

When in Doubt, Consult: Expert Debate for Sexism Detection via Confidence-Based Routing

Anwar Alajmi^{1,2}, Gabriele Pergola¹

¹Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK

²College of Business Studies, Public Authority of Applied Education and Training, Kuwait
{anwar.alajmi, gabriele.pergola.1}@warwick.ac.uk

Abstract

Online sexism increasingly appears in subtle, context-dependent forms that evade traditional detection methods. Its interpretation often depends on overlapping linguistic, psychological, legal, and cultural dimensions, which produce noisy and sometimes contradictory signals in annotated datasets. These inconsistencies, combined with label scarcity and class imbalance, result in unstable decision boundaries and cause fine-tuned models to overlook subtler, underrepresented forms of harm. To address these challenges, we propose a two-stage framework that unifies (i) targeted training procedures to better regularize supervision to scarce and noisy data with (ii) selective, reasoning-based inference to handle ambiguous or borderline cases. First, we stabilize the training combining class-balanced focal loss, class-aware batching, and post-hoc threshold calibration, strategies for the first time adapted for this domain to mitigate label imbalance and noisy supervision. Second, we bridge the gap between efficiency and reasoning with a dynamic routing mechanism that distinguishes between unambiguous instances and complex cases requiring a deliberative process. This reasoning process results in the novel *Collaborative Expert Judgment* (CEJ) module which prompts multiple personas and consolidates their reasoning through a judge model. Our approach outperforms existing approaches across several public benchmarks, with F1 gains of +4.48% and +1.30% on EDOS Tasks A and B, respectively, and a +2.79% improvement in ICM on EXIST 2025 Task 1.1.

1 Introduction

Sexist content on social media is increasingly pervasive, often appearing in subtle, sarcastic, or

¹**Warning:** This paper includes examples that may contain explicit, offensive, or harmful language.

²Code is available at: <https://github.com/anonymous-project-2025/SexismDebate>.

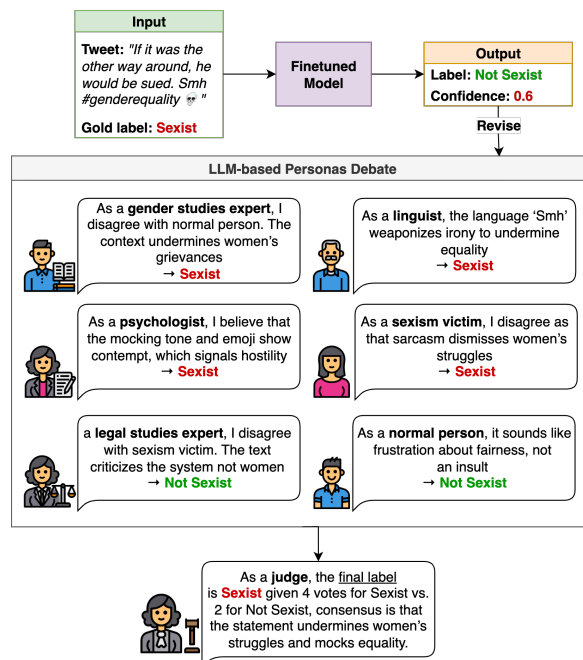


Figure 1: Overview of the proposed framework.

context-dependent forms that evade traditional classification methods. From seemingly innocuous stereotypes to overt hate speech, such content contributes to the emotional and psychological harm that disproportionately affects women (Abercrombie et al., 2023).

These diverse forms of sexism are shaped by distinct yet overlapping factors: linguistic, psychological, legal, and cultural, among others. As a result, both human annotators and automated systems must contend with mixed and sometimes contradictory signals: a message may appear harmless to a general audience, yet be perceived as harmful by a legal expert, a psychologist, or someone with lived experience of gender-based discrimination (Mostafazadeh Davani et al., 2022).

This multidimensionality creates several challenges for annotation and automatic detection. Even the largest and most reliable datasets (Kirk et al., 2023; Plaza et al., 2025) are typically small, highly imbalanced, and marked by substantial an-

notator disagreement not only due to ambiguity in individual examples, but also because different annotators may implicitly prioritize different criteria (e.g., linguistic form, psychological impact, cultural normativity, or legal definitions) (Khan et al., 2025a). These mixed signals can overshadow the representation of less explicit but equally damaging forms of sexism, introducing noise and instability into both the data and the models (Tanno et al., 2019; Jiang et al., 2024). At the same time, large language models (LLMs) are pretrained on vast, heterogeneous data sources, absorbing social norms, stereotypes, which are inconsistently represented, and at times in contradiction with human feedback. As a result, fine-tuning these large models on small, noisy sexism datasets is often insufficient to resolve these conflicts, and may inadvertently reinforce dominant or explicit patterns while neglecting subtler cases (Khan et al., 2025a). These limitations underscore the need for models that explicitly account for the combined effect of underrepresentation, noise, and conceptual ambiguity in both data and predictions.

To address these challenges, we introduce a two-stage framework that combines (i) targeted training procedures to adapt supervision to scarce and noisy data, with (ii) selective, reasoning-based inference to handle ambiguous cases. This joint design enables the specialized model to make efficient decisions when possible, while engaging with finer contextual reasoning when necessary. During (i) training, we mitigate long-tailed label distributions and inconsistent annotation through three well-known training strategies. These have never been adapted into an integrated framework for such tasks: the *class-balanced focal loss* (CB-Focal) (Cui et al., 2019), which amplifies the contribution of minority and difficult examples, and the *class-aware batching*, which ensures uniform exposure to underrepresented classes during optimization. We further employ *post-hoc threshold calibration* to stabilize the decision boundaries of the specialist classifier and improve robustness under noisy supervision. At (ii) inference time, we deploy confidence-aware dynamic routing, which filters the model predictions by difficulty: high-confidence samples are directly classified, while uncertain cases are escalated to a *Collaborative Expert Judgment* (CEJ) module. The CEJ module implements a structured reasoning pipeline composed of multiple *personas*, such as a linguist, a gender expert, a psychologist, and an everyday speaker, each prompted to analyze the same input from distinct perspectives. Their

deliberations are subsequently synthesized by a judge model, which integrates these perspectives into a final, interpretable decision (Figure 1). The aim is not to faithfully replicate human interaction, but rather to approximate its structure to activate the reasoning and knowledge that instruction-tuned LLMs already encode, yet rarely employ in single-step classification, as evidenced by the limited predictive power of zero-shot baselines (Kirk et al., 2023; Plaza et al., 2025; Khan et al., 2025a).

Our experimental evaluation on both the EDOS and EXIST 2025 datasets demonstrates that our methods outperform existing state-of-the-art approaches across almost all tasks. On EXIST 2025 (Task 1.1), the combination of targeted training procedures and confidence-aware routing yields an increase of over 5 F1 points, while also setting new benchmarks on ICM and ICM-Norm metrics. For EDOS, we observe significant improvements in both binary (Task A) and multi-class (Task B) classification: selective escalation with the CEJ module boosts macro F1 by 3.9 points on Task A and 1 point on Task B. Notably, our approach surpasses previous baselines without employing any ensemble or data augmentation methods.

Our contributions can be summarized as follows:

- *A unified framework with selective reasoning capability:* We introduce a unified framework for sexism detection that combines the unmatched efficiency of neural classifiers and targeted training procedures, such as class-balanced focal loss, class-aware batching, and post-hoc threshold calibration, with a reasoning-based inference mechanism that is selectively applied to challenging cases.
- *Collaborative Expert Judgment:* We introduce a novel module, *Collaborative Expert Judgment* (CEJ), that leverages multi-persona prompting to resolve ambiguous or borderline cases, enabling the activation of reasoning encoded in instruction-tuned LLMs.
- *Comprehensive experimental assessment:* We provide a comprehensive analysis of where multi-persona reasoning delivers the greatest benefit, particularly in handling ambiguous and minority cases, and discuss the limitations under severe class imbalance. We demonstrate that integrating these components achieves state-of-the-art results on almost all tasks of both the EXIST 2025 and EDOS datasets.

2 Related Work

Our work is aligned with at least two research lines on sexism detection and persona-based analyses.

Sexism Classification. Annotator disagreement, often shaped by cultural perspectives, complicates sexism detection (Mostafazadeh Davani et al., 2022). Initiatives like the EXIST 2024 shared task encouraged methods robust to inconsistent annotations (Khan et al., 2024). Successful approaches leveraged transformer-based models such as DistilBERT (Sanh et al., 2019), DeBERTa (He et al., 2020), RoBERTa (Liu et al., 2019), and hybrid ensembles that utilized data augmentation and multi-task learning. When it comes to LLMs, previous studies have employed adversarial evaluations and data augmentation to enhance detection accuracy. For example, Samory et al. (2021) highlighted standard model vulnerabilities through psychologically grounded adversarial datasets, emphasizing the need for psychologically informed robustness. Similarly, Khan et al. (2025a) addressed annotator disagreements through definition-driven data augmentation and ensemble methods, improving data-level robustness.

Beyond explicit abuse, a growing challenge lies in detecting nuanced or implicit forms of hate. The work by Zeng et al. (2025) shows that LLMs often fail to recognize metaphorical expressions of hate speech, highlighting substantial gaps in current moderation capabilities. Their findings emphasize that even state-of-the-art models struggle when hateful intent is obscured through indirect language, a phenomenon that is particularly relevant for sexist discourse. Building on this, Fasching and Lelkes (2025) demonstrate that hate speech detection outcomes vary significantly across LLM-based systems, underscoring how moderation decisions remain model-dependent and inconsistent. Our work directly responds to these challenges by integrating confidence-aware routing and multi-expert arbitration, ensuring that inconsistent or ambiguous cases are addressed through structured reasoning and synthesis.

Persona-based Approaches. More recent work integrates multi-agent architectures, expert simulations, and human-in-the-loop reasoning to improve transparency and social norm alignment (Pergola et al., 2019; Lyu and Pergola, 2024; Tan et al., 2025a,c,b). While expert persona-driven prompting strategies, such as those in Xu et al. (2023) and Long et al. (2024), have advanced interpretability, our method surpasses static personas by introduc-

ing dynamic, structured, and confidence-aware interactions among multiple expert roles. Moreover, our framework integrates a dedicated judge model to arbitrate conflicts, providing nuanced decisions beyond simple voting mechanisms.

Emerging research demonstrated the effectiveness of multi-expert deliberation in LLM-based decision-making. Lu et al. (2024) introduce a multi-agent framework for clinical triage where LLM agents assume distinct roles, engage in multi-round discussions with self-confidence scoring, and iteratively revise their assessments until reaching consensus. Similarly, Long et al. (2024) simulate multiple domain experts within a single LLM, aggregate their responses, and select the best one. Both approaches show that structured discussion among diverse perspectives yields more reliable outputs than single-viewpoint reasoning, as cross-perspective validation mitigates individual biases and enables principled conflict resolution before final judgment. Our framework extends this paradigm to sexism detection, where the inherently subjective nature of the task makes perspectival diversity essential.

Another notable study that effectively leverages expert agents is Expert-Token-Routing (Chai et al., 2024), which adopts a router-based approach similar to that of Jiang et al. (2023). In this method, a router model is trained to direct each question or instruction to a specialized expert, where the expert LLM is responsible for generating the next token. Their approach encodes expert LLMs as special tokens within the vocabulary of the meta-LLM, drawing inspiration from ToolkenGPT (Hao et al., 2023).

Our progressive prompt refinement aligns with iterative human-in-the-loop methods discussed in Shah (2025), systematically enhancing logical coherence and prompting effectiveness. In addition, we preserve the logical structure and enhance the prompt formatting for effective model reasoning as it is shown by Li et al. (2025)’s experiments.

3 Methodology

We introduce a two-stage framework for sexism detection. The two stages are based on the core premise that not all instances require complex reasoning. And while many can be better resolved by the specialist model, forcing indiscriminately an immediate decision on ambiguous cases leads to errors. We therefore implement a selective classification mechanism that acts as a gatekeeper between the (i) specialised classifier and the (ii) reasoning module. The specialized model is fine-tuned mov-

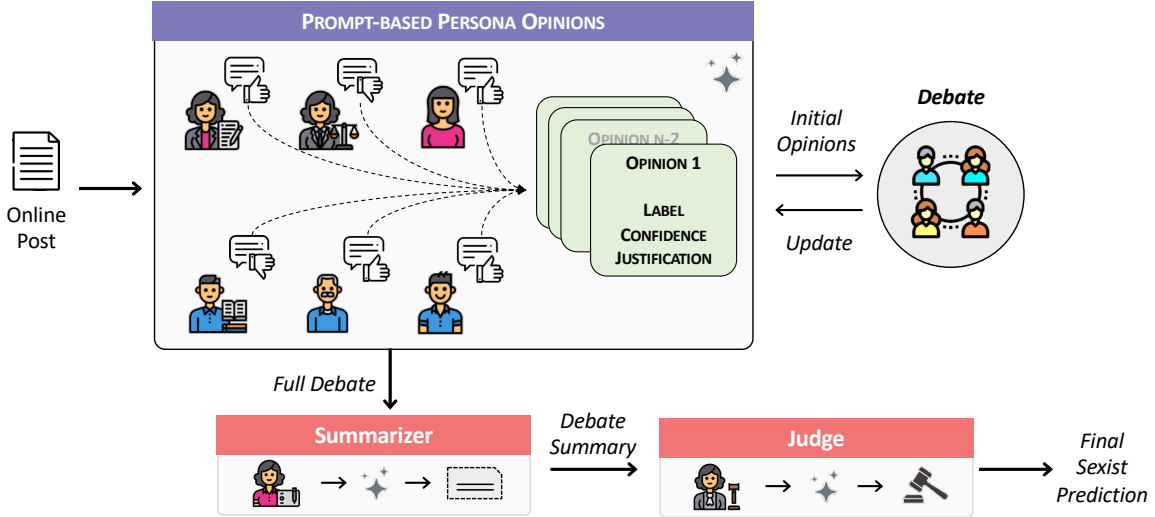


Figure 2: CEJ framework, where a collaborative reasoning discussion among personas with refinement is generated before the judge model makes a definitive decision.

ing beyond standard cross-entropy minimization by integrating three regularization strategies designed to stabilize learning on the unstable decision boundaries caused by the severe class imbalance and annotator disagreement. Then, we couple this mechanism with a confidence-aware routing mechanism which selects low-confidence instances the needs to be escalated to the *Collaborative Expert Judgment* (CEJ) module, where multiple personas are designed to reason, deliberate, and produce the final decision (Figure 2).

3.1 Domain-Tuned Model

We first fine-tune a small instruction-tuned LLM as a domain-specialized classifier. To handle severe class imbalance, we integrate three complementary strategies: the *Class-Balanced Cross-Entropy* (CB-CE) (Cui et al., 2019), the *Focal Loss* (Cui et al., 2019), and the *Class-Aware Batch Sampling* (CAB) (Henning et al., 2023). With the CB-CE, rather than weighing samples by raw frequency, which can be misleading in long-tailed distributions, we calculate the effective number of samples to capture the diminishing marginal benefit of additional samples:

$$w_y = \frac{1 - \beta}{1 - \beta^{n_y}} \quad (1)$$

where $\beta=0.999$ and n_y is the training count for label y . The loss function becomes:

$$\mathcal{L}_{\text{CB-CE}}(x, y) = -w_y \log p_\theta(y | x) \quad (2)$$

This is combined with the *Focal Loss* objective, which dynamically down-weights easy negatives

and focuses the gradient on hard examples, i.e., those that are misclassified or ambiguous:

$$\mathcal{L}_{\text{CB-Focal}}(x, y) = -w_y \cdot (1 - p_\theta(y | x))^\gamma \cdot \log p_\theta(y | x) \quad (3)$$

where $\gamma=2.0$ controls the down-weighting of high-confidence examples.

We further ensure consistent exposure to under-represented categories through CAB, decoupling batch composition from the dataset’s natural distribution, so that gradient updates are not dominated by majority class signals. Finally, we apply post-hoc threshold calibration via temperature scaling (Guo et al., 2017), so that the model’s output probabilities accurately align with its epistemic uncertainty, preventing it from assigning high confidence to ambiguous cases where it lacks evidence. For binary classification, we apply temperature scaling and tune a decision threshold on the development set; for multi-class classification, we use CB-Focal loss given the more severe imbalance.

3.2 Confidence-Aware Routing

Following the initial classification, we employ a routing mechanism to handle instances where the specialised model exhibits high epistemic uncertainty, i.e., ambiguous examples. To systematically identify these uncertain predictions, we utilize a selective classification rule (El-Yaniv et al., 2010). For an input x , the model produces a predicted label \hat{y}_s with the associated confidence score $c_s(x)$:

$$\hat{y}_s = \arg \max_k p_\theta(y=k | x) \quad (4)$$

$$c_s(x) = \max_k p_\theta(y=k | x) \quad (5)$$

If $c_s(x) \geq \tau_{\text{conf}}$, we accept \hat{y}_s ; otherwise, the input is routed to the CEJ module.

In multi-class settings, reliance on raw confidence scores alone is insufficient because a model may often assign a relatively high probability to the top class while remaining deeply conflicted between two competing categories (e.g., *Derogation* vs. *Animosity*). We thus define the margin between the two most probable labels as $m(x) = p_{(1)}(x) - p_{(2)}(x)$ where $p_{(1)}(x)$ and $p_{(2)}(x)$ denote the largest and second-largest posterior probabilities. An instance is then routed to CEJ if:

$$(c_s(x) < \tau_{\text{conf}}) \wedge (m(x) < \tau_{\text{margin}})^1 \quad (6)$$

3.3 Collaborative Expert Judgment (CEJ)

By acting as a gatekeeper, the aforementioned routing mechanism identifies the instances where statistical pattern matching fails, and ensures that only the subset of difficult cases triggers the accurate but computationally expensive Collaborative Expert Judgment (CEJ) module. This reasoning module is based on prior research on abusive and sexist language which has widely highlighted the importance of combining expert knowledge with lived experience (Sap et al., 2020; Mostafazadeh Davani et al., 2022). We therefore structure CEJ as a multi-perspective reasoning module based on a set of *personas* that act as distinct interpretive lenses. The ‘expert’ personas, such as *linguists*, *psychologists*, *gender studies scholars*, and *legal professionals* are chosen based on the literature showing that they are the main critical perspectives on language use, structural bias, emotional impact, and legal norms (Lazar, 2005; Vidgen and Derczynski, 2019). Yet, to still capture non-expert and experiential perspectives, we incorporate a *sexism victim* persona, reflecting direct experience of harm, and a *layperson* persona, capturing everyday interpretations often responsible for annotator disagreement (Table A3).

It is worth noting that the aim of this multi-persona prompting is not to simulate faithful and exhaustive expert panels, but rather to elicit and organize the knowledge already encoded within LLMs. This combined approach induces the CEJ module to trigger and control the LLMs’ domain understanding that could otherwise remain latent in single-step classification.

3.3.1 Structured Debate and Judge

The CEJ process comprises four sequential prompt-based stages, designed to elicit, interrogate, and

¹Both thresholds are tuned jointly on development data to maximize macro-F1 while minimizing unnecessary escalation.

synthesize domain-relevant knowledge:

Initial Opinions. Each persona independently analyzes the input, providing an initial classification, justification, with a confidence score intended (only) to inform the subsequent discussion. This initial stage establishes diverse baseline perspectives.

Structured Debate. Personas are subsequently exposed to all initial opinions and begging a critical evaluation of their peers’ reasoning. During this deliberation, each persona must: (i) engage with at least one other perspective, agreeing or disagreeing while offering supporting rationale; (ii) revise their own stance if confronted with compelling counter-arguments; and (iii) reassess the input through the interpretive lens of other personas, re-evaluating both intent and target. This process results in a revised classification, an updated justification, and an adjusted confidence score for each persona.

Summarization. Then, a dedicated summarization agent condenses the full debate into a concise synthesis highlighting the main arguments, points of consensus, and unresolved disagreements.

Final Judgment. Finally, the *judge* model synthesizes all available evidence, i.e., the original input, both initial and revised persona opinions, confidence scores, and the debate summary, to produce a final adjudication. Prompted to act impartially, the judge delivers a final classification, comprehensive justification, and a confidence score.

4 Experiments

We proceed describing the datasets, evaluation metrics, and results used to assess the proposed framework.

4.1 Experimental Setup

Datasets. We evaluate our methods on two widely used benchmarks for sexism detection: the EXIST 2025 Tweets dataset (Plaza et al., 2025) and the EDOS dataset (Kirk et al., 2023). The EXIST dataset incorporates annually updated social media posts, capturing emerging sexist language, evolving slang, and annotator disagreements that mirror ambiguities in human judgment. The EDOS dataset offers a complementary perspective with high quality fine-grained annotations at multiple levels of granularity, organized across three hierarchical classification tasks (see Appendix A.1).

Evaluation Metrics. Aligned with the official setup, we use the *Evaluate ALL 2.0* tool (de Educación a Distancia, UNED) for scoring under the EXIST 2025 Task 1.1 (Hard-Hard) configura-

#	Variant	Model	ICM	ICM-Norm	F1_YES	F1
Zero-shot Baselines						
1		GPT-5.2	0.3540	0.6779	0.7001	0.7307
2		QWEN2.5-72B-INSTRUCT	0.3765	0.6892	0.7112	0.7384
3		LLAMA-3.2-3B-INSTRUCT	-0.2009	0.3990	0.4804	0.5522
4		LLAMA-3.3-70B-INSTRUCT	0.3002	0.6509	0.6868	0.7145
5		COGITO-70B	0.3656	0.6837	0.7109	0.7348
6		COGITO-70B (reasoning)	0.3772	0.6896	0.7143	0.7384
EXIST 2025 Task 1.1 Leaderboard						
7		BERT (Nowakowski et al., 2025)	0.5727	0.7878	0.7802	–
8		XLM-RoBERTa (Pan et al., 2025)	0.5799	0.7915	0.7824	–
9		Dual-Transformer Fusion Network (Khan et al., 2025b)	0.5806	0.7918	0.7837	–
10		DeepSeek-R1-Distill-Llama-8B (Villarreal-Haro et al., 2025b)	0.6127	0.8079	0.7945	–
11		Ensemble Approach (Alajmi and Pergola, 2025)	0.6249	0.8141	0.7991	–
12		XLM-RoBERTa (Villarreal-Haro et al., 2025a)	0.6297	0.8165	0.7996	–
13		LLaMA-3.1-8B-Instruct (Lin Tian and Rizoio, 2025)	0.6774	0.8405	0.8167	–
Ours						
14	\mathcal{C}_1	LLAMA-3.2-3B-INSTRUCT-FT	0.5709	0.7913	0.7596	0.7934
15	\mathcal{C}_2	\mathcal{C}_1 + Task-Specific Optimizations	0.6595	0.8315	0.8089	0.8257
16	\mathcal{C}_3	\mathcal{C}_2 + Routing to CEJ $_{Q \rightarrow C}$	<u>0.6831</u>	<u>0.8433</u>	<u>0.8191</u>	<u>0.8348</u>
17	\mathcal{C}_4	\mathcal{C}_2 + Routing to CEJ $_{L \rightarrow C}$	0.6963	0.8500	0.8233	0.8389

Table 1: Results on the EXIST 2025 Task 1.1 test set with zero-shot baselines, leaderboard submissions, and our approach. "FT" denotes the finetuned model, "Q" and "L" denote the models used for personas debate (QWEN2.5-72B-INSTRUCT and LLAMA-3.3-70B, respectively), and "C" refers to the judge model COGITO-70B (reasoning mode). **Bold** = best score, underline = runner-up.

ration. The evaluation metrics include: (i) Inter-Consistency Measure (ICM) (Amigo and Delgado, 2022), which quantifies agreement between predicted and reference annotations by combining intra-system consistency and penalizing overlap; (ii) ICM-Norm, a normalized version of ICM scaled to $[0, 1]$; and the (iii) F1 score. As for the EDOS dataset, we follow the public evaluation protocol, reporting results for the Macro-F1 score, the unweighted mean of the F1-scores computed independently for each class across task A, B, and C (Appendix A.1).

Baselines. We evaluate our proposed framework against both zero-shot and fine-tuned LLM baselines. It worth noting that we focus on state-of-the-art approaches for sexism detection rather than aiming at an exhaustive exploration of all available LLMs, which would be beyond the scope of this work. As specialized models, we employ LLAMA-3.2-3B, adapted via LoRA fine-tuning (Hu et al., 2021) on both the EDOS and EXIST training datasets. Full task-specific training information is available in Appendix A.2. To evaluate instruction-following and reasoning abilities in a zero-shot setting, we test QWEN2.5-72B-INSTRUCT (Qwen et al., 2025), LLAMA-3.3-70B-INSTRUCT (Grattafiori et al., 2024), GPT-

5.2 (OpenAI, 2025), and COGITO-70B (Cogito, 2025) using consistent classification prompts (Appendix A.3). Pre-trained checkpoints were obtained from HuggingFace (Wolf et al., 2020). All experiments were conducted locally through LangChain (Chase, 2022) and Ollama (Team, 2023), on a system equipped with 3 Nvidia A100 GPUs.

4.2 Results

EXIST Dataset. Results for the binary classification task 1.1 of EXIST 2025 are reported in Table 1. The first group of models comprises LLMs evaluated in a zero-shot setting. Among these, COGITO-70B (reasoning mode) achieves the highest F1_YES (0.7143), outperforming by little QWEN2.5-72B-INSTRUCT (0.7112), standard COGITO-70B (0.7109), and GPT-5.2 (0.7043). Overall, all zero-shot models with simple prompts still underperform smaller yet specialised approaches, reflecting still the limitations of LLMs prompted with simple requests. The second group reports public leaderboard systems submitted to EXIST 2025 Task 1.1. Here, LLaMA-3.1-8B-Instruct (Lin Tian and Rizoio, 2025) leads in both ICM and F1_YES. Systems based BERT-variants still shows competitive results compare to larger

#	Variant	Model	Task A	Task B	Task C
Zero-shot Baselines					
1		GPT-5.2	0.7306	0.4543	0.3381
2		QWEN-2.5-72B-INSTRUCT	0.6190	0.4246	0.2576
3		LLAMA-3.2-3B-INSTRUCT	0.4889	0.2442	0.1248
4		LLAMA-3.3-70B-INSTRUCT	0.6421	0.4297	0.2775
5		COGITO-70B	0.6106	0.4214	0.2683
6		COGITO-70B (reasoning)	0.6332	0.4325	0.2826
SemEval 2023 Task 10.A Leaderboard					
7		DeBERTa-v3-large + twHIN-BERT-large (Zhou, 2023)	0.8746	–	–
8		RoBERTa-Large + ELECTRA	0.8740	0.7203	0.5487
9		DeBERTa Ensemble	0.8740	–	–
Data Augmentation and Ensemble Methods					
10		SEFM (Zhong et al., 2023)	0.8538	0.6619	0.4641
11		QCon (Feely et al., 2023)	0.8400	0.6400	0.4700
12		HULAT (Segura-Bedmar, 2023)	0.8298	0.5877	0.4458
13		CSE (Khan et al., 2025a)	0.8819	0.7243	0.5639
14		DDA (Khan et al., 2025a)	0.8769	0.7277	0.6018
15		PaLM Ensemble	–	<u>0.7326</u>	–
Ours					
16	\mathcal{C}_1	LLAMA-3.2-3B-INSTRUCT-FT	0.7824	0.5941	0.3717
17	\mathcal{C}_2	\mathcal{C}_1 + Task-Specific Optimizations	0.8986	0.6279	0.4189
18	\mathcal{C}_3	\mathcal{C}_2 + Routing to CEJ _{Q→C}	<u>0.9195</u>	0.7324	0.5842
19	\mathcal{C}_4	\mathcal{C}_2 + Routing to CEJ _{L→C}	0.9214	0.7421	<u>0.5904</u>

Table 2: Macro-F1 scores on the EDOS datasets with zero-shot baselines, leaderboard references, augmentation/ensemble methods, and our approach. "FT" denotes the finetuned model, "Q" and "L" denote the models used for personas debate (QWEN2.5-72B-INSTRUCT and LLAMA-3.3-70B, respectively), and "C" refers to the judge model COGITO-70B (reasoning mode). **Bold** = best score, underline = runner-up.

architectures with simple prompts.

The bottom of the Table 1 details the results of our routing framework. The initial specialised model, LLAMA-3.2-3B, achieves performance on par with lower-ranking leaderboard models (F1_YES: 0.7596). Applying simple but targeted regularization, indicated by \mathcal{C}_2 (multi-dataset training, class-balanced loss, threshold calibration), yields already substantial improvements: 0.8089 F1_YES and 0.8257 macro-F1, with outperforming several leaderboard submissions. Further gains are realized through the dynamic routing based on confidence calibrations. Configuration \mathcal{C}_3 integrates QWEN2.5-72B-INSTRUCT for persona-based debate, reaching 0.8191 F1_YES, while \mathcal{C}_4 , leveraging LLAMA-3.3-70B for personas and COGITO-70B as judge, achieves the highest scores across all reported metrics. This configuration, achievement state-of-the-art performance on the task, demonstrates the importance of controlling different LLMs to leverage wider capabilities for the task.

EDOS Dataset. For the EDOS dataset, we report in Table 2 the Macro-F1 across the three classification tasks of increasing granularity.

The first group reports the results of LLMs evaluated in a zero-shot setting. Performance re-

mains limited across all tasks when using simple prompts, with GPT-5.2 achieving the highest performance across all tasks, followed by the open-weight LLAMA-3.3-70B-INSTRUCT. COGITO-70B in reasoning mode consistently outperforms its standard variant. However, all zero-shot models particularly struggle with fine-grained classification (Task C). Among official leaderboard systems from SemEval 2023 Task 10.A, ensemble methods based on DeBERTa and RoBERTa architectures are also in this case strong baselines. The next group reports systems employing data augmentation, model ensembling, or both. CSE (Khan et al., 2025a) and DDA (Khan et al., 2025a) achieve notable gains: CSE reaches 0.8819 on Task A and 0.5639 on Task C, motivating the data expansion and leveraging perspectives from multiple models.

Our approach in the base configuration, \mathcal{C}_1 , i.e., fine-tuned LLAMA-3.2-3B, achieves competitive results for binary and multi-class tasks despite its simplicity, but falls short on fine-grained tasks. Introducing targeted training regularization also here, \mathcal{C}_2 , yields again marked improvements, outperforming the zero-shot baselines and most of the specialised systems. Finally, confidence-aware escalation to the CEJ module provides substantial gains. Configuration \mathcal{C}_3 , which employs

Category	n	C_2	C_3	C_4	Gain
<i>Task B</i>					
1. Threats/Harm	89	.681	.848	.857	+17.6
2. Derogation	454	.727	.608	.745	+1.9
3. Animosity	333	.563	.663	.519	+10.1
4. Prejudiced Disc.	94	.541	.811	.848	+30.7
<i>Task C</i>					
Cat. 1.1	16	.217	.300	.250	+8.3
Cat. 1.2	73	.626	.771	.767	+14.5
Cat. 2.1	205	.568	.670	.646	+10.2
Cat. 2.2	192	.537	.571	.561	+3.4
Cat. 2.3	57	.426	.514	.528	+10.2
Cat. 3.1	182	.700	.739	.728	+3.9
Cat. 3.2	119	.556	.735	.724	+18.0
Cat. 3.3	18	.231	.973	.973	+74.2
Cat. 3.4	14	.100	.933	.933	+83.3
Cat. 4.1	21	.408	.069	.188	-22.1
Cat. 4.2	73	.238	.152	.198	-4.1

Table 3: Class-wise F1 on EDOS Tasks B and C. C_2 , C_3 , and C_4 , are defined in Table 2. n is the number of instances for each class. C_2 : domain-tuned model; C_3 : CEJ routing (Qwen); C_4 : CEJ routing (LLaMA). Gain: improvement from C_2 to best CEJ configuration (percentage points).

QWEN2.5-72B-INSTRUCT for persona-based debate, achieves 0.9195 on Task A and 0.7324 on Task B, with the strongest results are obtained by C_4 , where LLaMA-3.3-70B is used for the persona debate and COGITO-70B as the judge. Both configurations outperform all previous systems on binary and multi-class classification on task A and B. On Task C, both routing variants remain competitive, though they struggle compare to the DDA data augmentation method. This gap can be attributed to the cascading effect of baseline limitations: despite employing task-specific optimization techniques, (C_2) exhibits progressive macro-F1 score degradation as class granularity increases (0.8986 \rightarrow 0.6279 \rightarrow 0.4189 across Tasks A, B, and C), indicating that these optimization techniques only partially mitigate the challenge of learning 11 fine-grained categories with severe class imbalances. In contrast, DDA addresses data scarcity directly by synthesizing training examples that populate underrepresented categories, providing a stronger foundation that inference-level refinement alone cannot replicate.

Table 3 presents class-wise F1 scores for Tasks B and C. CEJ shows an inverse relationship between class frequency and effectiveness: rarer categories benefit most, with gains of +30.7 (*Prejudiced Discussions*, $n=94$) and +17.6 (*Threats/Harm*, $n=89$) in Task B. This pattern holds in Task C, where the smallest classes 3.3 ($n=18$) and 3.4 ($n=14$) achieve gains of +74.2 and +83.3 points. This suggests that

Task	Total	Specialist	CEJ	% Escalated
EXIST - Task 1.1	2,076	1,558	518	24.95%
EDOS - Task A	4,000	3,532	468	11.70%
EDOS - Task B	970	578	392	40.41%
EDOS - Task C	970	558	412	42.47%

Table 4: Routing distribution between the specialist and the Collaborative Expert Judgment module (CEJ) across tasks.

multi-perspective debate compensates for representation bias in underrepresented categories.

However, CEJ degrades performance for categories 4.1 (-22.1) and 4.2 (-4.1). These sub-categories, 4.1 (*Supporting mistreatment of individual women*) and 4.2 (*Supporting systemic discrimination*), require subtle distinctions: 4.1 separates victim-blaming from neutral risk statements, while 4.2 separates endorsement of discrimination from good-faith debate about the progress of women’s rights. The domain-tuned model learns these annotation-specific boundaries directly from training data. CEJ, however, introduces multiple interpretive perspectives that may not align with these boundaries, adding variance where the task requires consistency with annotator conventions.

Figure 3 presents per-persona gains after the debate phase. The legal studies persona shows the largest improvement (+13.1% ICM, +4.1% F1), as legal assessments inherently require synthesizing intent, context, and harm dimensions enriched through cross-persona deliberation. Psychologist (+8.7%) and gender studies expert (+7.9%) also benefit substantially.

Table 4 reports how samples are distributed between the specialist and CEJ stages. Across all tasks, the specialist handles the majority of examples directly, with escalation rates of 11.70% and 24.95% on binary detection (EDOS and EXIST Task 1.1) and 40.41% and 42.47% on fine-grained categorization (Tasks B and C). The gap between the two binary tasks is worth noting: EXIST is rebuilt each year with new samples, and its annotation guidelines tend to include more ambiguous, context-dependent cases where annotator disagreement is higher, which naturally pushes more samples below the confidence threshold. The steeper escalation on Tasks B and C is expected, as classifying the specific type of sexism involves greater category overlap and annotation subjectivity than binary detection alone.

Ablation Study Table 5 presents the prompt ablation results for the defined prompts (detailed definitions and examples are presented in Appendix A.4).

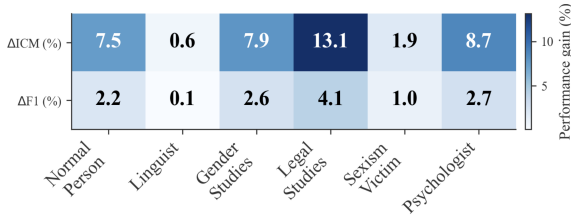


Figure 3: Performance gain after the debate phase.

Persona	\mathcal{P}_1	\mathcal{P}_2	\mathcal{P}_3	\mathcal{P}_4	\mathcal{P}_5
Normal Person	.442	.459	.561	.593	.598
Linguist	.439	.446	.584	.563	.586
Gender Studies Expert	.433	.443	.577	.576	.583
Legal Studies Expert	.466	.453	.558	.566	.573
Sexism Victim	.433	.469	.568	.571	.592
Psychologist	.406	.446	.574	.593	.599

Table 5: Prompt ablation (ICM scores, LLAMA-3.3-70B). \mathcal{P}_1 : baseline few-shot with role identity. \mathcal{P}_2 : \mathcal{P}_1 + specialized expert persona. \mathcal{P}_3 : \mathcal{P}_2 + formal sexism definitions. \mathcal{P}_4 : \mathcal{P}_3 + nuanced multilingual examples. \mathcal{P}_5 : \mathcal{P}_4 + refined guidelines from error analysis.

The full prompt configuration (\mathcal{P}_5) consistently outperforms ablated variants across all personas. The largest degradations occur when removing sexism definitions (\mathcal{P}_2) or persona-specific context (\mathcal{P}_1), with the *Psychologist* exhibiting the greatest sensitivity (+19.3% ICM gain from \mathcal{P}_1 to \mathcal{P}_5). Notably, ICM and ICM-Norm improvements exceed F1 gains, indicating enhanced alignment with human annotation distributions rather than mere accuracy.

4.3 Qualitative Analysis

Figure 4 illustrates successful error correction via confidence-based routing. Here, the specialist misclassifies the input as YES (sexist) with confidence below threshold τ , triggering CEJ escalation. The expert panel classifies it as NO, reasoning that the text targets physical appearance rather than gender, lacks patriarchal power dynamics, and occurs in a casual context. The judge concurs, overriding the false positive. This case demonstrates how routing captures borderline errors: low-confidence predictions are delegated to multi-perspective deliberation, where richer contextual and intent analysis corrects specialist mistakes. Further examples are presented in Appendix A.6.

5 Conclusion

In this work we addressed some of the structural issues in sexism detection, including data scarcity, class imbalance, and annotation noise; factors that

<p>ID: 400341 Tweet: @Amer1can_Barbie Exactly how I feel. I'm gonna be alone for sure, it's Gold Label: NO</p>
<p>Specialist Prediction:</p> <p>Value: YES Confidence: $< \tau$</p>
<p>Debate Summary:</p> <p>All experts maintained NO stance throughout, with slightly increased confidence levels. Consensus & Disagreements: - Strong consensus (NO) across all personas - Minor disagreement on unconscious biases in casual humor (Linguist vs others) Key Arguments: Normal Person: Focuses on appearance rather than gender roles - Legal Studies Expert: No legal threshold of discrimination met - Sexism Victim: Lacks patriarchal power dynamics reinforcement - Linguist: Potential unconscious stereotypes, but edge case - Psychologist: Casual jokes normalize attitudes without direct harm - Gender Studies Expert: Crass humor lacks systemic gender inequality implications</p>
<p>Judge Prediction:</p> <p>Value: NO Justification: The tweet describes a dream scenario where the speaker comments on Bill Gates' appearance using colloquial language. While the language used could be seen as objectifying or disrespectful, the context of a dream and the lack of direct reference to gender roles, stereotypes, or discrimination based on sex suggest that the primary intent is not sexist. The focus is on Bill Gates' appearance rather than his gender or any gender-related issue.</p>

Figure 4: Error correction via confidence-based routing. The specialist's false positive (confidence $< \tau$) is overridden by CEJ deliberation.

often undermine the reliability of standard fine-tuning. We proposed a unified framework that augments targeted training with confidence-calibrated routing and multi-persona deliberation. Through a combination of robust loss functions, balanced batching, and post-hoc calibration, our system improves specialist reliability, while selectively escalating ambiguous cases to a structured, multi-perspective judgment module, CEJ. The CEJ module uses complementary analytical lenses (linguistic, psychological, legal, and sociocultural) to reason over the same input. Evaluations on EXIST 2025 and EDOS show that this approach not only advances state-of-the-art performance, but also proves especially effective on cases marked by ambiguity or underrepresentation.

Limitations

While our frameworks demonstrate strong performance and interpretability through expert-guided multi-persona reasoning, several limitations remain. Our approaches utilize prompt engineering and LLM-generated content, which can reflect biases

present in the models’ pretraining data. Additionally, the scalability of multi-agent interactions poses a practical constraint: as the number of simulated personas increases, so does computational cost, which may hinder applicability in real-time or resource-limited settings.

Moreover, our system’s performance degrades with highly granular classification schemes. The domain-tuned baseline deteriorates with increasing class granularity and severe class imbalances, and this limitation propagates through the routing pipeline. This suggests that reasoning-based refinement is better suited for binary or moderate-granularity tasks, while extreme class fragmentation with heavy imbalances may require complementary data-level interventions such as the DDA method. Importantly, these approaches are not mutually exclusive; CEJ could potentially be combined with data augmentation strategies to leverage both enriched training distributions and structured inference-time reasoning. We leave exploration of such hybrid architectures to future work.

Acknowledgments

This work was supported by a scholarship from the Public Authority for Applied Education and Training (PAAET), Kuwait. This work made use of the Scientific Computing Research Technology Platform (SCRTP) at the University of Warwick for access to the Avon HPC cluster, and the Sulis Tier 2 HPC platform at HPC Midlands+, funded by the Engineering and Physical Sciences Research Council (EPSRC) under grant EP/T022108/1. This project partially supported by the Police STAR Fund 25/26 through the project “STARIST: Stalking Threat AI Recognition (and) Identification Support Tool”, funded by the Office of the Police Chief Scientific Adviser (OPCSA), and in part by the ESRC Digital Good Research Fund through the project “Making Harm Visible: Survivor-Centred Analysis of Domestic Homicide Reviews Using AI”.

References

Gavin Abercrombie, Aiqi Jiang, Poppy Gerrard-abbott, Ioannis Konstas, and Verena Rieser. 2023. Resources for automated identification of online gender-based violence: A systematic review. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, Toronto, Canada. Association for Computational Linguistics.

Anwar Alajmi and Gabriele Pergola. 2025. Leveraging model confidence and diversity: a multi-stage framework for sexism detection. In *Proceedings of*

the Conference and Labs of the Evaluation Forum Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. Springer Lecture Notes.

- Enrique Amigo and Agustín Delgado. 2022. Evaluating extreme hierarchical multi-label classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5819.
- Ziwei Chai, Guoyin Wang, Jing Su, Tianjie Zhang, Xuanwen Huang, Xuwu Wang, Jingjing Xu, Jianbo Yuan, Hongxia Yang, Fei Wu, and 1 others. 2024. [An expert is worth one token: Synergizing multiple expert llms as generalist via expert token routing](#). Preprint, arXiv:2403.16854.
- Harrison Chase. 2022. Langchain: Building applications with llms through composability. <https://www.langchain.com/>. Accessed: 2025-06-02.
- Deep Cogito. 2025. Introducing cogito v2 preview. <https://www.deepcogito.com/research/cogito-v2-preview>. Includes model cogito-v2-preview-llama-70B; accessed 2025-09-24.
- Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. [Class-balanced loss based on effective number of samples](#). In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9260–9269.
- Universidad Nacional de Educación a Distancia (UNED). 2025. Evall 2.0 (evaluateall 2.0). <https://evall.uned.es/en>. Accessed: 30 July, 2025.
- Ran El-Yaniv and 1 others. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5).
- Neil Fasching and Yphtach Lelkes. 2025. [Model-dependent moderation: Inconsistencies in hate speech detection across LLM-based systems](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22271–22285, Vienna, Austria. Association for Computational Linguistics.
- Weston Feely, Prabhakar Gupta, Manas Ranjan Mohanty, Timothy Chon, Tuhin Kundu, Vijit Singh, Sandeep Atluri, Tanya Roosta, Viviane Ghaderi, and Peter Schulam. 2023. [QCon at SemEval-2023 task 10: Data augmentation and model ensembling for detection of online sexism](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1260–1270, Toronto, Canada. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. [The llama 3 herd of models](#). Preprint, arXiv:2407.21783.

- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.
- Shibo Hao, Tianyang Liu, Zhen Wang, and Zhiting Hu. 2023. Toolkengpt: Augmenting frozen language models with massive tools via tool embeddings. *Advances in neural information processing systems*, 36:45870–45894.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. [Deberta: Decoding-enhanced bert with disentangled attention](#). *Preprint*, arXiv:2006.03654.
- Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. 2023. A survey of methods for addressing class imbalance in deep-learning based natural language processing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia. Association for Computational Linguistics.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *CoRR*, abs/2106.09685.
- Aiqi Jiang, Nikolas Vitsakis, Tanvi Dinkar, Gavin Abercrombie, and Ioannis Konstas. 2024. Re-examining sexism and misogyny classification with annotator attitudes. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15103–15125.
- Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178.
- Sahrish Khan, Arshad Jhumka, and Gabriele Pergola. 2025a. [Explaining matters: Leveraging definitions and semantic expansion for sexism detection](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16553–16571, Vienna, Austria. Association for Computational Linguistics.
- Sahrish Khan, Arshad Jhumka, and Gabriele Pergola. 2025b. Multilingual sexism detection through domain adaptation and label-augmented translation: Notebook for the exist lab at clef 2025. In *Proceedings of the Conference and Labs of the Evaluation Forum Information Access Evaluation meets Multilinguality, Multimodality, and Visualization*. Springer Lecture Notes.
- Sahrish Khan, Gabriele Pergola, and Arshad Jhumka. 2024. Multilingual sexism identification via fusion of large language models. In *Conference and Labs of the Evaluation Forum (CLEF 2024)*, volume 3740 of *CEUR Workshop Proceedings*, pages 1070–1079. CEUR-WS.org.
- Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. [SemEval-2023 task 10: Explainable detection of online sexism \(edos\)](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.
- Michelle M. Lazar. 2005. *Feminist critical discourse analysis: Gender, power and ideology in discourse*. Palgrave Macmillan.
- Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Eric Tang, Sumanth Hegde, Kourosh Hakhamaneshi, Shishir G Patil, Matei Zaharia, and 1 others. 2025. [Llms can easily learn to reason from demonstrations structure, not content, is what matters!](#) *Preprint*, arXiv:2502.07374.
- Johanne R. Trippas Lin Tian and Marian-Andrei Rizoiu. 2025. Mario at exist 2025: A simple gateway to effective multilingual sexism detection. In *Proceedings of the Conference and Labs of the Evaluation Forum Information Access Evaluation meets Multilinguality, Multimodality, and Visualization*. Springer Lecture Notes.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *Preprint*, arXiv:1907.11692.
- Do Xuan Long, Duong Ngoc Yen, Anh Tuan Luu, Kenji Kawaguchi, Min-Yen Kan, and Nancy F Chen. 2024. [Multi-expert prompting improves reliability, safety, and usefulness of large language models](#). *Preprint*, arXiv:2411.00492.
- Meng Lu, Brandon Ho, Dennis Ren, and Xuan Wang. 2024. [TriageAgent: Towards better multi-agents collaborations for large language model-based clinical triage](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5747–5764, Miami, Florida, USA. Association for Computational Linguistics.
- Chen Lyu and Gabriele Pergola. 2024. Society of medical simplifiers. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 61–68, Miami, Florida, USA. Association for Computational Linguistics.
- Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. [Dealing with disagreements: Looking beyond the majority vote in subjective annotations](#). *Transactions of the Association for Computational Linguistics*, 10:92–110.
- Nathan Nowakowski, Lorenzo Calogiuri, Előd Egyed-Zsigmond, Diana Nurbakova, Johan Erhani, and Sylvie Calabretto. 2025. Grootwatch at exist 2025: Automatic sexism detection on social networks – classification of tweets and memes: Notebook for the exist lab at clef 2025. In *Proceedings of the Conference and Labs of the Evaluation Forum Information Access Evaluation meets Multilinguality, Multimodality, and Visualization*. Springer Lecture Notes.

- OpenAI. 2025. Introducing GPT-5.2. <https://openai.com/index/introducing-gpt-5-2/>.
- Ronghao Pan, Tomás Bernal-Beltrán, José Antonio García-Díaz, and Rafael Valencia-García. 2025. Umuteam at exist 2025: Multimodal transformer architectures and soft-label learning for sexism detection: Notebook for the exist lab at clef 2025. In *Proceedings of the Conference and Labs of the Evaluation Forum Information Access Evaluation meets Multilinguality, Multimodality, and Visualization*. Springer Lecture Notes.
- Gabriele Pergola, Lin Gui, and Yulan He. 2019. Tdam: A topic-dependent attention model for sentiment analysis. *Information Processing & Management*, 56(6):102084.
- Laura Plaza, Jorge Carrillo de Albornoz, Iván Arcos, Paolo Rosso, Damiano Spina, Enrique Amigó, Julio Gonzalo, and Roser Morante. 2025. Overview of exist 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025)*.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. “call me sexist, but...”: Revisiting sexism detection using psychological scales and adversarial samples. In *Proceedings of the international AAAI conference on web and social media*, volume 15, pages 573–584.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. [Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter](#). *Preprint*, arXiv:1910.01108.
- Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. [Social bias frames: Reasoning about social and power implications of language](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.
- Isabel Segura-Bedmar. 2023. [HULAT at SemEval-2023 task 10: Data augmentation for pre-trained transformers applied to the detection of sexism in social media](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 184–192, Toronto, Canada. Association for Computational Linguistics.
- Chirag Shah. 2025. From prompt engineering to prompt science with humans in the loop. *Communications of the ACM*, 68(6):54–61.
- Xingwei Tan, Chen Lyu, Hafiz Muhammad Umer, Sahrish Khan, Mahathi Parvatham, Lois Arthurs, Simon Cullen, Shelley Wilson, Arshad Jhumka, and Gabriele Pergola. 2025a. SafeSpeech: A comprehensive and interactive tool for analysing sexist and abusive language in conversations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 361–382, Albuquerque, New Mexico. Association for Computational Linguistics.
- Xingwei Tan, Mahathi Parvatham, Chiara Gambi, and Gabriele Pergola. 2025b. InrEx: A dataset for modeling engagement in educational conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21830–21845, Suzhou, China. Association for Computational Linguistics.
- Xingwei Tan, Yuxiang Zhou, Gabriele Pergola, and Yulan He. 2025c. Cascading large language models for salient event graph generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2223–2245, Albuquerque, New Mexico. Association for Computational Linguistics.
- Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C. Alexander, and Nathan Silberman. 2019. [Learning from noisy labels by regularized estimation of annotator confusion](#). In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11236–11245.
- Ollama Team. 2023. Ollama: Run and deploy large language models locally. <https://ollama.com/>. Accessed: 2025-06-02.
- Bertie Vidgen and Leon Derczynski. 2019. [Challenges and frontiers in abusive content detection](#). In *Proceedings of the Third Workshop on Abusive Language Online (ALW3)*, pages 80–93. Association for Computational Linguistics.
- Kapioma Villarreal-Haro, Fernando Sánchez-Vega, and Adrián Pastor López-Monroy. 2025a. Knowledge expansion guided by justification for improved sexism categorization: Notebook for the exist lab at clef 2025. In *Proceedings of the Conference and Labs of the Evaluation Forum Information Access Evaluation meets Multilinguality, Multimodality, and Visualization*. Springer Lecture Notes.
- Kapioma Villarreal-Haro, Judith Tavarez-Rodríguez, Guillermo Segura-Gómez, Alejandro Rosales-Pérez, Fernando Sánchez-Vega, and Adrián Pastor López-Monroy. 2025b. Leveraging reasoning of auto-revealed insights via knowledge injection and evolutionary prompting for sexism analysis: Notebook for the exist lab at clef 2025. In *Proceedings of the Conference and Labs of the Evaluation Forum Information Access Evaluation meets Multilinguality, Multimodality, and Visualization*. Springer Lecture Notes.

- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, and Patrick von Platen. 2020. Hugging face: The ai community building the future. <https://huggingface.co>. Accessed: 2025-06-02.
- Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. [Expertprompting: Instructing large language models to be distinguished experts](#). *Preprint*, arXiv:2305.14688.
- Jingjie Zeng, Liang Yang, Zekun Wang, Yuanyuan Sun, and Hongfei Lin. 2025. [Sheep’s skin, wolf’s deeds: Are LLMs ready for metaphorical implicit hate speech?](#) In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16657–16677, Vienna, Austria. Association for Computational Linguistics.
- Tianyun Zhong, Runhui Song, Xunyuan Liu, Juelin Wang, Boya Wang, and Binyang Li. 2023. [UIRISC at SemEval-2023 task 10: Explainable detection of online sexism by ensembling fine-tuning language models](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2082–2090, Toronto, Canada. Association for Computational Linguistics.
- Mengyuan Zhou. 2023. [PingAnLifeInsurance at SemEval-2023 task 10: Using multi-task learning to better detect online sexism](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2188–2192, Toronto, Canada. Association for Computational Linguistics.

A Appendix

A.1 Datasets Details and Task Structure

The EDOS dataset (SemEval-2023 Task 10) consists of 14,000 training, 2,000 development, and 4,000 test examples for Subtask A (binary sexism detection) including 970 sexist examples for Subtasks B and C. It defines a three-level hierarchical classification problem:

- **Task A:** *Binary classification* — distinguishing between *sexist* and *not sexist* texts.
- **Task B:** *Four-way categorization of sexist content*, assigning each sexist instance to one of the following categories:
 1. Threats, plans to harm and incitement
 2. Derogation
 3. Animosity
 4. Prejudiced discussions
- **Task C:** *Fine-grained subcategorization*, where each sexist instance is further assigned to one of the following subcategories:
 - 1.1 Threats of harm
 - 1.2 Incitement and encouragement of harm
 - 2.1 Descriptive attacks
 - 2.2 Aggressive and emotive attacks
 - 2.3 Dehumanising attacks & overt sexual objectification
 - 3.1 Casual use of gendered slurs, profanities, and insults
 - 3.2 Immutable gender differences and gender stereotypes
 - 3.3 Backhanded gendered compliments
 - 3.4 Condescending explanations or unwelcome advice
 - 4.1 Supporting mistreatment of individual women
 - 4.2 Supporting systemic discrimination against women as a group

For Tasks B and C, we train on filtered subsets where `label_sexist = "Sexist"` from Task A predictions. Each task uses a separate specialist model with task-specific label mappings.

In addition, we evaluate on the EXIST 2025 dataset, which focuses on sexism detection in a cross-lingual, cross-platform setting. It contains 10,034 binary-labeled samples (YES or NO) in English and Spanish. The dataset includes 6,920 training samples (3,260 English, 3,660 Spanish), 1,038

development samples (489 English, 549 Spanish), and 2,076 test samples (978 English, 1,098 Spanish). The task is structured as follows:

- **Task 1.1 (EXIST):** Binary classification of social media posts as *Sexist* vs. *Not Sexist*. This task is analogous to EDOS Task A, but differs in scale, language diversity (English and Spanish), and domain coverage (Twitter and Gab), thereby presenting a more challenging generalization setting.

A.2 Training Details

A.2.1 Shared Training Infrastructure Details (Binary and Multi-Class)

We employ LLAMA-3.2-3B as our base model with 4-bit quantization to enable efficient fine-tuning on consumer hardware while maintaining model quality. The quantization configuration is as follows:

- **Quantization type:** NF4 (Normal Float 4-bit)
- **Double quantization:** Enabled
- **Compute dtype:** bfloat16 (if supported) or float16

LoRA adapters are applied to all attention and feed-forward projection layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`). Configuration varies by task (see Table A1). This reduces trainable parameters to < 1% of the full model while maintaining competitive performance across all tasks.

A.2.2 Binary Classification Optimizations

Class-balanced cross-entropy (CB-CE) is employed to address class imbalance by re-weighting loss contributions based on effective sample counts. For each class y with n_y training examples, the effective number accounts for information overlap:

$$EN_y = \frac{1 - \beta^{n_y}}{1 - \beta}, \quad \text{where } \beta = 0.999 \quad (7)$$

Raw class weights are computed as $w_y^{\text{raw}} = 1/EN_y$. These are normalized to unit mean:

$$w_y^{\text{norm}} = w_y^{\text{raw}} \cdot \frac{C}{\sum_{c=1}^C w_c^{\text{raw}}} \quad (8)$$

To prevent destabilization from extreme weight ratios, we apply clamping:

$$w_y = \text{clip}(w_y^{\text{norm}}, w_{\min}, w_{\max}) \quad (9)$$

with $w_{\min} = 0.25$ and $w_{\max} = 4.0$, followed by re-normalization to unit mean. Hence the loss function is formulated as follows:

$$\mathcal{L}_{\text{CB-CE}}(x, y) = -w_y \sum_{c=1}^C \tilde{y}_c \log \text{softmax}(\mathbf{z}(x))_c \quad (10)$$

Temperature Scaling After training, we calibrate model confidence using temperature scaling to improve probability estimates. We fit a single temperature parameter T by minimizing negative log-likelihood on the development set:

$$T^* = \arg \min_T \sum_{(x,y) \in \mathcal{D}_{\text{dev}}} \left[-\frac{z_y(x)}{T} + \log \sum_{c=1}^C \exp\left(\frac{z_c(x)}{T}\right) \right] \quad (11)$$

Calibrated probabilities are then computed as:

$$p_{\theta}^T(y | x) = \text{softmax}(\mathbf{z}(x)/T^*)_y \quad (12)$$

Temperature scaling is particularly important for binary classification as it corrects systematic over-confidence or under-confidence in model predictions without requiring retraining.

Decision Threshold Tuning Threshold t is tuned on calibrated probabilities to maximize development set macro-F1:

$$t^* = \arg \max_{t \in [0,1]} \text{F1}_{\text{macro}}(\hat{\mathbf{y}}_t), \hat{\mathbf{y}}_t = (\hat{y}_t(x))_{x \in \mathcal{D}_{\text{dev}}} \quad (13)$$

This approach balances precision and recall across both classes, particularly important when minority class performance is critical. In practice, t^* typically falls in the range $[0.3, 0.6]$ depending on class imbalance severity.

A.2.3 Multi-Class Classification Optimizations

For multi-class settings with long-tailed distributions (Tasks B and C), we adopt Class-Balanced Focal Loss, which combines effective-number weighting with focal loss modulation. The CB-Focal loss for a sample (x, y) is:

$$\mathcal{L}_{\text{CB-Focal}}(x, y) = -w_y (1 - p_y)^\gamma \log(p_y) \quad (14)$$

where:

- w_y is the CB weight for class y (computed as in the previous section)
- p_y is the predicted probability for the true class
- γ is the focusing parameter (typically $\gamma \in [1, 3]$, we use $\gamma = 2$)

The $(1 - p_y)^\gamma$ term down-weights easy examples (high p_y) while amplifying hard or misclassified ones (low p_y). This dual mechanism:

- Re-weights by class frequency (via w_y)
- Re-weights by example difficulty (via $(1 - p_y)^\gamma$)

This loss is particularly effective for Tasks B and C, where rare subcategories (e.g., *backhanded compliments*, *condescending explanations*) are overwhelmed by frequent classes (*derogation*). By combining class re-weighting and hard-example mining, CB-Focal improves macro-F1 by 3–5 points, ensuring minority classes are not neglected while reducing over-confidence on majority classes.

Class-Aware Batching. Standard uniform sampling produces batches dominated by majority classes (e.g., 12–14 majority, 2–4 minority instances), leading to inefficient learning on minority classes. Given per-device batch size B and C classes, the quota per class is:

$$k = \lfloor B/C \rfloor \quad (15)$$

For Task B with $C = 4$ classes and $B = 16$, we sample $k = 4$ instances per class per batch. For each training batch:

1. Partition training data into class-specific subsets $\mathcal{D}_1, \dots, \mathcal{D}_C$
2. For each class $c \in \{1, \dots, C\}$, randomly sample k instances from \mathcal{D}_c with replacement
3. Concatenate samples to form batch \mathcal{B} of size $k \cdot C \leq B$
4. Shuffle combined batch to avoid positional biases

This ensures minority classes contribute equally despite smaller pool sizes. Without replacement would exhaust small classes quickly, leading to epoch-boundary effects.

Hyperparameter	Task 1.1/Task A	Task B	Task C
Learning rate	2×10^{-4}	6×10^{-5}	2×10^{-5}
Batch size	16	16	16
Gradient accumulation	2	2	2
Training epochs	5	8	12
Warmup ratio	0.1	0.1	0.1
Weight decay	0.01	0.01	0.01
Max sequence length	512	512	512
Label smoothing (ϵ)	0.05	0.05	0.05
CB-CE β	0.999	0.999	0.999
CB-CE w_{\min}	0.25	0.25	0.25
CB-CE w_{\max}	4.0	4.0	4.0
Focal loss γ	—	2.0	2.0
LoRA rank (r)	32	96	96
LoRA alpha (α)	64	192	192
LoRA dropout	0.1	0.2	0.2

Table A1: Hyperparameter configuration across tasks.

A.2.4 Training Hyperparameters

Table A1 summarizes the hyperparameters selected across tasks.

A.3 Zero-shot Baselines

For the zero-shot baselines, LLMs including QWEN2.5-72B-INSTRUCT, LLAMA-3.3-70B-INSTRUCT, and COGITO-70B were evaluated using task-specific simple prompts. Each model was queried with a zero-shot classification instruction tailored to the corresponding EDOS task as follows:

- **EXIST Task 1.1/ EDOS Task A:** “Classify the following text as sexist or not sexist.”
- **Task B:** “Classify the following text into **one** of the sexism categories: 1) Threats, plans to harm and incitement; 2) Derogation; 3) Animosity; 4) Prejudiced discussions.”
- **Task C:** “Classify the following text into **one** of the sexism subcategories: 1.1) Threats of harm; 1.2) Incitement and encouragement of harm; 2.1) Descriptive attacks; 2.2) Aggressive and emotive attacks; 2.3) Dehumanising attacks & overt sexual objectification; 3.1) Casual use of gendered slurs, profanities, and insults; 3.2) Immutable gender differences and gender stereotypes; 3.3) Backhanded gendered compliments; 3.4) Condescending explanations or unwelcome advice; 4.1) Supporting mistreatment of individual women; 4.2) Supporting systemic discrimination against women as a group.”

A.4 Personas Progressive Prompt Design

Our progressive prompt construction method incrementally integrates role-conditioning, reasoning mechanisms, expert knowledge, and guideline refinements into our prompt design approach. Each iteration aims at improving performance, interpretability, and alignment with human annotator behavior as listed below:

- \mathcal{P}_1 : A basic prompt with few-shot examples and role identity. This sets the task framing and label expectations, testing the model’s generalization capabilities.
- \mathcal{P}_2 : Enhances reasoning by expanding the role identity and background into a specialized expert persona. This stage encourages more structured reasoning by clearly separating the task into sub-steps,
- \mathcal{P}_3 : Integrates formal sexism definition aligned with annotation guidelines.
- \mathcal{P}_4 : Adds nuanced and multilingual examples to strengthen generalization.
- \mathcal{P}_5 : Based on manual error analysis, the prompt incorporates refined guidelines, including considerations of the tweet author’s intent and target audience, as well as the inclusion of edge cases (e.g., profanity and slang) to guide borderline decisions (examples provided in Table A2).

The selected personas are prompted with unique descriptions as shown in Table A3. The six expert personas used in our classification framework are described, each representing a distinct interpretive lens on sexism. These roles were carefully constructed to reflect diverse perspectives, including legal, psychological, linguistic, and lived experience. To increase realism, demographic and educational attributes were heuristically aligned with annotator metadata from the EXIST 2025 dataset, mimicking the diversity and subjectivity found in real-world human annotation.

A.4.1 Preliminary Investigation

Table A4 reports a preliminary investigation conducted on relatively smaller-scale LLMs (7–8B parameters), namely MISTRAL-7B, DOLPHIN3-8B, and COGITO-7B. The table compares the baseline prompt (“Classify the given text as sexist or not sexist”) against our first enhanced persona-based prompt, denoted as \mathcal{P}_1 , evaluated on the EXIST

#	Example
\mathcal{P}_1	You are a psychologist. Classify the tweets as sexist or not sexist. Here are some examples...
\mathcal{P}_2	+ (45 y/o female, Argentina, Ph.D. Psychology) Focus on language, cognitive biases, and the psychological effects of sexism. (1) Analyze the text carefully; (2) Think before responding; (3) Classify the tweet as sexist (1) or not sexist (0).
\mathcal{P}_3	+ Sexism is gender-based prejudice, stereotyping, or discrimination, typically against women. Label as sexist (1) if the tweet: (a) is sexist itself, (b) describes a sexist situation, or (c) criticizes sexist behavior.
\mathcal{P}_4	+ "She got promoted because they needed 'more women in leadership.'" → Sexist (1). "No tengo nada contra las mujeres, pero en cargos altos siempre rinden menos." → Sexist (1).
\mathcal{P}_5	+ (1) Consider the author's intent (insult, joke, venting, shaming). (2) Profanity alone does not indicate sexism — consider the context carefully.

Table A2: Examples illustrating the structure and progression of each prompt stage.

2025 development set. In contrast to the baseline, \mathcal{P}_1 incorporates few-shot examples and an explicit role identity (e.g., psychologist, linguist, normal person) to guide model reasoning.

Overall, the results confirm that even modest prompt enhancements yield consistent improvements across metrics. For MISTRAL-7B and COGITO-7B, \mathcal{P}_1 produced reliable gains in ICM, normalized ICM, and F1, while DOLPHIN3-8B, despite its weak baseline performance, exhibited the largest relative improvements. Importantly, the absolute best scores are consistently achieved by COGITO-7B, suggesting that stronger base models are less sensitive to prompt design but still benefit incrementally. In contrast, weaker models (e.g., DOLPHIN3-8B) depend more heavily on the additional structure introduced by persona framing.

Finally, performance varied across personas: roles such as Normal Person and Legal Studies Expert saw more pronounced gains, while Sexism Victim and Linguist produced only modest improvements. These exploratory findings were not used as final results but rather as guidance for subsequent large-scale experiments, helping us refine prompt design choices and identify promising personas for sexism detection.

A.5 Collaborative Expert Judgment (CEJ) Algorithm and Complexity

Algorithm 1 summarizes the CEJ pipeline for a dataset of N instances and a persona set of size K . CEJ requires K persona inferences, one discussion, one summarization, and one judgment per instance, yielding $N \times (K + 3)$ total LLM calls with complexity $O(NK)$. The cost therefore scales linearly with both the dataset size and the number of personas.²

Algorithm 1 Collaborative Expert Judgment (CEJ)

Input: Dataset $\mathcal{D} = \{(x_i, \text{id}_i)\}_{i=1}^N$, persona set $\mathcal{Y} = \{y_1, \dots, y_K\}$, Large Language Model \mathcal{M}
Output: Opinion matrix $\mathcal{O} = \{O_{i,k}\}$, summary S_i , and final judgment J_i for each x_i

```

1: for each instance  $(x_i, \text{id}_i) \in \mathcal{D}$  do
2:   Initial Persona Classification
3:   for each persona  $y_k \in \mathcal{Y}$  do
4:     Construct persona-specific classification
       prompt  $\mathcal{P}_{\text{cls}}(x_i, y_k)$ 
5:      $O_{i,k} \leftarrow \mathcal{M}.\text{invoke}(\mathcal{P}_{\text{cls}})$ 
6:   end for
7:   Simulated Panel Discussion
8:   Construct discussion prompt
        $\mathcal{P}_{\text{disc}}(x_i, \{O_{i,k}\}_{k=1}^K, \mathcal{Y})$ 
9:    $\mathcal{D}_i \leftarrow \mathcal{M}.\text{invoke}(\mathcal{P}_{\text{disc}})$ 
10:  Summary Generation
11:  Construct summarization prompt
        $\mathcal{P}_{\text{sum}}(\mathcal{D}_i, \mathcal{Y})$ 
12:   $S_i \leftarrow \mathcal{M}.\text{invoke}(\mathcal{P}_{\text{sum}})$ 
13:  Final Judgment with Justification
14:  Construct judgment prompt
        $\mathcal{P}_{\text{judge}}(x_i, \{O_{i,k}\}_{k=1}^K, S_i)$ 
15:   $J_i \leftarrow \mathcal{M}.\text{invoke}(\mathcal{P}_{\text{judge}})$ 
16: end for

```

A.6 CEJ Examples

Figures A1 and A2 illustrate the Collaborative Expert Judgment (CEJ) framework using a representative tweet from EXIST 2025. Figure A1 shows the initial classification stage, where six expert personas independently assess the input. Each provides a binary label (YES/NO), a domain-grounded justification, and a confidence score. Most personas classify the tweet as sexist, citing trivialization of female emotional expression or reinforcement of gendered stereotypes; only the *Legal Expert* disagrees, arguing the statement lacks discriminatory

²The analysis below concerns the interaction-level complexity of the CEJ procedure rather than low-level runtime or token usage.

Persona	Perspective	Role Description (Prompt)
Layperson	General public perspective	An average person (18–22 y/o male, Portugal, B.A.) with no specialized expertise in law, psychology, or linguistics.
Linguist	Language patterns and bias	A linguist (23–45 y/o male, Poland, B.A. Linguistics) specializing in semantics, pragmatics, and discourse analysis, with a focus on gendered language.
Psychologist	Emotional and cognitive effects	A psychologist specializing in language, cognitive biases, and the psychological effects of sexism.
Legal Expert	Legal view on gender-based discrimination	A legal expert (46+ y/o male, Portugal, M.A. Law) specializing in anti-discrimination laws, workplace regulations, and gender equality.
Gender Expert	Structural and ideological analysis	A gender studies expert (46+ y/o female, UK, B.A. Gender Studies) with deep knowledge of gender theories, power dynamics, and social structures.
Sexism Victim	Personal impact of experiencing sexism	A person (18–22 y/o female, South Africa, H.S. diploma) who has personally experienced sexism and understands its emotional and social impact.

Table A3: Persona perspectives and detailed role descriptions.

intent under legal criteria.

Figure A2 presents the structured debate stage, where personas reflect on others’ reasoning and may revise their stance and confidence. This deliberative process demonstrates how LLM-based agents engage in structured argumentation—aligning with or challenging one another based on shared evidence and domain-specific reasoning. While most personas maintain their original YES classification, their justifications become more nuanced and confidence scores are recalibrated. The *Legal Expert* again maintains NO, citing insufficient legal grounds.

B Prompts

This section presents the prompt templates used throughout the CEJ framework. Initial persona judgments are collected using the prompt shown in Figure A3. This is followed by the structured debate prompt (Figure A4), which enables each persona to reflect on peer opinions, revise its stance if persuaded, and recalibrate its confidence. The discussion is then summarized and passed to the judge prompt (Figure A5), which concludes the process by generating a final label along with a justification and calibrated confidence score. All prompts rely on consistent definitions and structured output formats (e.g., JSON objects with fields such as `label`, `justification`, and `confidence`) to support interpretability and downstream evaluation.

Persona	Model	ICM ₁	ICM ₂	ICM%↑	Norm ₁	Norm ₂	Norm%↑	F1 ₁	F1 ₂	F1%↑
Psychologist	MISTRAL-7B	0.2388	0.4059	+69.96%	0.6195	0.7031	+13.49%	0.7433	0.8019	+7.89%
	DOLPHIN3-8B	-0.2299	-0.2299	0.00%	0.3850	0.3850	0.00%	0.5245	0.5245	0.00%
	COGITO-7B	0.4230	0.4703	+11.16%	0.7116	0.7353	+3.33%	0.8052	0.8233	+2.25%
Sexism Victim	MISTRAL-7B	-0.0258	0.3559	+1478.68%	0.4871	0.6781	+39.16%	0.6672	0.7882	+18.13%
	DOLPHIN3-8B	-0.2265	-0.1298	+42.70%	0.3867	0.4351	+12.51%	0.5244	0.5866	+11.86%
	COGITO-7B	0.4089	0.4505	+10.16%	0.7046	0.7254	+2.95%	0.8024	0.8148	+1.54%
Linguist	MISTRAL-7B	0.1882	0.3435	+82.56%	0.5941	0.6718	+13.11%	0.7327	0.7800	+6.44%
	DOLPHIN3-8B	-0.3184	-0.3184	0.00%	0.3407	0.3407	0.00%	0.4724	0.4724	0.00%
	COGITO-7B	0.3967	0.4055	+2.22%	0.6985	0.7029	+0.63%	0.7957	0.8011	+0.68%
Legal Studies Expert	MISTRAL-7B	0.2317	0.3586	+54.75%	0.6159	0.6794	+10.30%	0.7494	0.7858	+4.86%
	DOLPHIN3-8B	-0.2300	0.0427	+118.57%	0.3849	0.5214	+35.45%	0.5261	0.6693	+27.26%
	COGITO-7B	0.4375	0.4772	+9.08%	0.7189	0.7387	+2.75%	0.8117	0.8254	+1.69%
Normal Person	MISTRAL-7B	0.3445	0.3952	+14.72%	0.6723	0.6977	+3.78%	0.7816	0.7985	+2.16%
	DOLPHIN3-8B	-0.2395	-0.0675	+71.82%	0.3802	0.4662	+22.62%	0.5212	0.6213	+19.20%
	COGITO-7B	0.3982	0.4863	+22.11%	0.6992	0.7433	+6.31%	0.7978	0.8287	+3.87%
Gender Studies Expert	MISTRAL-7B	0.2875	0.3956	+37.57%	0.6438	0.6979	+8.41%	0.7590	0.7986	+5.22%
	DOLPHIN3-8B	-0.2493	-0.0769	+69.15%	0.3753	0.4615	+22.99%	0.5188	0.6164	+18.78%
	COGITO-7B	0.3826	0.4339	+13.41%	0.6914	0.7171	+3.72%	0.7930	0.8098	+2.12%

Table A4: Preliminary results of persona-based prompts, comparing baseline prompt with \mathcal{P}_1 . Metrics include ICM, normalized ICM, and F1, with relative improvements (%↑).

<p>Tweet: "Bella can play the victim card all she wants, she won't still win the show. Why didn't she walk out from the toilet? did sheggz tie her legs there? #BBnaja"</p>		
<p>Normal Person</p> <ul style="list-style-type: none"> • Label: YES • Justification: The tweet implies that Bella is using a 'victim card' to gain sympathy, which can be seen as undermining her agency and credibility. • Confidence: 0.75 	<p>Sexism Victim</p> <ul style="list-style-type: none"> • Label: YES • Justification: The tweet dismisses Bella's actions by suggesting she is playing a 'victim card,' which trivializes her experiences and reinforces gendered power dynamics. • Confidence: 0.85 	<p>Psychologist</p> <ul style="list-style-type: none"> • Label: YES • Justification: The tweet uses the term 'victim card' to suggest that Bella is manipulating others, which can be harmful to women who are often accused of being overly emotional or manipulative. • Confidence: 0.82
<p>Linguist</p> <ul style="list-style-type: none"> • Label: YES • Justification: The phrase 'victim card' implies that Bella is exaggerating her situation for personal gain, which can be seen as a gendered stereotype that undermines women's experiences. • Confidence: 0.80 	<p>Legal Studies Expert</p> <ul style="list-style-type: none"> • Label: NO • Justification: The tweet may be offensive, but it does not demonstrate the explicit discriminatory intent or reference to protected categories required to constitute unlawful speech that can be seen as sexist. • Confidence: 0.78 	<p>Gender Studies Expert</p> <ul style="list-style-type: none"> • Label: YES • Justification: The tweet reinforces the stereotype that women often play the 'victim card' to manipulate situations, which is a common form of gendered dismissal and undermines women's credibility. • Confidence: 0.90

Figure A1: An example output from the initial classification stage of the CEJ framework.

<p>Tweet: "Bella can play the victim card all she wants, she won't still win the show. Why didn't she walk out from the toilet? did sheggz tie her legs there? #BBnaija"</p>		
<p>Normal Person</p> <ul style="list-style-type: none"> • Intent: To mock and dismiss Bella's experience • Reaction: Agrees with Sexism Victim because dismissing a woman's concerns as mere "playing" shows insensitivity to gendered power dynamics • Updated Reasoning: The tweet undermines her agency by implying she's faking distress for attention, which feels disrespectful to women's experiences in reality TV settings. • Final Stance: YES (unchanged) • Updated Confidence: 0.80 	<p>Sexism Victim</p> <ul style="list-style-type: none"> • Intent: To shame and silence Bella's expression of distress • Reaction: Agrees with Gender Studies Expert because it reflects common patterns of dismissing women's experiences. • Updated Reasoning: The tweet demonstrates how women are often discouraged from speaking up by being labeled as manipulative or attention-seeking. • Final Stance: YES (unchanged) • Updated Confidence: 0.85 	<p>Psychologist</p> <ul style="list-style-type: none"> • Intent: To invalidate and minimize Bella's feelings • Reaction: Agrees with Normal Person because using "card" implies manipulation rather than genuine emotion. • Updated Reasoning: This language can damage self-esteem by suggesting women's distress is performative or fake, reinforcing harmful stereotypes about female emotional expression. • Final Stance: YES (unchanged) • Updated Confidence: 0.90
<p>Linguist</p> <ul style="list-style-type: none"> • Intent: To delegitimize Bella's experience through rhetorical questions • Reaction: Agrees with Psychologist because the sarcastic questioning creates a dismissive tone. • Updated Reasoning: The rhetorical structure presupposes the situation wasn't serious, undermining Bella's agency and experience. • Final Stance: YES (unchanged) • Updated Confidence: 0.95 	<p>Legal Studies Expert</p> <ul style="list-style-type: none"> • Intent: To express opinion about reality TV behavior • Reaction: Disagrees with Sexism Victim because while it might be dismissive, it doesn't meet legal thresholds for harassment. • Updated Reasoning: The tweet shows poor taste but lacks explicit discriminatory intent required for legal action. • Final Stance: NO (unchanged) • Updated Confidence: 0.65 	<p>Gender Studies Expert</p> <ul style="list-style-type: none"> • Intent: To reinforce patriarchal power structures by dismissing female experience • Reaction: Disagrees with Legal Studies Expert because it exemplifies systemic dismissal of women's concerns. • Updated Reasoning: The tweet perpetuates gendered stereotypes about women being manipulative or attention-seeking, reflecting broader social patterns of invalidating female experiences. • Final Stance: YES (unchanged) • Updated Confidence: 0.98

Figure A2: An example output from the structured debate stage of the CEJ framework.

You are {personas_description} tasked with classifying the following tweet for sexism.

Tweet: {tweet_text}

Your task:

1. Read the guidelines below carefully
2. Analyze the tweet carefully for sexism.
3. Think before responding.
4. Decide the final label: 1 (sexist) or 0 (not sexist).
5. Provide a short justification for their label based on their role.
6. Output a confidence score between 0.0 and 1.0 reflecting your certainty.

Sexism Definition

{definition}

Objective:

{objective}

Here are some examples:

{examples}

Output Example: Provide only a valid JSON object like the following example:

```
{
  "persona": "Normal Person",
  "label": "1",
  "justification": "The tweet stereotypes women's intelligence.",
  "confidence": "0.87"
}
```

Figure A3: Template for initial persona classification in the CEJ framework.

You are continuing the expert panel discussion on the following tweet:

Tweet: {tweet_text}

Initial Opinions: {persona_opinions}

Now, each persona must:

1. Read all other personas' initial opinions.
2. Reflect on whether their own reasoning is still the strongest.
3. Engage with at least one other persona by agreeing or disagreeing with their argument.
4. Update their stance if persuaded, or affirm their original decision.
5. Reassess and adjust their confidence accordingly.

Sexism Definition

{definition}

Objective:

{objective}

Here are some examples:

{examples}

Important Notes:

- Confidence can be increased if supported by solid reasoning, or reduced if uncertainty arises.
- Final answers must state if the stance is changed or unchanged.

Output Example (per persona):

```
{
  "persona": "Sexism Victim",
  "intent": "The author's intent is to shame the woman by dismissing her distress.",
  "reaction": "Agree with Linguist because their interpretation highlights the use of gendered stereotypes.",
  "updated_reasoning": "While my initial view focused on tone, I now realize the tweet uses the 'victim card' trope to discredit women's emotional responses.",
  "final_stance": 1 (changed from 0),
  "updated_confidence": 0.72
}
```

Figure A4: Structured discussion prompt for reasoning and stance revision in the CEJ framework.

You are an impartial judge of the detection of sexist language. Review the following expert panel summary and the original tweet being discussed then classify it for sexism.

You have the following information:

Tweet: {tweet_text}

Discussion summary:

{summary}

Your task:

1. Read the guidelines below carefully
2. Review expert panel summary.
3. Analyze the tweet for sexism.
4. Think before responding.
5. Decide the final label: 1 (sexist) or 0 (not sexist).
6. Provide a brief justification for your decision.
7. State your confidence score (0.0–1.0).

Sexism Definition:

{definition}

Objective:

{objective}

Here are some examples:

{examples}

Output Example: Provide only a valid JSON object like the following:

```
{
  "label": 0,
  "justification": "Although the tweet contains profanity and strong language, it is not
    clearly directed at women. Based on the personas' disagreement and the classification
    guidelines, the tweet is offensive but not sexist.",
  "confidence": 0.79
}
```

Figure A5: Prompt for the judge revision following persona discussion in the CEJ method.