Leveraging Reasoning of Auto-Revealed Insights via Knowledge Injection and Evolutionary Prompting for Sexism Analysis

Notebook for the EXIST Lab at CLEF 2025

Kapioma Villarreal-Haro[1,*,†], Judith Tavarez-Rodríguez[1,*,†], Guillermo Segura-Gómez[1,*,†], Alejandro Rosales-Pérez[3], Fernando Sánchez-Vega[1,2] and Adrián Pastor López-Monroy[1]

[1]*Computer Science Department, Mathematics Research Center (CIMAT), Jalisco S/N Valenciana, 36023, Guanajuato, Guanajuato, México*

[2]*Secretaría de Ciencia, Humanidades, Tecnología e Innovación (Secihti), Av. Insurgentes Sur 1582, Col. Crédito Constructor, 03940, CDMX, México*

[3]*Mathematics Research Center (CIMAT), Monterrey, Av. Alianza Centro 502, Apodaca, 66628, Nuevo León, México.*

Abstract

This paper addresses the CIMAT-CS-NLP team's participation in the three EXIST 2025 text-based tasks (hard settings), consisting of sexism detection, intent analysis, and sexism categorization in tweets. The proposed method is based on a single multitask query to a large language model (LLM), with a prompt that first generates auto-revealed insights for the analysis, and then requests answers for all tasks simultaneously. To automate query refinement, we apply evolutionary computation, optimizing the F1-macro on a development subset. We then utilize the LLM to generate justification texts supporting the categorization obtained with the refined prompt, and use them to fine-tune a multilingual RoBERTa-based model for each task. Our experiments show that the evolved prompt boosts some classification metrics, in comparison with the initial prompt. Experiments with DeepSeek-R1-Distill-Llama-8B and Gemini-1.5-Flash show that (i) explicit reasoning can be induced even in models not originally optimized for it, and (ii) Gemini-1.5-Flash achieves the highest scores in a few-shot scenario, while DeepSeek-R1-Distill-Llama-8B offers a competitive, smaller and open-source alternative. Our findings highlight the advantage of inducing reasoning in an LLM to contextualize tweets and subsequently using them in fine-tuning to analyze, to various degrees, sexism in social media text.

Keywords

Online Sexism, Multilingual Sexism Detection, Sexism Source Intention, Sexism Categorization, LLMs, Reasoning Models, Few-Shot Learning, Prompt Engineering, Evolutionary Computation

1. Introduction

In 1975, the UN proclaimed International Women's Year, placed gender equality on the global agenda, and declared March 8 International Women's Day [1, 2]. Despite all the efforts made over more than half a century, gender inequality remains a problem in society. The persistence of sexist attitudes legitimizes exclusion and violence directed against women, which is reflected in current figures on gender gaps and disparities, where women are clearly the most affected [3].

The expansion of digital media has facilitated the widespread dissemination of sexist messages. Recent United Nations reports state that, in a wide range of countries, between 16 and 58 percent of women and girls have been subjected to online gender-based violence [4]. Unfortunately, practicing digital violence is easier with the use of social media, which transcends all geographical barriers and has the potential for messages to come from anonymous sources. In particular, the social network X

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

*Corresponding author.

†These authors contributed equally.

kapioma.villarreal@cimat.mx (K. Villarreal-Haro); judith.tavarez@cimat.mx (J. Tavarez-Rodríguez); guillermo.segura@cimat.mx (G. Segura-Gómez); alejandro.rosales@cimat.mx (A. Rosales-Pérez); fernando.sanchez@cimat.mx (F. Sánchez-Vega); pastor.lopez@cimat.mx (A. P. López-Monroy)

(formerly Twitter) allows for almost instant virality, while the short format makes it difficult to perceive the full context, which complicates content moderation. This problem raises the need for automated systems capable of detecting, interpreting, and categorizing sexism in tweets with high precision and in real time [5].

Furthermore, traditional binary approaches (sexist/non-sexist) are insufficient to address all the nuances of today's society, so a more in-depth analysis becomes necessary given the complexity of the problem. Detecting the intent of messages, as well as identifying them at a more granular level, helps us understand the problem of sexism on social media, and, based on this understanding, they can lead to real solutions. In this context, several initiatives have emerged to combine efforts within the scientific community and advance toward better detection of sexism, both in binary and finer-grained terms. EXIST (sEXism Identification in Social neTworks) 2025 is the fifth edition of this shared task [6, 7] and its fundamental objective is to make sexism detection a more far-reaching endeavor, attacking sexism from its explicit forms, such as misogyny, to its more subtle expressions, such as micro-violence. This year's edition features several multimodal challenges: text, images, and videos. Each one with three tasks: identification, intent analysis, and fine-grained categorization.
Moreover, recent technologies such as large language models (LLMs) have demonstrated a surprising ability to generalize in classification tasks without requiring fine-tuning for the specific task. However, the performance of an LLM can vary dramatically depending on the wording of the prompt, the selection of examples for few-shot schemes, and the output format of the response text. Identifying the optimal configuration often involves manual trial and error, a process that is unsystematic and difficult to reproduce.

In this work, we address the three text tasks, focusing in the hard-settings, as last year's edition showed that there is still room for improvement, especially in the intention analysis of sexist content and fine-grained categorization tasks. The main tool we use to perform the classifications are specific requests to LLMs, where the prompt was designed to let the model extract relevant insights from the tweet and the context of the tasks before issuing a categorization. With this prompt, we seek to extract the answers to the three text tasks in a single query. Furthermore, the prompt was refined using evolutionary computation, resulting in some improvements in the development set. We subsequently prompt the LLM to extract reasoning texts explaining its own predictions, which are injected in the fine-tuning process of a RoBERTa-based model to perform the final classifications for each task. Our results suggest that using a single multi-task prompt that let the model identify relevant elements of the tweet by itself helps language models better respond to the tasks. Besides, the model's reasoning encoded in justifications to the answer helps in the fine-tuning of RoBERTa-based models. Additionally, evolutionary computation-assisted prompt engineering is a promising avenue for leveraging the power of LLMs and avoiding manual prompt refinement, helping to improve classification metrics.
Our main contributions are:

1. We designed a single prompt that encourages the model to reason deeply about relevant information and insights for the analysis of the tweets and then answer the three tasks in a single query.
2. We used evolutionary computation to evolve the prompt and thus optimize the F1-macro on a development subset, automating what previously relied on manual intuition.
3. We compared the efficiency of the evolved prompt on two models: DeepSeek-R1-Distill-Llama-8B, and Gemini-1.5-Flash, only with few-shot examples within the prompt itself, demonstrating that reasoning can be induced in non-reasoning models, the latter being better in performance.
4. We found that fine-tuning a RoBERTa-based model with rationales that Gemini-1.5-Flash produced from DeepSeek-R1-Distill-Llama-8B answers yields performance on par with using rationales generated entirely by Gemini-1.5-Flash. The key benefit is that the DeepSeek pipeline relies on a smaller, fully open-source model.

2. Related Work

Sexism Detection Since the first edition of the EXIST forum, transformer fine-tuning has been the preferred tactic for classification tasks [8, 9, 10, 11]. Some of the best strategies from past years and strategies found in literature include ensembles of ensembles of monolingual BERT and RoBERTa transformers [12], fine-tuning RoBERTa-based models using information from annotators [13, 14, 15] and fine-tuning of GPT-NeoX and BERTIN-GPT-J-6B models [16] with the latter achieving first place in the hard setting for tasks 1 and 2 in 2023. Thus, it is promising to explore different and more recent LLMs and techniques for sexism identification, analysis of intention and categorization tasks.
Strategies with LLMs were tested in EXIST 2024, were several teams relied on prompt engineering processes to perform zero-shot or few-shot classifications [17, 18, 19]. In fact, in [20], authors generated rationales for each tweet with Meta-Llama-3-8B and used these rationales to fine-tune a Twitter-specialized XLM-RoBERTa, yielding better results than fine-tuning using only the tweets.
In this work, we employed the use of rationales for solving the three tasks in one single query to the model, and subsequently we refined performance through an additional fine-tuning phase on a RoBERTa-based model.

Evolutionary Optimization for Prompt Engineering in Sensitive Tasks LLMs are powerful tools for performing complex classification tasks, as they can generate accurate results without requiring modifications to the model architecture. LLMs leverage extensive pretraining on massive corpora, consisting of billions of words, to model language in unprecedented ways. However, this also induces them to have certain biases or prejudices towards topics that could be considered sensitive. Recent work has demonstrated that LLMs exhibit gender bias [21], and reinforce negative stereotypes [22]. In fact, there is evidence of an amplification of gender bias by LLMs, reflecting and potentially aggravating the societal perceptions from which these models learn [23]. Essentially, they mirror collective societal thinking.

It has been found that LLMs exhibit a significant dependence on the prompt used [24]. By altering the prompt, it is even possible to mitigate some of the biases that LLMs may display. Studies have shown that certain prompting strategies, such as *chain-of-thought*, can reduce the impact of gender bias in these models [25, 26]. For this reason, it becomes crucial to focus on identifying the features that help us better detect gender bias from the prompt itself. Training a language model is highly resource-intensive, and building a model free of bias is even more complex due to the substantial amount of biased data, especially regarding gender, found across the Internet [23].

Efforts in prompt engineering for sexism detection have combined various architectures to improve detection accuracy [27]. Recent studies have opted to use multilingual models like XLM-R or mBERT, paired with language-adapted prompts, while others have explored zero-shot and few-shot techniques alongside different LLM variants. It has been noted that using well-chosen examples, explicit prompts, and structured formats can significantly enhance performance in tasks such as binary classification or intention detection [17, 18]. Nevertheless, these methods often depend heavily on the expertise of the individual designing them, and they generally lack a systematic approach for exploring useful prompt variants. This lack of automation limits their scalability, particularly in multilingual or finely annotated scenarios, such as those found in the EXIST framework.

To address the challenge of prompt optimization, automatic methods based on evolutionary algorithms, such as EvoPrompt [28] and PromptBreeder [29], have recently been proposed. These approaches enable the exploration of effective prompt variants without requiring human intervention. Although they have shown promising results in general tasks, their application in sensitive areas like sexism detection remains limited.

Most research on prompt optimization has been carried out on general tasks, but recent studies have started to focus specifically on sensitive contexts such as sexism. For instance, it has been suggested to enrich prompt statements with semantic definitions and explicit descriptions of the sexist phenomenon, which helps in identifying subtle cases that models often overlook [30]. Another line of work, more geared toward analyzing biases within LLMs themselves, involves the use of *soft prompts*, as in [31],

where specific embeddings are adjusted to assess how models react to sensitive inputs with varying gender cues.

However, these studies often concentrate on qualitative analysis, partial evaluations, or require internal modifications to the models. In contrast, our approach in the first phase preserves the original model and focuses on systematically enhancing prompt formulation, without additional training, by aligning advances in evolutionary optimization with a real and complex issue such as sexism. Moreover, we propose a multitask and scalable strategy that not only performs classification but also induces reasoning, enabling broader generalization to other sensitive contexts.

Reasoning in Generative LLMs While focusing on improving prompt formulation and structure is one way to enhance LLM performance, alternative approaches involve modifying how the models process and respond to those prompts by leveraging their reasoning capabilities. Simple techniques to encourage the models to "reason" before answering improve the already impressive results achieved by Generative LLMs in classification tasks across several domains. For instance, chain-of-thought prompting, which involves adding a few demonstrations that guide reasoning towards a solution, helps models reproduce such reasoning before generating the final answer, ultimately improving performance [32].

While chain-of-thought prompting is useful, it has been shown that explicit demonstrations of how to conduct the reasoning are not necessary, as models are capable of reproducing this behavior by themselves. This is achieved by simply indicating the model to "reason" by adding "let's think step by step" before each answer. Inspired by these ideas, new models have been trained to generate internal chains of thought before producing final answers, with popular models like DeepSeek demonstrating competitive performance against other state-of-the-art approaches [33].

Whether or not models are "really thinking" and despite the limitations that they face [34], the richness and knowledge encoded in their "reasoning" can not be neglected and can be leveraged by other models. There is work where smaller models can benefit from chain-of-thought reasoning by learning in a distillation process to reproduce both the reasoning and final outputs generated by larger models [35]. Other work leverages this knowledge by incorporating reasoning produced by larger models as knowledge expansion of the original input to improve classification metrics [20]. In this approach, although the reasoning is not reproduced, the knowledge embedded in the reasoning is still utilized to improve model performance [20]. Other more direct approaches focus only on using final categorical answers as a means of knowledge expansion, where, despite not incorporating reasoning, it serves as an indirect knowledge transfer coming from larger models [17]. Other approaches do not rely solely on either reasoning or final answer categories but instead incorporate reasoning generated after producing answers to justify the output, leveraging both aspects [36]. This work exemplifies that the reasoning capabilities of LLMs are valuable not only because they lead to more accurate final answers, but also because the knowledge and insights expressed in the process can be leveraged effectively by smaller models in different settings.

3. Methodology

The systems developed to address the three text classification tasks in the hard setting combine generative LLMs with fine-tuning techniques in a multi-stage process. The methodology consists of three main stages: (1) prompt-based classification with evolutionary optimization, (2) justification generation to extract reasoning behind answers, and (3) fine-tuning with knowledge injection.

3.1. Prompt-Based Classification with Auto-Discovery of Relevant Elements

The first stage of the model consists of an auto-discovery and classification approach that lets the model identify relevant elements autonomously. It has been shown that guiding the model by decomposing the task into parts can improve performance in domains like hate speech detection, but it faces the

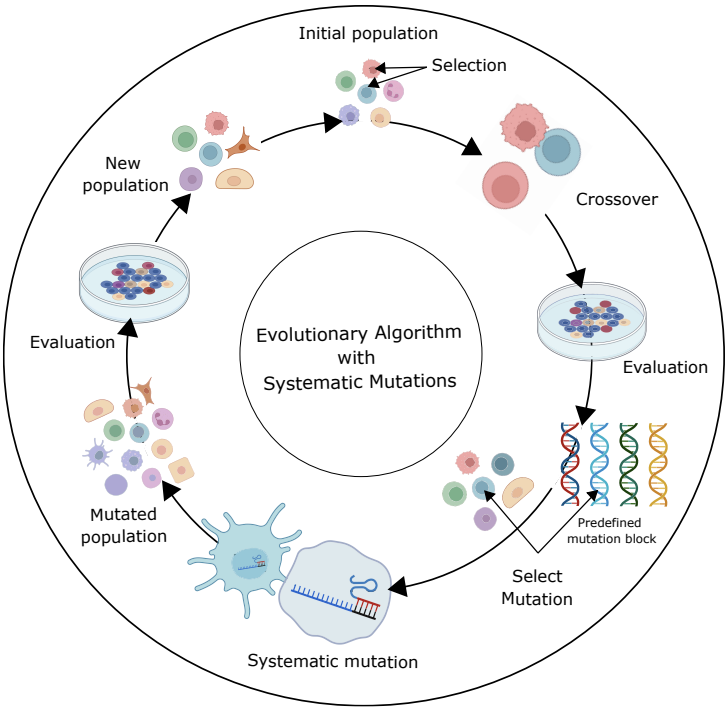

Figure 1: Schematic representation of the genetic algorithm with systematic mutations, inspired by the biological metaphor of evolutionary computation. The process begins with an initial population of prompts and follows a classic evolutionary loop of selection, crossover, and evaluation. Rather than applying random mutations, predefined mutation blocks are injected to introduce controlled, task-aware variations. In each generation, prompts are evaluated and the top-performing variants are retained for the next iteration.

difficulty of carefully selecting its constituents [37]. To avoid the manual definition of components, we aim for the models to discover them on their own.

We define a prompt, where we instruct generative LLMs to self-identify relevant elements that help understand the tweet and are relevant for the tasks, and then to perform classification on the three tasks, all in a single query. The prompt incorporates the task definitions and few-shot examples, which are optimized through an evolutionary process described in the following subsection. The output includes highlighted elements related to the input tweet, and classification predictions for all three tasks.

Performance of the optimized prompt at this stage is evaluated using as core models two different generative LLMs: DeepSeek-R1-Distill-Llama-8B *(AD-Dsk)*, a smaller model optimized for reasoning and Gemini-1.5-Flash *(AD-Gem)*, a larger non-reasoner general-purpose model.

3.1.1. Evolutionary Optimization

The optimization process adopts a classic genetic algorithm framework, incorporating selection, crossover, and mutation phases. The initial population includes both handcrafted prompts and variants generated using GPT-4. To retain the highest-performing prompts across generations, we use tournament and top-k selection strategies. Unlike conventional methods that rely on purely stochastic mutations, we implemented a curated library of structured mutations, specifically designed to improve the efficiency of the prompt in this context. This structured mutation strategy draws inspiration from the concept of *systematic mutation*, recently introduced by Segura-Gómez et al. [38], in which each mutation type is defined as a task-aware transformation intended to enhance prompt performance. In our configuration, each mutation was selected based on empirical trends observed in prompt engineering and validated through ablation studies. The complete evolutionary process, which encapsulates the behavior of systematic mutations, is illustrated in Figure **??**.

According to the proposed approach, we define a single prompt designed to address the reasoning of the language model and solve all three tasks in a single call. This initial prompt or *template* includes for-

matting instructions, text generation, and most importantly, specific instructions for each subtask. While this complete prompt is used as-is in the final system, during the optimization phase, it is segmented by subtask, with each section optimized independently. This strategy enables the evolutionary process to focus on a specific task at a time, thereby enhancing both its efficiency and stability. Additionally, it allows for fine-grained control over the modifications applied to each functional block of the prompt.

Optimization Setup The evolutionary process was performed on a subset of 500 randomly selected examples from the dev set. This subset is based on the collection of tweets originally introduced in EXIST 2023 and follows the "Learning With Disagreement" paradigm, in which each tweet is annotated by six distinct individuals [7]. These annotators represent a diverse range of perspectives, differing in gender, age range, ethnicity, education level, and country of origin.

The selected subset includes tweets in both English and Spanish. Each tweet is annotated with both consensus labels, computed via majority vote, and the full distribution of individual annotator labels. For the purposes of the evolutionary process, only the consensus labels were used. This choice aligns with the requirements of the fitness function, which operates on predefined classes: a single target label in binary classification, or a reduced set of categories in multiclass tasks. This subset was used consistently throughout the evolutionary cycle to evaluate the effectiveness of the generated prompts.

The procedure was executed over 10 generations with a population of 10 prompts. In each generation, the best-performing candidates were selected via tournament selection, combined through crossover, and modified using a structured mutation. Prompts were evaluated on the dev set using the Meta-Llama-3.1-8B model to generate outputs, which were then scored using macro F1 as the fitness function. The process prioritized interpretability and stability over aggressive exploration, aiming to generate prompt variants that remained robust across different tasks and could generalize well when integrated into the full pipeline. A detailed description of the prompt evolution system and intermediate prompt versions is provided in Appendix A.

3.2. Process justifications before fine tuning

To leverage the model's existing knowledge in the fine-tuning step, we further prompt a generative LLM to justify the answers to the tasks generated during the Prompt-Based Classification. Thereby, we obtain justification texts that explain the answers provided, showcasing the internal reasoning of the models in a posterior process. We use Gemini-1.5-Flash to generate justification for both model outputs in the previous steps.

Gemini justifications over *AD-Gem* answers serve as a way to extract reasoning from Gemini's own classification. While we can extract reasoning directly from DeepSeek, this approach presented several challenges: consistency in the answers was not guaranteed, labels were not always present, and early development results were unfavorable. Instead, we extract Gemini justifications over *AD-DSk* serve to standardize answers in DeepSeek while also incorporating reasoning of Gemini explaining DeepSeek's classification.

These justifications will serve as knowledge injection to expand tweets in the next classification stage.

3.3. Justification Enhanced (JE) Fine Tuning

The final stage consists of a fine-tuning process with knowledge injection obtained from the justifications obtained from the previous step, following the approach presented in [36]. We focus on an XLM-RoBERTa model trained on multilingual tweets (T-XLM-RoBERTa) [39].

The knowledge injection process is implemented by concatenating justifications with the original tweets, creating a more informative representation for classification. The base *T-XLM-RoBERTa* model is then fine-tuned separately to predict each of the three tasks using the enriched input. A final correction process is applied to tasks 1.2 and 1.3 based on the answers of task 1.1: if the answer to 1.1 was negative, we correct tasks 1.2 and 1.3 answers to reflect this dependency.

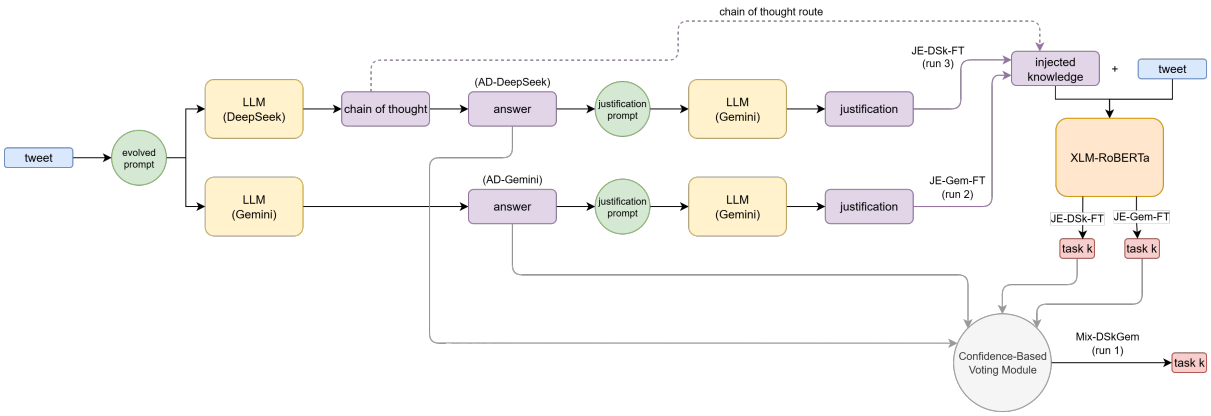

Figure 2: Overall methodology: we first evaluate the base prompt, extracting answers and then justifications using generative LLMs. Justifications are then injected into a classification model, which is fine-tuned. We also aggregate answers from different models via confidence-based voting.

These fine-tuned models are identified as *(JE-DSk-FT)* and *(JE-Gem-FT)*, depending on the model used to generate the answers that serve as guide to the justifications (*AD-DSk* and *AD-Gem* respectively).

Figure 2 outlines the overall methodology for individual systems, including the alternative tested in the early stages.

3.4. Summary of the submitted models

Models submitted correspond to the individual fine-tuned versions *JE-DSk-FT* and *JE-Gem-FT*. A third model is obtained by combining multiple models, expecting to obtain a more robust panorama of the answers. The models considered are those obtained in the Prompt-Based Classification with Auto-Discovery of Relevant Elements (*AD-DSk* and *AD-Gem*), as well as the ones obtained in Justification Enhanced (JE) Fine Tuning (*JE-DSk-FT* and *JE-Gem-FT*).

The ensemble model, identified as *MixDSkGem*, is produced by assigning a confidence score according to the performance of each model in the dev set. Final answers are then weighted using these confidence scores to finally output the answer with the highest confidence. For task 1.1 (binary classification) and task 1.2 (multiclass classification), the highest-scored label is submitted, while for task 1.3 (multi-label classification), we treated each label as a binary problem to decide if it is included in the labels submitted.

Table 1 contains a summary of the systems and the corresponding runs in the official leaderboard.

Table 1
Summary of evaluated systems and their submission number. The ensemble corresponds to the weighted sum of responses from the systems AD-DSk, AD-Gem, JE-DSk-FT, and JE-Gem-FT for all three text tasks in a hard setting.

System	Submitted Run	System Type
AD-DSk	–	Prompting only
AD-Gem	–	Prompting only
JE-DSk-FT	Run 3	Fine-tuned
JE-Gem-FT	Run 2	Fine-tuned
Mix-DSkGem	Run 1	Ensemble (Prompting + Fine-tuned)

4. Results and Discussion

4.1. Results on dev

During the prompt evolution phase, we evaluated candidate prompts using Meta-Llama-3.1-8B as the inference model and macro F1 score as the fitness function. Table 2 summarizes the results obtained

using the best evolved prompt versus the original base prompt across the three tasks on a 500-example subset of the development data.

Table 2
Macro F1 Performance Across Tasks During Prompt Evolution with LLaMA 3.1 8B on a 500-example Dev Subset

Generation	Task 1.1	Task 1.2	Task 1.3
Gen 0	0.8098	0.3109	0.0381
Gen 4	0.8250	0.4350	0.1650
Gen 6	0.8390	0.4850	0.1960
Gen 8	0.8470	0.5350	0.2250
Gen 10	0.8478	0.5724	0.2477

The evolved prompt led to consistent improvements across all tasks, with the most notable gain observed in Task 1.2, which benefited from clearer intention phrasing. These results validate the effectiveness of our systematic mutation strategy under a controlled evaluation setup. While these scores are not directly comparable to the ICM-Hard-Norm metrics reported for Gemini-based evaluations, they were essential for guiding the prompt optimization process and served as a foundation for the final prompt adopted in downstream components of the system. A detailed description of the prompt evolution process and intermediate versions is provided in Appendix A.

Final prompt evolution performance was evaluated using the complete dev dataset with Gemini-1.5-Flash. Seed prompts performance show improvement from ranges of ICM-Norm from $[0.71 - 0.73]$, $[0.54 - 0.57]$, and $[0.48 - 0.49]$ in tasks 1.1, 1.2, and 1.3 respectively to $0.74, 0.55$, and 0.50 compared to the final prompt version. We observe an overall improvement of the final prompt in tasks 1.1 and 1.3 compared to all seed prompts, while task 1.2 shows minimal improvement but still outperforms the worst-performing seeds.

Table 3 shows the performance of auto-discovery prompting-based models and fine-tuned ones on the dev dataset. From these results, we observed that fine-tuning the T-XLM-RoBERTa model with justifications generated by LLMs significantly improves classification performance across all tasks, compared to using only the classifications produced directly through prompting. Ensemble model based on confidence voting *Mix-DSkGem* improves performances compared to individual models in tasks 1.1 and 1.2, and experiments a loss in task 1.3 compared with the fine-tuned models, suggesting the merging process for this task is not optimal.

Furthermore, in the literature [17], we found that prompting with Gemini-1.0-pro without Auto Discovery in the prompt (class_definitions_refined_prompt), achieved an ICM-Hard-Norm metric of 0.7318 in the dev set for task 1, and fine-tuning a T-XLM-RoBERTa model with only the tweets achieved a 0.7940 ICM-Hard-Norm metric for task 1. This tells us that our system has the potential to obtain competitive metrics in the test, specially the JE-Gem-FT model, with the advantage that the three tasks can be obtained directly from a single request to the model, which helps reduce latency.

Table 3
ICM-Hard-Norm on dev for each task

System	Task 1.1	Task 1.2	Task 1.3
AD-DSk	0.61	0.37	0.35
AD-Gem	0.74	0.55	0.50
JE-DSk-FT	0.79	0.61	0.58
JE-Gem-FT	0.81	0.64	0.59
Mix-DSkGem	0.82	0.64	0.54
class_definitions_refined_prompt	0.73	-	-
FT T-XLM-RoBERTa	0.79	-	-

4.2. Results on test

Tables 4 and 5 show the results obtained by our systems in the test set, for all tweets and divided by language, respectively; and its position in the EXIST 2025 leader-board. Table 4 also presents the performance of systems submitted to EXIST 2024 (N-LLM-R-Stack-Ra [20] and Resp_aware_in + FT XLM-RoBERTa [17]), which are closely related to our proposed systems.

The majority of our submissions ranked in the top ten of all the evaluations, specially those for tasks 1 and 2. Our best submitted system ranked 4th for task 2 and it consists of the T-XLM-RoBERTa model which was fine-tuned with the tweets plus the justifications generated by the Gemini-1.5-Flash model. It is worth to notice that the ranking order achieved in the dev set is not preserved in the test set for task 1, since the JE-Gem-FT system obtained a lower ICM-Hard metric. Still, the difference is only 0.0005 points for the F1 metric, and for task 2, the difference is only 0.0021. This tells us that **the fine-tuning process helps an 8B parameter model like DeepSeek contribute similarly to a larger model like Gemini-1.5-Flash, but with the advantage of being open source**.

Regarding the results of N-LLM-R-Stack-Ra and Resp_aware_in + FT XLM-RoBERTa, these systems correspond to fine-tuned XLM-RoBERTa models with tweets + generated-text-from-LLMs.
N-LLM-R-Stack-Ra consisted in asking to a Meta-Llama-3-8B-Instruct model to generate analysis that supported the idea that the tweet was not related to sexism (negative reasoning). This analysis was concatenated along with the tweet itself and then used as input for fine-tuning a T-XLM-RoBERTa model. The Resp_aware_in + FT XLM-RoBERTa system consisted in prompting a Gemini-1.0-pro model to classify the tweet as sexist or not. Then, the YES or NO answer was concatenated along with the tweet and a XLM-RoBERTa model was fine-tuned with this new input.
As we can notice, **incorporating justifications into the fine-tuning inputs significantly improves classification performance, more than adding negative reasoning or adding the LLM classification directly to the tweet**. The previous demonstrates that the RoBERTa-based model benefits from the contextual information extracted during LLM Auto Discovery prompting and it contributes to more accurate detection of sexist tweets and a better identification of intent.

Table 4
System performance on the test set - Language: ALL

Task	System	ICM-Hard	ICM-Hard Norm	F1	Ranking
1.1	JE-DSk-FT	0.6127	0.8079	0.7945	5
	JE-Gem-FT	0.6076	0.8054	0.7940	6
	Mix-DSkGem	0.5246	0.7637	0.7652	35
1.2	JE-Gem-FT	0.4264	0.6386	0.5461	4
	JE-DSk-FT	0.4118	0.6339	0.5482	5
	Mix-DSkGem	0.3619	0.6177	0.5266	8
1.3	JE-Gem-FT	0.3980	0.5924	0.6125	8
	Mix-DSkGem	0.3353	0.5779	0.6039	14
	JE-DSk-FT	0.1310	0.5304	0.5861	23
1.1	N-LLM-R-Stack-Ra	0.5407	0.7718	0.7694	-
1.1	Resp_aware_in + FT XLM-RoBERTa	0.5486	0.7757	0.7746	-
1.2	Resp_aware_in + FT T-XLM-RoBERTa	0.2643	0.5859	0.5171	-

Table 5
System Performance on the test set - Languages: ES (Spanish) and EN (English)

Task	System	ICM-Hard	ICM-Hard Norm	F1
Spanish (ES)				
1.1	JE-Gem-FT	0.6013	0.8007	0.8044
	JE-DSk-FT	0.6012	0.8006	0.8052
	Mix-DSkGem	0.5060	0.7530	0.7800
1.2	JE-Gem-FT	0.5256	0.6642	0.5852
	JE-DSk-FT	0.5161	0.6612	0.5904
	Mix-DSkGem	0.4623	0.6444	0.5752
1.3	JE-Gem-FT	0.4710	0.6052	0.6346
	Mix-DSkGem	0.4079	0.5911	0.6227
	JE-DSk-FT	0.2389	0.5534	0.6063
English (EN)				
1.1	JE-DSk-FT	0.6143	0.8135	0.7802
	JE-Gem-FT	0.6018	0.8071	0.7800
	Mix-DSkGem	0.5312	0.7711	0.7455
1.2	JE-Gem-FT	0.2894	0.6001	0.4896
	JE-DSk-FT	0.2685	0.5929	0.4860
	Mix-DSkGem	0.2277	0.5788	0.4587
1.3	JE-Gem-FT	0.2961	0.5726	0.5805
	Mix-DSkGem	0.2344	0.5575	0.5744
	JE-DSk-FT	-0.0095	0.4977	0.5604

5. Conclusions and Limitations

Our results confirm that incorporating LLM-generated justifications during fine-tuning boosts the performance of our systems to competitive positions on the EXIST 2025 leader-board. Moreover, the prompting method yields to a framework for solving the three task in a single query to the LLM, leading to better metrics when coupled with the fine-tuning process.

The use of evolutionary computation proved promising for optimizing the prompt. The final evaluation metrics did not change significantly, but we hypothesized that using this technique with smaller prompts may be more beneficial in the prompt engineering process.

The experiments show that incorporating smaller open-source models, such as DeepSeek-R1-Distill-Llama-8B, can achieve results similar to much larger models like Gemini-1.5-Flash for the process of enrichment with reasoning and explanations.

Validation of the stability of the observed improvements remains to be explored, as the slight variation between the dev and test rankings suggests that our models show room for improvement in terms of generalization.

Acknowledgments

Villarreal-Haro acknowledges *Secretaría de Ciencia, Humanidades, Tecnología e Innovaciones* (SECIHTI) for its support provided by the program *Becas Nacionales Para Estudios de Posgrados* (CVU 1309535). Segura-Gómez acknowledges financial support from SECIHTI through the graduate scholarship provided by the program *Becas Nacionales Para Estudios de Posgrados* number 4006758 and CVU 1308651. Tavarez-Rodríguez acknowledges SECIHTI and *Centro de Investigación en Matemáticas* (CIMAT) for the support through the PhD scholarship (CVU 859147). Sanchez-Vega acknowledges SECIHTI for its support

through the program *"Investigadoras e Investigadores por México"* (Project ID.11989, No.1311). Rosales-Pérez acknowledges SECIHTI for its support through the grant project "Búsqueda de arquitecturas neuronales eficientes y efectivas" (CBF2023-2024-2797). The authors gratefully acknowledge SECIHTI and CIMAT for the computing resources provided by the CIMAT Bajio Supercomputing Laboratory (#300832), and the support of Google Cloud Platform for the credits granted under its trial program.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly in order to: Drafting content, Text Translation, Grammar and spelling checking, Citation management. After using these tool(s)/service(s), the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] United Nations General Assembly, Resolution 3010 (xxvii): International women's year, https://undocs.org/en/A/RES/3010(XXVII), 1972. Proclaimed 1975 as International Women's Year and called for the first World Conference on Women.

[2] United Nations, Background – international women's day, 2025. URL: https://www.un.org/en/observances/womens-day/background, states that the UN began celebrating International Women's Day on 8 March during International Women's Year (1975).

[3] World Economic Forum, Global gender gap report 2024, 2024. URL: https://www.weforum.org/reports/global-gender-gap-report-2024, the report shows that women continue to be the most affected by disparities in economic participation, political representation, health, and education, highlighting how sexist attitudes legitimize exclusion and gender-based violence.

[4] UN Women, Creating safe digital spaces free of trolls, doxing and hate speech, 2023. URL: https://www.unwomen.org/en/news-stories/explainer/2023/11/creating-safe-digital-spaces-free-of-trolls-doxing-and-hate-speech, explainer article, UN Women Headquarters.

[5] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, Automatic classification of sexism in social networks: An empirical study on twitter data, IEEE Access 8 (2020) 219563–219576. doi:10.1109/ACCESS.2020.3042604.

[6] 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.

[7] 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: CLEF 2025 Working Notes, 2025.

[8] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: sexism identification in social networks, Procesamiento del Lenguaje Natural 67 (2021) 195–207.

[9] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2022: sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022) 229–240.

[10] L. Plaza, J. C. de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023-learning with disagreement for sexism identification and characterization (extended overview)., CLEF (Working Notes) (2023) 813–854.

[11] L. Plaza, J. C. de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante,

D. Spina, Overview of exist 2024–learning with disagreement for sexism identification and characterization in tweets and memes (extended overview), in: Conference and Labs of the Evaluation Forum, 2024.
[12] E. Villa-Cueva, F. Sanchez-Vega, A. P. López-Monroy, Bi-ensembles of transformer for online bilingual sexism detection., in: IberLEF@ SEPLN, 2022.
[13] Y.-Z. Fang, L.-H. Lee, J.-D. Huang, Nycu-nlp at exist 2024–leveraging transformers with diverse annotations for sexism identification in social networks, Working Notes of CLEF (2024).
[14] M. P. Jimenez-Martinez, I. H. Lopez-Nava, M. Montes-y Gómez, An analysis of the impact of gender and age on perceiving and identifying sexist posts, in: E. Mezura-Montes, H. G. Acosta-Mesa, J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad, J. A. Olvera-López (Eds.), Pattern Recognition, Springer Nature Switzerland, Cham, 2024, pp. 308–318.
[15] M. P. Jimenez-Martinez, I. H. Lopez-Nava, M. Montes-y Gómez, Enhancing the detection of sexist messages through a multi-profile-based ensemble approach, Computación y Sistemas 29 (2025) 283–294.
[16] L. Tian, N. Huang, X. Zhang, Efficient multilingual sexism detection via large language model cascades., in: CLEF (Working Notes), 2023, pp. 1083–1090.
[17] J. Tavarez-Rodríguez, F. Sánchez-Vega, A. Rosales-Pérez, A. P. López-Monroy, Better together: Llm and neural classification transformers to detect sexism, Working Notes of CLEF (2024).
[18] A. Azadi, B. Ansari, S. Zamani, S. Eetemadi, Bilingual sexism classification: fine-tuned xlm-roberta and gpt-3.5 few-shot learning, arXiv preprint arXiv:2406.07287 (2024).
[19] S. Khan, G. Pergola, A. Jhumka, Multilingual sexism identification via fusion of large language models, Working Notes of CLEF (2024).
[20] K. Villarreal-Haro, F. Sánchez-Vega, A. Rosales-Pérez, A. P. López-Monroy, Stacked reflective reasoning in large neural language models, Working Notes of CLEF (2024).
[21] S. Soundararajan, S. J. Delany, Investigating gender bias in large language models through text generation, in: Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024), 2024, pp. 410–424.
[22] R. Ostrow, A. Lopez, Llms reproduce stereotypes of sexual and gender minorities, arXiv preprint arXiv:2501.05926 (2025).
[23] H. Kotek, R. Dockum, D. Sun, Gender bias and stereotypes in large language models, in: Proceedings of the ACM collective intelligence conference, 2023, pp. 12–24.
[24] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, A. Chadha, A systematic survey of prompt engineering in large language models: Techniques and applications, arXiv preprint arXiv:2402.07927 (2024).
[25] A. Sant, C. Escolano, A. Mash, F. D. L. Fornaciari, M. Melero, The power of prompts: Evaluating and mitigating gender bias in mt with llms, arXiv preprint arXiv:2407.18786 (2024).
[26] S. Dwivedi, S. Ghosh, S. Dwivedi, Breaking the bias: Gender fairness in llms using prompt engineering and in-context learning, Rupkatha Journal on Interdisciplinary Studies in Humanities 15 (2023).
[27] M. Siino, I. Tinnirello, Prompt engineering for identifying sexism using gpt mistral 7b, Working Notes of CLEF (2024).
[28] Q. Guo, R. Wang, J. Guo, B. Li, K. Song, X. Tan, G. Liu, J. Bian, Y. Yang, Connecting large language models with evolutionary algorithms yields powerful prompt optimizers, arXiv preprint arXiv:2309.08532 (2023).
[29] C. Fernando, D. Banarse, H. Michalewski, S. Osindero, T. Rocktäschel, Promptbreeder: Self-referential self-improvement via prompt evolution, arXiv preprint arXiv:2309.16797 (2023).
[30] S. Khan, A. Jhumka, G. Pergola, Explaining matters: Leveraging definitions and semantic expansion for sexism detection, arXiv preprint arXiv:2506.06238 (2025).
[31] J.-J. Tian, D. Emerson, S. Z. Miyandoab, D. Pandya, L. Seyyed-Kalantari, F. K. Khattak, Soft-prompt tuning for large language models to evaluate bias, arXiv preprint arXiv:2306.04735 (2023).
[32] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Proceedings of the 36th

International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2022.

[33] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al., Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, arXiv preprint arXiv:2501.12948 (2025).

[34] P. Shojaee, I. Mirzadeh, K. Alizadeh, M. Horton, S. Bengio, M. Farajtabar, The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity, arXiv preprint arXiv:2506.06941 (2025). URL: https://arxiv.org/abs/2506.06941, submitted 7 June 2025.

[35] L. H. Li, J. Hessel, Y. Yu, X. Ren, K.-W. Chang, Y. Choi, Symbolic chain-of-thought distillation: Small models can also" think" step-by-step, arXiv preprint arXiv:2306.14050 (2023).

[36] K. Villarreal-Haro, F. Sánchez-Vega, A. P. López-Monroy, Knowledge expansion guided by justification for improved sexism categorization, Working Notes of CLEF (2025).

[37] B. AlKhamissi, F. Ladhak, S. Iyer, V. Stoyanov, Z. Kozareva, X. Li, P. Fung, L. Mathias, A. Celikyilmaz, M. Diab, ToKen: Task decomposition and knowledge infusion for few-shot hate speech detection, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 2109–2120. URL: https://aclanthology.org/2022.emnlp-main.136/. doi:10.18653/v1/2022.emnlp-main.136.

[38] G. Segura-Gómez, A. P. Lopez-Monroy, F. Sanchez-Vega, A. Rosales-Pérez, Nlp-cimat at semeval 2025 task 11: Prompt optimization for llms via genetic algorithms and systematic mutation applied on emotion detection, in: Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Association for Computational Linguistics, Vienna, Austria, 2025. To appear.

[39] F. Barbieri, L. Espinosa Anke, J. Camacho-Collados, XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 258–266. URL: https://aclanthology.org/2022.lrec-1.27.

[40] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38–45.

[41] T. Dettmers, M. Lewis, Y. Belkada, L. Zettlemoyer, Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale, Advances in neural information processing systems 35 (2022) 30318–30332.

[42] E. Frantar, S. Ashkboos, T. Hoefler, D. Alistarh, Gptq: Accurate post-training quantization for generative pre-trained transformers, arXiv preprint arXiv:2210.17323 (2022).

[43] C. Centro de Investigación en Matemáticas A.C, Laboratorio de supercómputo del bajío, n.d. URL: https://supercomputobajio.cimat.mx/, accessed on June 15, 2025.

A. Appendix: Evaluated Prompts

This appendix presents the top-performing prompt variations at different stages of the evolutionary process. Each version includes the full phrasing used for the three EXIST 2025 subtasks.

All evaluations, prompt generations, and structured mutations were performed using *LLaMA 3.1 8B* in an 8-bit quantized configuration. The implementation leveraged `PyTorch`, the `transformers` library [40], and `bitsandbytes` [41, 42] for efficient low-precision inference. All experiments were executed on two NVIDIA Titan RTX GPUs (24 GB DDR6), provided by the Supercomputing Laboratory at CIMAT Bajío [43].

Due to computational constraints, the evolutionary search was limited to 10 generations with a population size of 10. The entire optimization process was inference-based, with no gradient updates or additional training. Below, we present the best prompt per generation, showing how task-specific instructions evolved through systematic mutation.

Prompt - Gen 0 (Seed)

Macro F1: T1.1 = 0.8098, T1.2 = 0.3109, T1.3 = 0.0381

> **TASK 1:** Is there any gender-based assumption or judgment in this message?
> **TASK 2:** Identify the intention behind the tweet: is it a direct sexist statement, a report of sexism, or a judgement against sexism?
> **TASK 3:** How would you classify the kind of sexism this tweet reflects — is it about stereotypes, objectification, or maybe something more violent?

Prompt - Gen 4

Macro F1: T1.1 = 0.8250, T1.2 = 0.4350, T1.3 = 0.1650

> **TASK 1:** Evaluate the given tweet for the presence of sexism, categorizing it as 'Direct Sexism' if sexism is directly expressed or stated, 'Described Sexism' if sexism is described or referenced, 'Criticized Sexism' if sexism is being criticized or challenged, or 'NO' if no sexism is detected, and also assess the tone and context to identify language that targets, belittles, or promotes sexist situations based on a person's gender.
> **TASK 2:** As a social media analyst, consider the context and underlying message of the tweet: does it actively promote or perpetuate sexism, serve as a means of documenting sexist behavior, narrate or bear witness to a sexist moment, or express disapproval towards sexist attitudes by explicitly promoting sexism, serving as a means of documenting sexist behavior, or pushing back against and objecting to it; determine whether the tweet embodies sexism directly (DIRECT), reports a sexist event that has occurred (REPORTED), or delivers a critical judgment of sexism (JUDGEMENTAL).
> **TASK 3:** Analyze the tweet for sexism and categorize it based on the specific type it exhibits. Classify the sexism as one of three types: (1) Mocking feminism, which involves ridiculing or disparaging feminist ideologies or individuals. (2) Objectifying women by reducing them to body parts, where women are represented or addressed solely in terms of their physical appearance. (3) Expressing direct aggression, which involves using threatening, derogatory, or violent language towards women. Tag the corresponding sexism class accordingly, indicating the type of sexism the tweet embodies.

Prompt - Gen 6

Macro F1: T1.1 = 0.8390, T1.2 = 0.4850, T1.3 = 0.1960

TASK 1: Evaluate the given tweet as a Social Media Content Moderator for the presence of sexism, categorizing it as 'Direct Sexism' if sexism is directly expressed or stated, 'Described Sexism' if sexism is described or referenced, 'Criticized Sexism' if sexism is being criticized or challenged, or 'NO' if no sexism is detected, and also assess the tone and context to identify language that targets, belittles, or promotes sexist situations based on a person's gender, including microaggressions or sexist undertones.
TASK 2: As a social media analyst, consider the context and underlying message of the tweet: does it actively promote or perpetuate sexism, serve as a means of documenting sexist behavior, narrate or bear witness to a sexist moment, or express disapproval towards sexist attitudes by explicitly promoting sexism, serving as a means of documenting sexist behavior, or pushing back against and objecting to it; determine whether the tweet embodies sexism directly (DIRECT), reports a sexist event that has occurred (REPORTED), or delivers a critical judgment of sexism (JUDGEMENTAL).
TASK 3: Analyze the tweet for sexism and classify it into one of three categories: (1) Mocking feminism by ridiculing or disparaging feminist ideologies or individuals, (2) Objectifying women by reducing them to body parts where women are represented or addressed solely in terms of their physical appearance, (3) Expressing direct aggression by using threatening, derogatory, or violent language towards women. Tag the corresponding sexism class accordingly, indicating the type of sexism the tweet embodies.

Prompt - Gen 8

Macro F1: T1.1 = 0.8470, T1.2 = 0.5350, T1.3 = 0.2250

TASK 1: Evaluate the given tweet for the presence of sexism... provide specific examples of sexist language and offer suggestions for how the author could rephrase the tweet to avoid perpetuating sexism, be clear about the criteria used for categorization, and provide a detailed breakdown of the tone and context.
TASK 2: As a social media analyst, determine the tweet's intent: DIRECT, REPORTED, or JUDGEMENTAL.
TASK 3: As a Social Media Analyst, analyze the tweet for sexism, classifying it into one of three categories: (1) Mocking feminism by ridiculing or disparaging feminist ideologies or individuals, (2) Objectifying women by reducing them to body parts where women are represented or addressed solely in terms of their physical appearance, or (3) Expressing direct aggression by using threatening, derogatory, or violent language towards women, and indicate the type of sexism the tweet embodies with concise reasoning.

Prompt - Gen 10 (Final)

Macro F1: T1.1 = 0.8478, T1.2 = 0.5724, T1.3 = 0.2477

TASK 1: Evaluate the tweet, assigning YES if sexism is detected (direct, described, or criticized), and categorize it as Direct Sexism, Described Sexism, or Criticized Sexism if sexism is present, or NO if not.
TASK 2: Given a tweet that contains a sexist message, identify the intention behind the message. Choose one of the following categories: DIRECT, REPORTED, JUDGEMENTAL.
TASK 3: If the tweet was not classified as sexist in TASK 1, respond with: NO. Otherwise, identify and list all the types of sexism that apply: IDEOLOGICAL-INEQUALITY, STEREOTYPING-DOMINANCE, OBJECTIFICATION, SEXUAL-VIOLENCE, MISOGYNY-NON-SEXUAL-VIOLENCE.

Final Prompt Template

The final prompt is designed to elicit detailed and structured responses from the language model in a JSON format. It consolidates general tweet-level metadata with task-specific classifications for sexism detection, intention, and category. The evolved instructions are highlighted below.

TWEET: { }
INSTRUCTIONS: Return a JSON object with the following attributes. Answer based on the content of the tweet.

"RELEVANT_INFO": Include 4 different sections where you add relevant information and insights for the analysis.

"DISCLAIMERS_AND_WARNINGS": Warnings and disclaimers on potential bias or sensitive content in the analysis.

"TASK1": Evaluate the tweet, assigning YES if sexism is detected (direct, described, or criticized), and categorize it as Direct Sexism, Described Sexism, or Criticized Sexism if sexism is present, or NO if not.
Examples for TASK1:
YES: "It's less of #adaywithoutwomen and more of a day without feminists, which, to be quite honest, sounds lovely"
YES: "I'm sorry but women cannot drive, call me sexist or whatever but it is true"
YES: "You look like a whore in those pants - My brother of 13 when he saw me in a leather pant"
NO: "Where all the white women at?"
NO: "The shocking video of a woman at the wheel who miraculously escapes an assassination attempt"
NO: "Congratulations to the local high school basketball team on their big championship win! The whole town is so proud"

"TASK2": Given a tweet that contains a sexist message, identify the intention behind the message. Choose one of the following categories: DIRECT: The author expresses sexist ideas or incites sexism. REPORTED: The author is sharing or describing a sexist event they witnessed or experienced JUDGEMENTAL: The author is condemning or criticizing sexist behavior.
Examples for TASK2:
DIRECT: "A woman needs love, to fill the fridge, if a man can give this to her in return for her services, I don't see what else she needs."
REPORTED: "I doze in the subway; I open my eyes feeling something weird: the hand of the man sat next to me on my leg #SquealOnYourPig."
JUDGEMENTAL: "As usual, the woman was the one quitting her job for the family's welfare."

"TASK3": If the tweet was not classified as sexist in TASK 1, respond with: NO. Otherwise, identify and list all the types of sexism that apply. Choose from the following categories: IDEOLOGICAL-INEQUALITY: The tweet discredits the feminist movement, rejects gender inequality, or portrays men as victims of gender-based oppression. STEREOTYPING-DOMINANCE: The tweet promotes traditional gender roles, portrays women as less capable or suitable for certain tasks, or claims male superiority. OBJECTIFICATION: The tweet reduces women to their physical attributes or treats them as objects without personal agency. SEXUAL-VIOLENCE: The tweet includes sexual harassment, explicit sexual suggestions, or references to sexual assault. MISOGYNY-NON-SEXUAL-VIOLENCE: The tweet expresses hatred or non-sexual violence toward women.

Examples for TASK3:
IDEOLOGICAL-INEQUALITY: "#Feminism is a war on men, but it's also a war on women. It's a war on female nature, a war on femininity"
IDEOLOGICAL-INEQUALITY: "I think the whole equality thing is getting out of hand. We are different, that's how we're made!"
STEREOTYPING-DOMINANCE: "Most women no longer have the desire or the knowledge to develop a high quality character, even if they wanted to."
STEREOTYPING-DOMINANCE: "I feel like every time I flirt with a girl they start to imagine all the ways they can utilize me"
OBJECTIFICATION: "No offense but I've never seen an attractive African American hooker. Not a single one"
OBJECTIFICATION: "Most of the times I just want women for sex just because every time I try to make a relationship out of it they always break my heart"
SEXUAL-VIOLENCE: "Wanna touch your tits... you can't imagine what I can do on your body"
SEXUAL-VIOLENCE: "Fuck that cunt, I would with my fist"
MISOGYNY-NON-SEXUAL-VIOLENCE: "Domestic abuse is never okay... Unless your wife is a bitch"
MISOGYNY-NON-SEXUAL-VIOLENCE: "Some women are so toxic they don't even know they are draining everyone around them in poison. If you lack self-awareness you won't even notice how toxic you really are"