of annotated data and optimize the selection of representative examples, two strategies based on *Learning with Disagreement* (LwD) were applied, tailored specifically to the requirements of pipelines A and B.

Pipeline A: Supervised Label Curation The supervised model based on XGBoost required a single final label per instance. However, since each meme was annotated multiple times with sometimes conflicting labels, a robust aggregation process was applied consisting of:

- Identifying the majority label for each subtask (2.1 and 2.2), discarding UNKNOWN or invalid responses.
- For subtask 2.3 (multi-label), considering labels that surpassed a minimum frequency threshold (at least 50% of annotators), only if the instance had been previously classified as sexist in subtask 2.1.

This procedure was implemented via the `compute_final_labels_from_df` algorithm, which produces a dictionary of consensus final labels per instance.

Subsequently, the consistency and accuracy of each annotator were evaluated based on their agreement with the final labels. This was achieved using an algorithmic function that computes metrics such as accuracy (for subtasks 2.1 and 2.2) and Jaccard similarity (for 2.3), generating a global ranking. This analysis was key to filtering out unreliable or inconsistent annotators. The ranking of annotators from best to worst, including demographic variables for contextualizing individual performance, is presented in Table 1.

Table 1
Annotator ranking based on accuracy for subtasks and Jaccard similarity for subtask 2.3

Annotator ID	2.1 Acc.	2.2 Acc.	2.3 Jaccard	Global	Gender	Age	Ethnicity	Study Level	Country
Annotator_740	0.741	0.920	1.000	**0.887**	M	46+	White or Caucasian	Bachelor's	Portugal
Annotator_708	0.926	0.880	0.781	0.862	F	46+	White or Caucasian	Master's	United Kingdom
Annotator_630	0.889	1.000	0.692	0.860	M	23-45	White or Caucasian	Bachelor's	United Kingdom
Annotator_821	0.741	0.917	0.917	0.858	M	23-45	Black or African American	Bachelor's	Zimbabwe
Annotator_689	0.889	0.846	0.833	0.856	F	18-22	White or Caucasian	Bachelor's	Estonia
Annotator_859	0.222	0.381	0.095	0.233	F	18-22	Other	High school	South Africa
Annotator_850	0.222	0.333	0.114	0.223	M	18-22	White or Caucasian	Bachelor's	France
Annotator_162	0.296	0.182	0.123	0.200	M	23-45	Hispano or Latino	Bachelor's	Chile
Annotator_736	0.074	0.040	0.333	0.149	F	23-45	Black or African American	High school	South Africa
Annotator_300	0.037	0.053	0.000	**0.030**	M	23-45	Hispano or Latino	Bachelor's	Chile

Pipeline B: Few-Shot Example Selection Pipeline B, based on large language models (Qwen-VL 2.5 and Mistral), required high-quality training examples per class. To select these examples, a strategy focused on identifying instances with full annotator consensus was employed:

- Instances where all six annotators assigned exactly the same label for each subtask were selected.
- For subtask 2.3 (multi-label), instances with identical normalized label sets were considered matching.

This process was implemented using the `get_top_agreed_instances` function, which returns the 10 most representative instances per class and subtask, thus ensuring the quality and clarity of examples used in few-shot training.

Summary of Impact The Learning with Disagreement approach not only increased the reliability of the training data but also improved the interpretability and coherence of the developed models. By reducing annotator noise and prioritizing quality over quantity, the approach enabled better utilization of the available dataset, especially in a context where multiple annotations were rich but inconsistent.

4.4. Model Training and Evaluation

4.4.1. Baselines

We employed traditional unimodal baselines using widely adopted Transformer architectures such as **BERT-base** and **XLM-RoBERTa-base**, fine-tuned exclusively on the textual content extracted

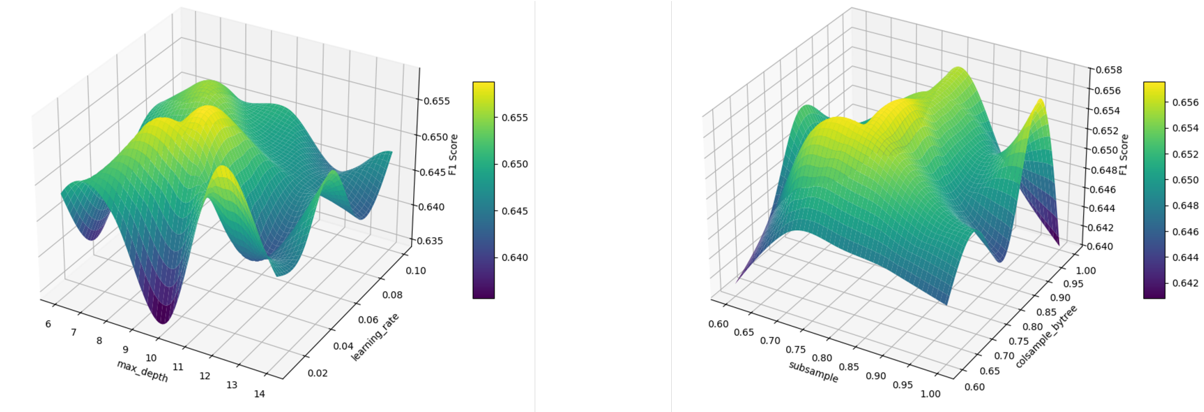

Figure 3: The F1 scores obtained from validation are interpolated using cubic smoothing, offering a compact visual summary of interaction effects and optimal regions in the search space for each class-specific model.

from memes. These baselines serve as foundational reference points to benchmark the performance of subsequent multimodal approaches.

4.4.2. Training Setup

Models were trained using standardized hyperparameters and optimization schemes. For Transformer-based baselines, we utilized AdamW with a learning rate of 3×10^{-5}, batch size of 32, sequence length truncated to 128 tokens, weight decay of 0.01, and a total of 5 training epochs. Early stopping was implemented based on validation loss plateauing to avoid overfitting. Multimodal pipelines integrating BLIP2 embeddings with XGBoost classifiers were optimized via grid search over key hyperparameters such as `max_depth`, `learning_rate`, `subsample`, and `colsample_bytree`. Figure 3 presents the hyperparameter tuning landscape for the XGBoost classifier trained for the `YES` class in Subtask 2.1. Each sub-plot illustrates the effect of two key hyperparameters (e.g., `max_depth` vs. `learning_rate`, and `subsample` vs. `colsample_bytree`), with remaining parameters fixed.

4.4.3. Evaluation Metrics and Protocol

Performance was assessed using task-specific metrics: binary classification subtasks employed macro-averaged F1 score, intent classification utilized balanced precision and recall, and soft multilabel subtasks considered micro-averaged F1 scores. To ensure fairness and comparability across pipelines, evaluation adhered to uniform protocols with identical test splits. Additionally, extensive error analysis and qualitative review of misclassifications were conducted to identify systematic weaknesses and guide model refinement.

5. Official Results Summary

In the 2025 edition of EXIST, two main pipelines, `I2C-UHU-Altair_1` and `I2C-UHU-Altair_2`, were evaluated across three subtasks under both Soft-Soft and Hard-Hard modalities.

- **Task 2.1 (Soft-Soft):** Pipeline `I2C-UHU-Altair_1` achieved a competitive performance, ranking **5th place**.
- **Task 2.1 (Hard-Hard):** Both pipelines attained solid results, placing among the top-ranked systems.
- **Task 2.2 (Soft-Soft):** Pipeline `I2C-UHU-Altair_1` showed strong performance, securing **3rd place**.
- **Task 2.2 (Hard-Hard):** The pipelines maintained competitive results despite the increased evaluation difficulty.

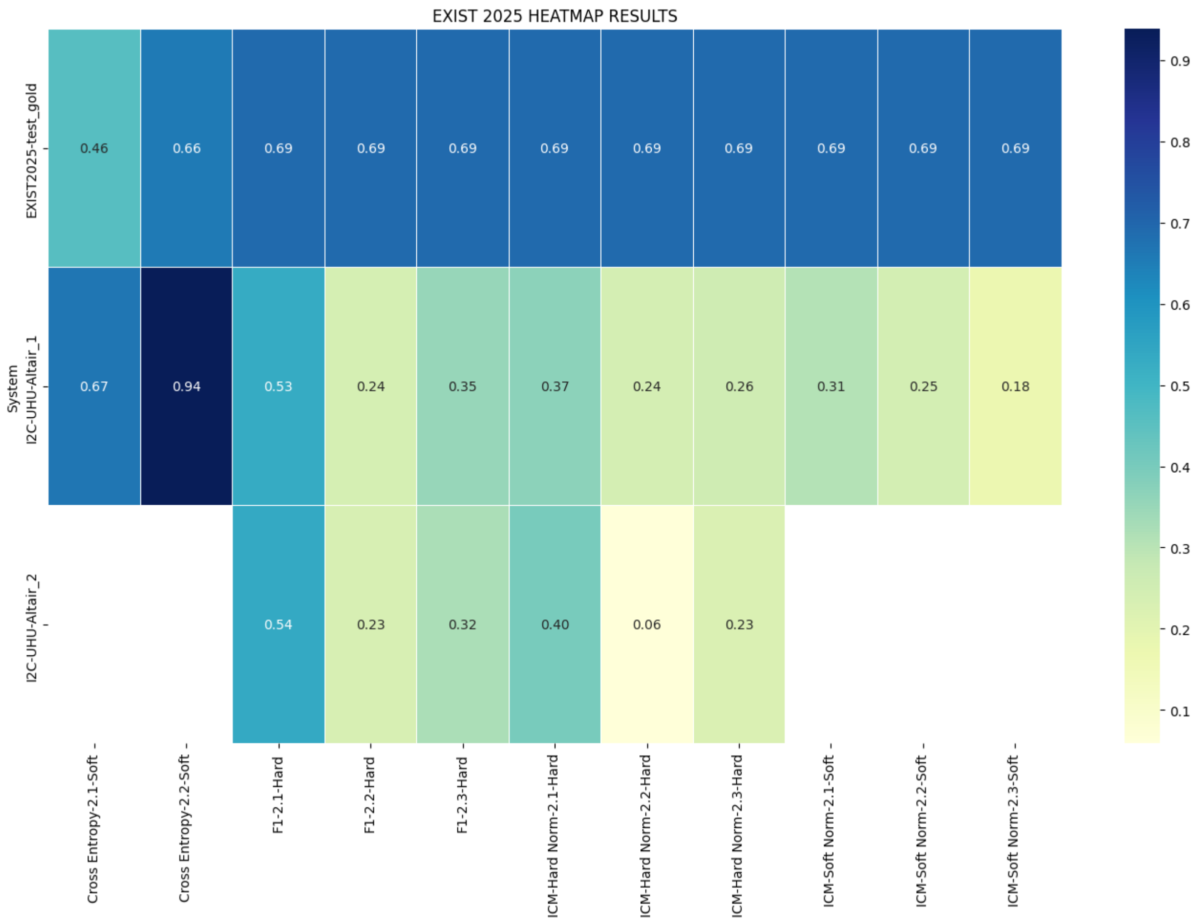

Figure 4: The heatmap underscores the relative strengths and weaknesses of each pipeline, offering a comprehensive view of their behavior across subtasks 2.1, 2.2, and 2.3

- **Task 2.3 (Soft-Soft):** Pipeline `I2C-UHU-Altair_1` ranked **3rd place**.
- **Task 2.3 (Hard-Hard):** Achieved **6th place** with consistent F1 scores.

Figure 4 presents a heatmap visualization of all evaluation metrics across different systems and subtasks within the EXIST 2025 challenge. This representation facilitates a holistic comparison by emphasizing both performance disparities and consistent trends across soft-soft and hard-hard evaluations. The color gradient enables clearer interpretation of metric magnitudes, particularly in the presence of extreme values that obscure middle-range variations in standard bar plots.

5.1. Key Conclusions

The experimental evaluation across Tasks 2.1 to 2.3 demonstrates that both proposed pipelines — *I2C-UHU-Altair_1* (based on BLIP2 and XGBoost) and *I2C-UHU-Altair_2* (based on QWEN for caption generation and Mistral for classification) — achieve competitive performance, frequently ranking among the top submissions in multiple tracks. The following key conclusions can be drawn:

- **Effectiveness of probabilistic modeling in soft evaluations:** The I2C-UHU-Altair_1 pipeline consistently outperformed in the soft-soft evaluation tracks, obtaining second place in Task 2.2 and third in Task 2.3. These results indicate that the combination of vision-language embeddings from BLIP2 with probabilistic reasoning via XGBoost yields well-calibrated outputs capable of capturing semantic ambiguity and subtle expressions of misogyny in memes.
- **Robustness in hard classification by language modeling:** In contrast, I2C-UHU-Altair_2 shows improved performance in hard evaluation tasks, such as Task 2.1 (F1 YES = 0.7125) and Task

2.3 (Macro F1 = 0.3786), ranking within the top 10. This suggests that the use of QWEN to enrich visual input with detailed captions provides Mistral with more context for firm classification decisions. However, this approach appears less effective in probabilistic calibration, potentially due to the literal or contextually imprecise nature of generated captions.
- **Difficulty of modeling irony and indirect misogyny:** Performance drops significantly on minority or nuanced classes (e.g., indirect misogyny or sexualized humor) as evidenced in Task 2.3, where even the best models experience a marked decline in Macro F1. This reflects the inherent challenge in meme classification where irony, satire, and multimodal ambiguity demand high-level pragmatic reasoning.
- **Multiclass prediction favors vision-language fusion:** Task 2.3 further highlights the advantage of Altair_1, whose higher Macro F1 suggests better coverage across minority classes. This points to the strength of deep vision-language representation learning (via BLIP2) combined with tree-based decision modeling (XGBoost) for handling nuanced and class-imbalanced scenarios.
- **Complementarity of soft and hard evaluation metrics:** The observed divergence between soft and hard performance underlines the importance of using both evaluation paradigms. Soft metrics reward nuanced probabilistic reasoning, while hard metrics assess decisiveness and class-level discrimination. A comprehensive evaluation of system behavior benefits from jointly considering both.

Overall, the BLIP2 + XGBoost pipeline (Altair_1) demonstrates higher robustness and semantic sensitivity, particularly in tasks requiring fine-grained or probabilistic interpretation. In contrast, the QWEN + Mistral pipeline (Altair_2) delivers competitive hard predictions but may be limited by the variability and literalness of generated meme captions. These results validate the efficacy of multimodal fusion for tackling the inherently ambiguous and context-sensitive task of misogyny detection in memes.

6. Conclusions and Future Work

This work has demonstrated competitive performance in the 2025 edition of the EXIST challenge by leveraging a hybrid visual-textual approach combining the BLIP model with XGBoost classifiers. The pipelines `I2C-UHU-Altair_1` and `I2C-UHU-Altair_2` consistently achieved strong results across multiple subtasks and evaluation modalities (Soft-Soft and Hard-Hard), validating the effectiveness of multimodal fusion strategies. Key achievements include:

- Securing top rankings such as 3rd place in Task 2.2 and Task 2.3 under Soft-Soft evaluation.
- Obtaining competitive F1 scores and normalized ICM metrics in Hard-Hard evaluations, comparable with state-of-the-art complex systems.
- Demonstrating robustness and generalization potential of hybrid models that integrate complementary visual and textual features.

These results confirm the viability of combining pre-trained multimodal transformers with gradient boosting techniques for complex misinformation detection tasks.

6.1. Future Work

Future research directions include both the refinement of the current pipeline architectures and the exploration of advanced multimodal approaches that integrate additional modalities such as audio and video.

- **Pipeline optimization:** The BLIP2 + XGBoost pipeline could be enhanced by replacing the XGBoost classifier with transformer-based models, such as RoBERTa or DeBERTa, fine-tuned on meme-specific textual embeddings. Incorporating contextual prompt generation using BLIP-2's visual-question answering capabilities may also contribute to a more nuanced understanding of implicit meaning in multimodal content.

- **Caption generation improvements:** In the case of the QWEN + Mistral pipeline, future iterations may benefit from the use of instruction-tuned vision-language models (e.g., InstructBLIP, MiniGPT-4) for the generation of socially grounded and contextually relevant captions. Classification could be performed using multimodal large language models such as LLaVA or GPT-4V, which allow for more integrated reasoning over vision and text inputs.
- **Contrastive representation learning:** To improve the detection of subtle phenomena such as irony or covert hate speech, contrastive learning approaches (e.g., CLIP-style encoders) tailored to affective or sociocultural dimensions could be adopted. The use of curated datasets with fine-grained annotations would facilitate training of models with enhanced cultural sensitivity.
- **Incorporation of dynamic modalities:** Given the increasing prevalence of misogynistic content in short-form videos (e.g., TikTok, Instagram Reels), extending current approaches to support audio-visual content is essential. Multimodal models capable of aligning visual, audio, and textual streams—such as VideoCLIP, VATT, or Flamingo—should be explored for robust temporal and semantic fusion.
- **Development of real-time moderation systems:** Future work could involve the deployment of low-latency systems for real-time content moderation. Lightweight multimodal architectures, optimized for inference on edge devices, may enable integration into content platforms for early detection and prevention of harmful speech.
- **Explainability and ethical alignment:** To enhance transparency, attention should be directed toward incorporating interpretability mechanisms such as multimodal attention heatmaps, natural language rationales, or model critique techniques (e.g., Chain-of-Thought or Reflexion prompting). These approaches may assist in aligning system outputs with ethical guidelines and user expectations.

Overall, future efforts are expected to advance towards context-aware, temporally sensitive, and ethically aligned multimodal architectures capable of addressing the evolving landscape of online misogyny, particularly in formats beyond static image-text memes.

Acknowledgments

This paper is part of the I+D+i Project titled *"Conspiracy Theories and hate speech online: Comparison of patterns in narratives and social networks about COVID-19, immigrants, refugees and LGBTI people [NON-CONSPIRA-HATE!]"*, PID2021-123983OB-I00, funded by *MCIN/AEI/10.13039/501100011033/* and by *"ERDF/EU"*.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] A. Uma, V. Prabhakaran, 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.

[2] L. Plaza, J. C. de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of exist 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[3] L. Plaza, J. C. de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of exist 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos (extended overview), in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[4] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL Student Research Workshop, 2016, pp. 88–93. URL: https://aclanthology.org/N16-2013.

[5] J. Li, D. Li, X. Xie, W. Lu, S. C. H. Wang, Y. Loh, J. Wang, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, arXiv preprint arXiv:2301.12597 (2023).

[6] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 785–794.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[8] Q.-V. Team, Qwen-vl and qwen-vl-chat: A billion-scale vision-language model with strong multi-modal capabilities, https://huggingface.co/Qwen/Qwen-VL, 2024. https://github.com/QwenLM/Qwen-VL.

[9] M. AI, Mistral 7b, https://huggingface.co/mistralai/Mistral-7B-v0.1, 2023. https://mistral.ai/news/announcing-mistral-7b.

[10] teknium, Openhermes 2.5 - mistral 7b model, 2024. https://huggingface.co/teknium/OpenHermes-2.5-Mistral-7B.

[11] N. Shazeer, Glu variants improve transformer, arXiv preprint arXiv:2002.05202 (2020).

[12] J. Su, Y. Lu, S. Pan, B. Wen, Y. Liu, Roformer: Enhanced transformer with rotary position embedding, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 115–124.