

Exploring LLM Performance for sexist hate speech detection on French social media

Shaden Shabayek¹, Camilla Penzo¹, Eunhyung Lee¹, Farah Benamara^{2,3},
Véronique Moriceau², Gohar Dashyan¹

¹PEReN, Paris, France

²Université de Toulouse, Toulouse INP, CNRS, IRIT, Toulouse, France

³IPAL, CNRS-NUS-A*STAR, Singapore

Correspondence: shaden.shabayek@peren.gouv.fr

Abstract

Deploying Large Language Models (LLMs) for content moderation on social media platforms introduces new challenges related to potential errors and biases, especially for speech-related policies. In this paper, we evaluate a diverse sample of 20 LLMs in zero-shot classification, for the specific case of sexist hate speech in French; a challenging task due to the scarcity of corpora in French and the English-centric research. We elaborate multiple protocols to explore the models' understanding of the task and sensitivity to the language of the prompts and class labels. We show that small or medium size models can outperform much larger models, supporting the paradigm that scale is not the main driver of performance. We also demonstrate that models' performance is not significantly altered when receiving specific guidelines about how to annotate potentially sexist content in French. Qualitative error analysis reveals that guidelines can help capture edge-cases (e.g. condemn sexism), but those are off-set by new misclassifications. This highlights a limitation of zero-shot prompting for content moderation, where specific instructions are crucial for adapting to different languages and cultures.

CONTENT WARNING: This paper contains some examples of abusive and hateful content that some readers may find offensive or distressing.

1 Introduction

With the recent development of generative AI, many social media platforms are scaling their content moderation by deploying generative AI technology (see the case of YouTube (O'Connor and Moxley, 2023) and Facebook (20xx)). These policies are implemented globally for users across diverse languages and cultures.

Standard content moderation rules are decided and implemented by platforms (Gillespie, 2018),

through a mix of human moderators and automated systems. Even before the integration of generative AI, these rules already struggled to achieve consensus among stakeholders, particularly regarding subjective speech rules (Kaye, 2019), such as hateful or misleading content. Building on these challenges, deploying Large Language Models (LLMs) for content moderation introduces a new layer of issues regarding induced errors and biases (see for example Motoki et al. (2024), Boelaert et al. (2025), Muti et al. (2025)). Addressing these issues requires evaluating the tools that are being deployed and covering the variety of speech related policies, since each can carry a distinct set of problems. Hence, our research tackles the specific case of sexist hate speech online in the French language.

The automatic detection of sexist hate speech in French is challenging for two primary reasons. First, existing research on online hate speech remains heavily English-centric (Nozza, 2021; Poletto et al., 2021), which limits transfer to other languages due to the cultural nuances. Second, defining hate speech itself is complex (Fortuna and Nunes, 2018; Vidgen et al., 2019; Siegel, 2020); operational definitions are needed to conduct research on the same topic, yet a diversity of definitions have emerged to capture specificities of sub-categories or related concepts like sexism or racism.

In the present research, we deliberately adopt the term “sexist hate speech”, to bridge definitions across three distinct institutional levels: (micro) specific legal codes in French-speaking countries, such as “sexist or sexual outrage”¹, (meso) regional frameworks such as the work conducted by the Council of Europe² on sexist hate speech and (macro) broader definitions such as the one formalized by the UN,³ where hate speech is “*offensive discourse targeting a group or an individual based*

¹ Articles 222-33-1-1 & R. 6925 of the French Penal Code.

² See [Combating Sexist Hate Speech](#).

³ See [What is hate speech?](#) by the UN.

on inherent characteristics (such as race, religion or gender) and that may threaten social peace” or specific operational hateful conduct policies by social media platforms.⁴

In this context, we investigate LLM performance in content moderation zero-shot classification tasks, through multiple research questions: *Can LLMs effectively detect sexism in French social media content? Do bigger models have better performance? To what extent does the prompting strategy (e.g., language of instructions, inclusion of annotation guidelines) impact results?*

We evaluate a diverse sample of LLMs, including models of different sizes by multiple providers (DeepSeek, Cohere, Google, Meta, Microsoft, Mistral, Qwen). We use the [Chiril et al. \(2020\)](#) dataset which is one of the few annotated datasets of social media sexist content in French. We elaborate multiple protocols to explore the models’ understanding of the task, when providing specific annotation guidelines and the models’ sensitivity to the language of the prompts and class labels. We focus on open-weight models under 70B parameters, omitting frontier models. From an economic and operational perspective, given the abundance of daily content online, it is far more likely that digital platforms deploy average-sized models for real-time moderation. For example, the largest version of the content safety model Llama Guard by Meta, has 12B parameters.⁵

Our contribution. First, we show that small or medium size models can outperform much larger models; for instance, Qwen2.5-7B-Instruct achieves the highest Macro F_1 score, supporting the paradigm that scale is not the main driver of performance. Second, we provide a critical discussion on evaluation metrics in the specific context of content moderation, arguing that both positive and negative classes must be accounted for, as under-detection undermines platform safety while over-detection harms user expression. Third, we provide a counter-intuitive finding: incorporating annotation guidelines so that the models better understand the task does not significantly improve alignment with the annotated data. Qualitative error analysis reveals that guidelines can help capture edge-cases

(e.g. condemn sexism), but those are off-set by new misclassifications. Finally, we inspect the role of language and show that there is not a significant change in performance when aligning the language of the prompt and the labels with the language of the dataset. Surprisingly, we observe an asymmetrical effect, where prompting the model in English but providing French labels (“oui” / “non”) does not lead to a significant change in performance whereas when prompting the model in French but providing English labels makes the performance of the majority of models drop.

Our research highlights the need for more datasets in French on hate speech online, preferably curated by stakeholders, to deepen the error analysis and disentangle potential systematic biases of LLMs from those of the annotators. For instance, our research provides examples of content where a majority of our sample of LLMs can be in disagreement with the annotators. Finally, hate speech online is inherently a dynamic topic, possibly linked to ongoing political or societal events offline ([Siegel, 2020](#)). Hence, more documentation by social media platforms about which generative AI-powered tools are integrated in content moderation pipelines can be useful for the academic community.

2 Related Work

2.1 Sexist Hate Speech Datasets

[Poletto et al. \(2021\)](#) provide a systematic review of resources and corpora for hate speech detection. They highlight the difficulties surrounding the study of hate speech, as it overlaps with related concepts such as abusive, offensive or toxic language. They attribute this diversity to varying topical focuses (racism, anti-semitism, sexism, misogyny and homophobia) where each resource targets a specific abusive phenomenon. For instance, [Abercrombie et al. \(2023\)](#) discuss in their systematic review of 63 resources,⁶ the challenges specific to sexism and gender-based violence (GBV). They provide a careful overview of information surrounding the annotation of datasets. They document that more than half of the datasets are annotated by either anonymous crowdworkers or by individuals within the academic sphere (authors, other researchers, undergraduate students). They find that inclusion of minority perspectives and interaction with stakeholders was limited, with one exception ([Vidgen](#)

⁴For example, Meta provides the following definition of hateful conduct: *direct attacks against people — rather than concepts or institutions — on the basis of what we call protected characteristics (...)*. See [link](#) (last visited July 2, 2026).

⁵See <https://huggingface.co/meta-llama/Llama-Guard-4-12B>. The same model is available with 1B and 8B parameters.

⁶See github.com/HWU-NLP/GBV-Resources.

et al., 2021) where information about whether the annotators themselves have been victims or not of online abuse is provided. They document that coverage of non-English languages is poor, where most available resources are exclusively in English (26 datasets, 41.3%) followed by Spanish (8 datasets, 12.7%), Arabic (8 datasets, 12.7%) and French (5 datasets, 7.9%).

However, there are notable regional initiatives, such as GermEval (Wiegand et al., 2018) for the identification of offensive language in German, Evalita (Basile et al., 2020) and IberEval (Fersini et al., 2018) both with a track on automatic misogyny detection. As for French, the few annotated datasets have distinct set-ups aimed at addressing a drawback in prior work. Chung et al. (2019) constructed the CONAN dataset, using an expert-based approach, where NGO professionals were asked to write “hate text”. These initial sentences were subsequently augmented by paraphrasing. This dataset includes a rich sample of hateful narratives, however they may not accurately reflect social media content due to the lack of context. Social media hateful content is typically related to real-world events (Olteanu et al., 2019; Kaye, 2019; Williams and Burnap, 2016; Müller and Schwarz, 2021). In contrast, Ousidhoum et al. (2019) constructed an ecologically valid, multilingual hate speech Twitter dataset including English (3,353 tweets), French (4,014 tweets) and Arabic (3,353 tweets), based on slurs and controversial topics. Their dataset was annotated via Amazon Mechanical Turk, using guidelines explaining that “offensive comments and hate do not necessarily come in the form of profanity”. Similarly, Chiril et al. (2020) developed a dataset of sexist hate speech from French tweets. They provide an annotation scheme to capture more nuance in hate speech, namely they distinguish direct sexism contained in tweets (targeting directly a person), descriptive sexist content (targeting a group in general, such as women) and tweets that report sexism. The dataset includes nearly 11,000 annotated tweets, which is 3 times higher than Ousidhoum et al. (2019).

2.2 Automatic Hate Speech Detection

Research on automatic hate speech detection can be characterized by three main waves. First, early research (pre-BERT) on hate speech detection relies on text mining techniques, such as bag-of-words, N-grams, topic classification and sentiment analysis; see the literature review by Fortuna and Nunes

(2018). In a more recent systematic literature review, Alkomah and Ma (2022) investigate 138 papers ranging from 2016 to 2021 and document the emergence of deep learning methods for detecting hate speech, namely Convolutional Neural Networks (CNN) (Zimmerman et al., 2018; Gambäck and Sikdar, 2017; Siino et al., 2021) and Bidirectional Encoder Representations from Transformers (BERT) (Caselli et al., 2021; Wiedemann et al., 2020), which require fine-tuning. To overcome data scarcity in non-English languages, new techniques have emerged such as data augmentation by translating corpora (e.g. Chung et al. (2020)). A third wave of detection techniques has emerged with the recent developments of generative AI. Wei et al. (2022), for example, discuss how scale can induce the emergence of new abilities for LLMs.

2.3 LLMs in Content Moderation

There is an emerging literature on the use of LLMs for content moderation on social media. For example, Bonagiri et al. (2025) evaluate the performance of four LLMs (Mistral-7B, Llama2-13B, GPT-3.5-Turbo and GPT-4o-Mini) in zero-shot detection of harmful content, based on a dataset of YouTube videos. They show that these models outperform, in terms of accuracy (maximum value at 0.7), proprietary tools such as Google’s widely used Perspective API and OpenAI Moderation. Similarly, Li et al. (2024) evaluate gpt-3.5-turbo on zero-shot classification of Hateful, Offensive, and Toxic (HOT) content, using a baseline prompt that includes the definition of HOT. They then compare the model’s performance to annotations provided by individuals recruited via Mechanical Turk. The authors document an inclination towards classifying content as HOT, where LLMs more frequently return the positive answer for the baseline protocol. Furthermore, Kath et al. (2026) propose a broad framework, which explains how LLMs are used in different phases of content moderation, ranging from data labeling and generation for training models, to detection and auditing methodologies.

However, the extensive literature on bias, fairness and safety (see Gallegos et al. (2024)) of LLMs calls for careful inspection of performance, especially when it comes to culturally embedded concepts like sexism or implicit expressions (Muti et al., 2025). For instance, Reuver et al. (2025) build a hybrid experimental protocol including nine sexism researchers and LLMs, where the re-

Table 1: Distribution of tweets (Chiril et al., 2020).

Class	Tweets	%
Non-sexist	7,394	65,5%
Sexist or reporting sexism	3,894	34,5%
Total	11,288	100%

searchers assess the model’s knowledge about sexism and even co-create with the model a common definition of sexism. Similarly, Di Bonaventura et al. (2024) design a controlled experiment where, AI experts are asked to evaluate the usefulness and trustworthiness of LLM-generated explanations related to abusive language. They find that the usefulness and trustworthiness, measured by Likert scores assigned to models before and after treatment (viewing the LLM generated text), drops significantly and the authors call for a greater attention to the cultural biases of the models.

Hence, this body of research calls for careful inspection of LLMs regarding sensitive topics and a broader evaluation of a larger sample, as typically only a very small number of models are inspected simultaneously. To bridge this gap, we design multiple experimental protocols, which we run for a large sample of models, to evaluate LLM understanding of sexism in French.

3 Methodology

3.1 Model Selection

For our evaluation, we construct a large sample of LLMs following multiple criteria. In order to have a diverse sample, we have included models from major providers, such as DeepSeek, Google, Meta, Qwen, Mistral, and Cohere. Where possible, we included, for each provider, models with different capabilities. We also included models with a high number of downloads in the previous month, to ensure that the models are actually used. The models can be divided into three groups by number of parameters: 5 small models ($\leq 4B$), 8 average models (7B to 12B), 7 large models (27B to 70B). The list of models is in Table 6 in the Appendix.

3.2 Evaluation Dataset

To evaluate the sample of LLMs, we use the French social media dataset curated by Chiril et al. (2020). This dataset is one of the very few available datasets in French with hateful social media content. The authors collected French tweets between Octo-

ber 2017 and May 2018 by using representative keywords (e.g. *femme* (woman), *fille* (girl), *enceinte* (pregnant), *cuisine* (kitchen)), names of individuals either targeted by sexism online or are known for perpetuating sexist content online (e.g. some influencers or politicians) and specific hashtags which are associated with social movements and awareness campaign (e.g. *#balancetonporc*, *#metoo*, etc.). They distinguish between sexist content (tweets directly addressed to a target individual or addressed to a target group) and reported assertions which consist of individuals relaying sexist content in order to condemn it. After accounting for repeated tweets with identical labels and removing conflicting annotations from the original 11,834 tweets, the final corpus contains 11,288 tweets. As shown in Table 1, we have two categories: (1) tweets which contain sexist content, including reported tweets as indicated in the annotation guidelines (34.5% of the annotated tweets), and (2) tweets which do not contain any sexist content (65.5% of the annotated tweets). Following Chiril et al. (2020), an example of a sexist tweet is “*t’es une femme pq tu veux parler de foot?*” (“You’re a woman why do you want to talk about football”), while a non-sexist baseline example is “*Laetitia Casta pas d’accord avec #balancetonporc*” (“Laetitia Casta disagrees with #SquealOnYourPig”).

3.3 Design of Protocols

Our baseline protocol consists of prompting a given model with the question: “*is the following tweet {} sexist?*”

To ensure that there is no variability in the formats of output, we enforce a strict binary schema (“yes” or “no”) by using guided decoding (structured outputs⁷). We design two distinct families of experimental protocols, one which explores the models’ understanding of the task and a second one which aims at analysing whether language (prompts or labels) affects performance. We provide the complete list of prompts in Appendix B.

3.3.1 Task Comprehension Protocols

First, we augment the baseline prompt with the original annotation guidelines from Chiril et al. (2020). These guidelines (see Appendix A) include a description of each class (sexist, non-sexist) with specific examples, which boils down to few-shot

⁷Implemented using the Outlines/vLLM structured generation framework.

prompting. Furthermore, to evaluate robustness of the models’ understanding of the task, we test a negative framing “*Is the following tweet {} non-sexist?*”. This protocol helps to identify whether there could be a bias towards “yes” or “no”. Finally, to isolate the impact of the quality of the guidelines and degraded performance due to language mismatch (since the dataset is in French), we test the protocols both in English and in French by translating the annotation guidelines.

3.3.2 Language Sensitivity Protocols

Second, we design complementary protocols to test the models’ sensitivity to language because our dataset is in French. This yields four protocols, where we cross prompt language (English/French) and label language (“yes”/“oui” or “no”/“non”), to test whether the choice of language and labels can interact with the models’ performance and whether aligning the language of the prompt and the dataset influences performance.

4 Results and Discussion

4.1 Baseline Protocol

We report the results of multiple evaluation metrics in Table 2.

(Accuracy) Given the mild imbalance of the Chiril et al. dataset (34.5% positive, see Table 1), a dummy classifier which only classifies as negative (non-sexist) has a baseline accuracy of 0.66. Only two evaluated LLMs (Phi-3.5-MoE-instruct and Qwen2.5-7B-Instruct) beat this score. When looking at balanced accuracy (the normalized metric which accounts for class distributions), the score is at 0.5 and all models outperform this threshold. However, only three models (Qwen2.5-7B-Instruct, Qwen3-30B-A3B-Instruct-2507, Phi-3.5-MoE-instruct) achieve a substantially higher score (> 0.6).

(False Positive Rate) Next we shift our focus to metrics that evaluate model performance on specific classes (positives or negatives). The False Positive Rate (FPR) gives the probability of misclassifying a non-sexist tweet as sexist. In the context of automated content moderation on social media, a high FPR means over-moderation by flagging non-harmful content. A random coin-flip dummy classifier gives a FPR of 0.5. In our sample, half of the evaluated models have an FPR which is strictly

lower than that threshold, possibly indicating a positive class (sexist) bias or higher precaution for flagging sexist content.

(Precision) Next we inspect precision. Precision measures the proportion of tweets that are actually sexist (based on human annotation) out of all tweets that the model classifies as sexist. A dummy classifier which always predicts positive (sexist) has a precision of 0.35. Among the evaluated LLMs, we observe that half of the models have a precision between 0.35 and 0.37. Finally, the model Phi-3.5-MoE-instruct has the highest precision (0.63) indicating that when it flags a tweet as sexist, it is correct three out of five times. It seems to be the only model operationally satisfying, if we want to ensure that a fair proportion of alerts actually correspond to sexist tweets.

(Recall) As a complement to precision, we inspect recall which gives the proportion of tweets correctly classified as sexist out of all the tweets annotated as sexist. We observe that 9 out of the 10 previously identified models with a precision between 0.35 and 0.37, belong to the top tier for the highest recall values (> 0.59). This overlap illustrates a precision-recall trade-off for these models. We observe a notable exception, the model Qwen2.5-7B-Instruct has a relatively low FPR (0.38), the second highest precision among the sample of LLMs (0.52) and a relatively high recall (0.77). The model Phi-3.5-MoE-instruct which scored the highest precision, has the second lowest recall within our sample, but struggles in identifying the positives.

(F_1 and Macro F_1) To overcome the precision-recall trade-off we turn to F_1 scores. The standard F_1 score (positive class), places in the top 5 models like Gemma-3-27b-it and Qwen2.5-3B-Instruct, which display high FPR (respectively 0.84 and 0.97) and a precision close to the dummy classifiers baseline. This indicates that such models systematically classify non-sexist tweets as sexist. In the context of content moderation, both classes (positives and negatives) matter, since not detecting harmful content is not optimal in terms of platform safety and over-moderating non-harmful content is a harm itself. Hence, we inspect the Macro F_1 score because it accounts for both classes equally and provides a balanced view of the model’s ability to correctly classify sexist and non-sexist content.

The model Qwen2.5-7B-Instruct ranks first for both F_1 and Macro F_1 , indicating that it

Model	Accuracy	Balanced Accuracy	FPR	Precision	Recall	F_1	Macro F_1
♣ Qwen2.5-7B-Instruct	0.67	0.69	0.38	0.52	0.77	0.62	0.67
♠ Qwen3-30B-A3B-Instruct-2507	0.65	0.64	0.32	0.50	0.59	0.54	0.63
♠ Gemma-3-27b-it	0.43	0.55	0.84	0.37	0.94	0.54	0.40
♣ Gemma-3-12b-it	0.51	0.57	0.62	0.39	0.76	0.52	0.51
◇ Qwen2.5-3B-Instruct	0.36	0.50	0.97	0.35	0.97	0.51	0.29
♣ DeepSeek-R1-Distill-Llama-8B	0.40	0.52	0.88	0.36	0.92	0.52	0.36
♣ DeepSeek-R1-Distill-Qwen-7B	0.39	0.51	0.89	0.35	0.91	0.51	0.35
♠ Phi-3.5-MoE-instruct	0.71	0.64	0.13	0.63	0.41	0.50	0.65
◇ Gemma-3-1b-it	0.42	0.51	0.80	0.35	0.82	0.49	0.40
◇ Llama-3.2-3B-Instruct	0.49	0.53	0.59	0.37	0.65	0.47	0.49
♣ Llama-3.1-8B-Instruct	0.53	0.53	0.47	0.38	0.54	0.44	0.52
♣ Mistral-7B-Instruct-v0.3	0.45	0.50	0.67	0.35	0.67	0.46	0.45
◇ DeepSeek-R1-Distill-Qwen-1.5B	0.45	0.52	0.69	0.36	0.73	0.48	0.45
♣ Ministral-8B-Instruct-2410	0.55	0.54	0.41	0.39	0.49	0.43	0.53
♠ Llama-3.3-70B-Instruct	0.59	0.56	0.32	0.42	0.43	0.43	0.56
♠ Aya-23-35B	0.46	0.52	0.67	0.36	0.71	0.48	0.46
◇ Qwen3-4B	0.56	0.53	0.37	0.38	0.44	0.41	0.53
♣ Aya-23-8B	0.52	0.51	0.45	0.35	0.46	0.40	0.51
♠ Mixtral-8x7B-Instruct-v0.1	0.54	0.54	0.45	0.38	0.52	0.44	0.53
♠ Aya-expansive-32b	0.60	0.55	0.27	0.42	0.36	0.39	0.55
Baseline Dummy Classifiers							
Coin-flip	0.50	0.50	0.50	0.35	0.50	0.41	0.49
Stratified random	0.55	0.50	0.35	0.35	0.35	0.35	0.50
Always Positive	0.35	0.50	1.00	0.35	1.00	0.52	0.26
Always Negative	0.66	0.50	0.00	–	0.00	0.00	0.40

Table 2: Baseline protocol, results are ranked by descending order of the F_1 scores. Models are marked with symbols indicating their category in terms of number of parameters: ♠ big, ♣ average, ◇ small. Top three models for each metric are in bold.

can find a balance between correctly identifying sexist tweets (recall) and limiting false flagging (precision). It also suggests that an average size model can out-perform much larger models for these safety-related tasks. The model Phi-3.5-MoE-instruct ranks second in terms of Macro F_1 score but only eighth in terms of F_1 indicating a high precision but a relatively low recall. For the remaining protocols, we use the Macro F_1 as our primary ranking metric.

We observe that small or medium size models can outperform large models within our sample. We test whether there is a positive relationship between a high F_1 or Macro F_1 and the number of model parameters (log scale). We find no statistically significant relationship between the number of parameters and both metrics (p-value = 0.404 for F_1 and 0.058 for Macro F_1). However, we inspect the number of monthly downloads (log scale) as it could indicate a community based measure of quality. As shown in Figure 1, this variable has a statistically significant relationship with the standard F_1 score at the 5% threshold.

Finally, to test the robustness of our results and support our methodological choice of selecting general purpose LLMs, we evaluate two

safety-tuned Llama-Guard models (1B and 8B) by Meta. We find that their Macro F_1 performance places them in the 14th and 6th positions respectively among the evaluated models (see Appendix D.1). Hence, a general-purpose smaller model like Qwen2.5-7B-Instruct out-performs safety-tuned models like Llama-Guard-8B on this specific task. This may be due to the broad safety taxonomies used to train safety-tuned models, whereas general-purpose models might better capture the nuanced context required for specific moderation tasks.

4.2 Task Comprehension

We take our analysis further, by inspecting whether model performance is related to a poor understanding of the definition of sexism, implying a poor understanding of the classification task. We test whether the performance can be improved by providing models with the annotation guidelines in (Chiril et al., 2020). The results are reported in Table 3. A two-sided Wilcoxon signed-rank test reveals no statistically significant difference in Macro F_1 scores between the baseline protocol and the protocol augmented with the English guidelines. While the number of LLMs we compare is large relative to existing literature, it may still be small

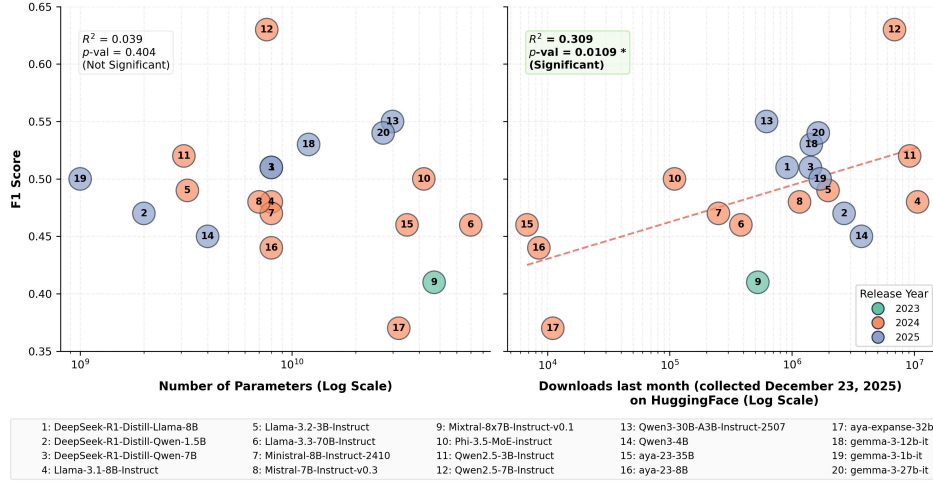


Figure 1: F_1 scores of models of different size and popularity. Left panel: F_1 scores plotted against the number of parameters of each model. Right panel: F_1 scores plotted against the number of downloads on Hugging Face.

from a statistical point of view. Hence, these results should be read with caution as non-significance may be linked to the sample size. Furthermore, an error analysis of a random sample of 60 predictions from our top-performing model revealed that introducing the guidelines helped produce correct predictions for ten cases, but equally resulted in 11 misclassifications (see Appendix D.3).

To identify whether the guidelines may have failed because the instructions themselves were poor or because the models struggle with the language mismatch since the dataset is in French, we translated these guidelines into French. Similarly, the Macro F_1 is not significantly different from the baseline protocol. Finally, the negative framing of the prompt in both languages, augmented with the annotation guidelines, does not significantly alter the performance compared to the baseline protocol.

These negative results can suggest that LLMs might be following implicit internal rules or safeguards (see e.g. Dong et al. (2025)) for sensitive topics, such as sexist content detection. There exists an emerging body of literature about using LLMs as research subjects, for example as survey respondents. Boelaert et al. (2025) show that models cannot replace research subjects for tasks related to opinions or attitudes, because they harbor a “machine” bias with very little variance regarding some topics. Hence in our case, an open question is whether we should expect LLMs to be proficient content moderators, since this task requires models to follow specific instructions, rather than relying on their own internal rules.

4.3 Language Sensitivity

Next, we further investigate the impact of language mismatch on performance, by varying the prompt language and the labels. The results are reported in Table 4. The experiment consisting of using the baseline prompt in French with French labels (“oui” or “non”) returns results which are not significantly different from the baseline. Similarly, keeping the baseline prompt in English but providing French labels does not result in a significant change in model performance measured by Macro F_1 . However, when providing the baseline prompt in French with English labels the results are significantly different from the baseline protocol at the 5% threshold (p-value at 0.028).

5 Conclusion

To the best of our knowledge, this paper is the first to evaluate a large sample of LLMs for sexist hate speech detection in French. We have designed two families of experimental protocols to identify models’ sensitivity to guidelines when running a classification for a sensitive topic and the impact of prompt and label language on performance. We believe that a greater focus on relatively small LLMs with an open-source licence (e.g., Apache 2.0) is very important for content moderation tasks. Although there is insufficient publicly accessible information about the type and size of models deployed for content moderation, given the abundance of content, it is highly probable that only small or average size models are practically viable for deployment by social media

model	Baseline Eng. prompt	Eng. Annotation Guidelines + Eng. prompt	Fr. Annotation Guidelines + Fr. prompt	Eng. Annotation Guidelines + Eng. ¬ prompt	Fr. Annotation Guidelines + Fr. ¬ prompt
♣ Qwen2.5-7B-Instruct	0.67	0.72	0.62	0.70	0.55
♠ Phi-3.5-MoE-instruct	0.65	0.58	0.70	0.38	0.54
♠ Qwen3-30B-A3B-Instruct-2507	0.63	0.69	0.39	0.50	0.44
♠ Llama-3.3-70B-Instruct	0.56	0.61	0.52	0.59	0.48
♠ aya-expanse-32b	0.55	0.59	0.54	0.59	0.52
♠ Mixtral-8x7B-Instruct-v0.1	0.53	0.49	0.57	0.50	0.53
◇ Qwen3-4B	0.53	0.46	0.33	0.49	0.44
♣ Ministral-8B-Instruct-2410	0.53	0.53	0.53	0.50	0.51
♣ Llama-3.1-8B-Instruct	0.52	0.49	0.52	0.31	0.52
♣ gemma-3-12b-it	0.51	0.28	0.35	0.58	0.64
♣ aya-23-8B	0.50	0.47	0.42	0.50	0.48
◇ Llama-3.2-3B-Instruct	0.49	0.50	0.46	0.52	0.50
♠ aya-23-35B	0.46	0.41	0.52	0.50	0.50
◇ DeepSeek-R1-Distill-Qwen-1.5B	0.45	0.48	0.44	0.46	0.44
♣ Mistral-7B-Instruct-v0.3	0.45	0.48	0.51	0.49	0.51
◇ gemma-3-1b-it	0.40	0.33	0.54	0.46	0.49
♠ gemma-3-27b-it	0.40	0.52	0.64	0.54	0.58
◇ DeepSeek-R1-Distill-Llama-8B	0.36	0.32	0.37	0.51	0.42
♣ DeepSeek-R1-Distill-Qwen-7B	0.35	0.35	0.52	0.48	0.37
◇ Qwen2.5-3B-Instruct	0.29	0.29	0.31	0.45	0.46
Wilcoxon Signed-Rank Test vs Baseline, p-value	-	0.407	0.965	0.324	0.896

Table 3: Macro F_1 of the **task comprehension protocols**. Models are marked with symbols indicating their category in terms of number of parameters: ♠ big, ♣ average, ◇ small.

model	Eng. prompt Eng. Labels	Fr. prompt Fr. Labels	Fr. prompt Eng. labels	Eng. prompt French Labels
♣ Qwen2.5-7B-Instruct	0.67	0.45	0.38	0.61
♠ Phi-3.5-MoE-instruct	0.65	0.63	0.34	0.55
♠ Qwen3-30B-A3B-Instruct-2507	0.63	0.48	0.41	0.52
♠ Llama-3.3-70B-Instruct	0.56	0.51	0.26	0.48
♠ aya-expanse-32b	0.55	0.50	0.49	0.44
♠ Mixtral-8x7B-Instruct-v0.1	0.53	0.47	0.50	0.48
◇ Qwen3-4B	0.53	0.33	0.35	0.48
♣ Ministral-8B-Instruct-2410	0.53	0.54	0.43	0.52
♣ Llama-3.1-8B-Instruct	0.52	0.52	0.48	0.47
♣ gemma-3-12b-it	0.51	0.44	0.37	0.46
♣ aya-23-8B	0.50	0.46	0.50	0.48
◇ Llama-3.2-3B-Instruct	0.49	0.40	0.48	0.48
♠ aya-23-35B	0.46	0.50	0.49	0.51
◇ DeepSeek-R1-Distill-Qwen-1.5B	0.45	0.50	0.29	0.50
♣ Mistral-7B-Instruct-v0.3	0.45	0.50	0.44	0.45
◇ gemma-3-1b-it	0.40	0.41	0.47	0.54
♠ gemma-3-27b-it	0.40	0.55	0.40	0.55
♣ DeepSeek-R1-Distill-Llama-8B	0.36	0.43	0.51	0.48
♣ DeepSeek-R1-Distill-Qwen-7B	0.35	0.35	0.36	0.36
◇ Qwen2.5-3B-Instruct	0.29	0.32	0.37	0.58
Wilcoxon Signed-Rank Test vs Baseline, p-value	-	0.191	0.028*	0.658

Table 4: Macro F_1 of the **language sensitivity protocols**. Models are marked with symbols indicating their category in terms of number of parameters: ♠ big models, ♣ average models, ◇ small models. The superscript * indicates a significant p-value at the 0.05 threshold.

platforms. Furthermore, relatively small LLMs can be scrutinized and deployed by regulators, civil society and the academic community, without the need for a substantial computational infrastructure and financial resources. Our research highlights the need to curate datasets in French, in coordination with stakeholders and NGOs actively addressing issues related to hate speech online. Finally, there is a need to run an evaluation of LLMs using a multi-modal hate speech dataset, since generative AI is being used for content moderation on multiple digital platforms where the majority of content is a mix of multimedia and text.

6 Limitations

We acknowledge multiple limitations of the current research. First, we evaluated our sample of LLMs on only one dataset in French targeting sexist hate speech. To generalize the findings across broader hate speech, future research should apply the same evaluation protocols to other sub-categories of hate speech (e.g. xenophobic or racist content). Furthermore, the dataset was released in 2020⁸ and it might be that it was included as training material for some models, even though only the Tweet IDs were made public by the authors to comply with the Terms of Service of the platform. Second, while we think that for technical reasons, social media platforms likely deploy small or average size LLMs for content moderation, our research does not include frontier models (>300B parameters); evaluating these models remains an interesting avenue for future research. Third, we detected specific instances where there was total consensus among LLMs but complete disagreement with the annotated data. For example, the tweet “*Votre pote huile smith c’est vraiment la plus grosse s****e du réseau*” (“Your Friend Huile⁹ Smith is the biggest s**t of the network.”) was classified as sexist by 20 LLMs, yet was annotated as non-sexist in the dataset. This divergence calls for future investigation into the alignment between LLMs and human annotators (see Figure 2) under controlled experimental setup. Finally, our research focuses strictly on zero-shot prompting and does not evaluate fine-tuned models. Fine-tuning is a promising avenue for improving performance, but it might not be practical for real-

⁸See <https://github.com/patriChiril/An-Annotated-Corpus-for-Sexism-Detection-in-French-Tweets>.

⁹The author of the tweet wrote “Huile” instead of Will (like the actor), it sounds like “Will” when read in French. Notice that “Huile” means oil in French.

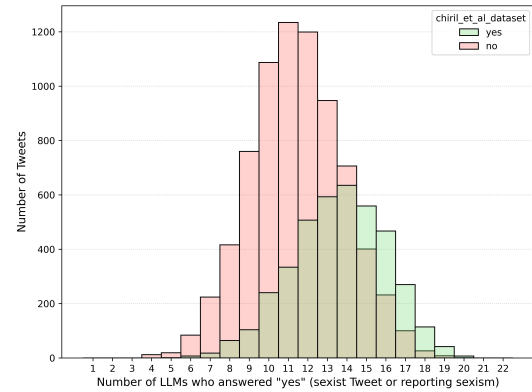


Figure 2: Distribution of the number of LLMs (1 to 20) predicting the positive (sexist) class per tweet, broken down by original dataset labels from Chiril et al. (2020), sexist (green) and non-sexist (red).

time moderation of ongoing events or emerging slang, as platforms would have to retrain models constantly.

7 Acknowledgments

This work was granted access to the HPC resources of IDRIS under the allocation 2025-AD011015776R1 made by GENCI. We acknowledge the constructive comments and feedback of colleagues at PEReN, as well as the two anonymous reviewers.

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)*, pages 170–186.
- Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6):273.
- Valerio Basile, Maria Di Maro, Danilo Croce, Lucia Passaro, and 1 others. 2020. Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian. In *CEUR Workshop Proceedings*, volume 2765. CEUR-ws.
- Julien Boelaert, Samuel Coavoux, Étienne Ollion, Ivaylo Petev, and Patrick Präg. 2025. Machine bias. how do generative language models answer opinion polls? *Sociological Methods & Research*, 54(3):1156–1196.
- Akash Bonagiri, Lucen Li, Rajvardhan Oak, Zeerak Babar, Magdalena Wojcieszak, and Anshuman Chhabra. 2025. Towards safer social media platforms: scalable and performant few-shot harmful

- content moderation using large language models. *arXiv preprint arXiv:2501.13976*.
- Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. Hatebert: Retraining bert for abusive language detection in english. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25.
- Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. An annotated corpus for sexism detection in french tweets. In *Proceedings of the twelfth language resources and evaluation conference*, pages 1397–1403.
- Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. Conan-counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. *arXiv preprint arXiv:1910.03270*.
- Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2020. Italian counter narrative generation to fight online hate speech. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLIC-it 2020)*, pages 82–88.
- Chiara Di Bonaventura, Lucia Siciliani, Pierpaolo Basile, Albert Meroño-Peñuela, and Barbara McGillivray. 2024. Is explanation all you need? an expert survey on llm-generated explanations for abusive language detection. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLIC-it 2024)*, pages 280–288.
- Yi Dong, Ronghui Mu, Yanghao Zhang, Siqi Sun, Tianle Zhang, Changshun Wu, Gaojie Jin, Yi Qi, Jinwei Hu, Jie Meng, and 1 others. 2025. Safeguarding large language models: A survey. *Artificial intelligence review*, 58(12):382.
- Help Center Facebook. 20xx. [How does facebook use artificial intelligence to moderate content?](#)
- Elisabetta Fersini, Paolo Rosso, Maria Anzovino, and 1 others. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. *Ibireval@sepln*, 2150:214–228.
- Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *Acm Computing Surveys (Csur)*, 51(4):1–30.
- Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.
- Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90.
- Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
- Suraj Kath, Sanket Badhe, Preet Shah, Ashwin Sampathkumar, and Shivani Gupta. 2026. Large language models in the abuse detection pipeline. *arXiv preprint arXiv:2604.00323*.
- David Kaye. 2019. Speech police: The global struggle to govern the internet.
- Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024. “hot” chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, 18(2):1–36.
- Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: measuring chatgpt political bias. *Public Choice*, 198(1):3–23.
- Karsten Müller and Carlo Schwarz. 2021. Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, 19(4):2131–2167.
- Arianna Muti, Chris Emmery, Debora Nozza, Alberto Barrón-Cedeño, Tommaso Caselli, and 1 others. 2025. The “r” in “woman” stands for rights. auditing llms in uncovering social dynamics in implicit misogyny. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics.
- Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914.
- Jennifer Flannery O’Connor and Emily Moxley. 2023. [Our approach to responsible ai innovation, youtube official blog](#).
- Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, 2:13.
- Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049*.
- Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection. *Language resources and evaluation*, 55(2):477–523.
- Myrthe Reuver, Indira Sen, Matteo Melis, and Gabriella Lapesa. 2025. [Tell me what you know about sexism: Expert-LLM interaction strategies](#)

and co-created definitions for zero-shot sexism detection. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8438–8467, Albuquerque, New Mexico. Association for Computational Linguistics.

Alexandra A Siegel. 2020. Online hate speech. *Social media and democracy: The state of the field, prospects for reform*, pages 56–88.

Marco Siino, Elisa Di Nuovo, Ilenia Tinnirello, and Marco La Cascia. 2021. Detection of hate speech spreaders using convolutional neural networks. In *CLEF (Working Notes)*, pages 2126–2136.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott A Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the third workshop on abusive language online*, pages 80–93.

Bertie Vidgen, Tristan Thrush, Zeerak Talat, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 1667–1682.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and 1 others. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Gregor Wiedemann, Seid Muhie Yimam, and Chris Biemann. 2020. Uhh-It at semeval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection. *arXiv preprint arXiv:2004.11493*.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the semeval 2018 shared task on the identification of offensive language.

Matthew L Williams and Pete Burnap. 2016. Cyberhate on social media in the aftermath of woolwich: A case study in computational criminology and big data. *British Journal of Criminology*, 56(2):211–238.

Steven Zimmerman, Udo Kruschwitz, and Chris Fox. 2018. Improving hate speech detection with deep learning ensembles. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Appendix

A Annotation Guidelines

The annotation guidelines used in several protocols were copied from Chiril et al. (2020). These guidelines provided three categories, sexist, non-sexist

and no decision. We made two minor changes to the guidelines:

- We dropped the last category no decision because the dataset to which we had access did not contain this class. In the following subsections we provide the exact text and the translation in French;
- We replaced the sentence “linguistic clues are underlined” by “Linguistic clues are enclosed between asterisks like this for the word clue *clue*” in order to help the models understand the instruction.

A.1 Annotation Guidelines in English

Given a tweet, annotation consists in assigning it one of the following three categories:

(i) **Sexist content:** it can be either direct (cf. (1) and (2)), descriptive (cf. (3) and (4)) or reporting (cf. (5) and (6)). The first two are real sexist messages but not the last one as reporting tweets must not be considered as sexist in a context of moderation. Direct sexist content, directly addressed to a woman or a group of women, generally uses second person pronoun/verb and imperatives, as shown in the examples below. (Linguistic clues are enclosed between asterisks like this for the word clue *clue*)

(1) *t’es* une femme pq tu veux parler de foot?

(2) *les femmes* qui sont en plus Dijonnaise ne *parlez* pas de foot s’voulai c’est comme si un aveugle manchot parler de passer le permis. In descriptive sexist content where the tweet describes a woman or women in general, clues can be presence of a named entity or use of generalizing terms.

(3) *Theresa May* succède à David Cameron. Pas mieux qu’*une femme* pour faire le ménage.

(4) La place de *la femme* dans la société moderne, est très précise elle est dans la cuisine.

When the sexist content is in fact a report of a sexism experience or a denunciation of a sexist behaviour, we observe the presence of reporting verbs, quotation and specific hashtags.

(5) I m’a *dit* “normal qu’elles soient moins payer

pcq enceinte=moins de temps au travail”.

(6) #RolandGarros : une petite *remarque* bien *#sexiste* sur l’émotivité des joueuses de tennis dans Stade2.

(ii) **Non sexist**: when the tweet has no sexist content (it may contain a specific hashtag but the content is not sexist), as in (7) and (8).

(7) Paris Match : journal d’investigation.

(8) Laetitia Casta pas d’accord avec #balancetonporc.

A.2 Annotation Guidelines Translated in French

Étant donné un tweet, l’annotation consiste à lui attribuer l’une des trois catégories suivantes:

(i) Contenu sexiste : ça peut être soit direct (cf. (1) et (2)), descriptif (cf. (3) et (4)) ou lié à un signalement ou témoignage (cf. (5) et (6)). Les deux premiers sont de vrais messages sexistes mais pas le dernier, car les tweets de témoignage ne doivent pas être considérés comme sexistes dans un contexte de modération. Le contenu sexiste direct, directement adressé à une femme ou à un groupe de femmes, utilise généralement le pronom/verbe à la deuxième personne et des impératifs, comme le montrent les exemples ci-dessous. (Les indices linguistiques sont encadrés par des astérisques comme ceci pour le mot indice indice)

(1) *t’es* une femme pq tu veux parler de foot ?

(2) *les femmes* qui sont en plus Dijonnaise ne parlez pas de foot s’il vous plaît c’est comme si un aveugle manchot parlait de passer le permis. Dans le contenu sexiste descriptif où le tweet décrit une femme ou les femmes en général, les indices peuvent être la présence d’une entité nommée ou l’usage de termes généralisants. (3) *Theresa May* succède à David Cameron. Pas mieux qu’*une femme* pour faire le ménage.

(4) La place de *la femme* dans la société moderne est très précise : elle est dans la cuisine.

Lorsque le contenu sexiste est en réalité le récit d’une expérience de sexisme ou la dénonci-

ation d’un comportement sexiste, on observe la présence de verbes de dénonciation, de guillemets et de hashtags spécifiques.

(5) Il m’a *dit* “normal qu’elles soient moins payées pcq enceintes = moins de temps au travail”

(6) #RolandGarros : une petite remarque bien #sexiste sur l’émotivité des joueuses de tennis dans Stade2.

(ii) **Non sexiste** : lorsque le tweet ne contient pas de contenu sexiste (il peut contenir un hashtag spécifique mais le contenu n’est pas sexiste), comme dans (7) et (8).

(7) Paris Match : journal d’investigation.

(8) Laetitia Casta pas d’accord avec #balancetonporc.’

B List Of Prompts

Table 5 provides the complete list of prompts used. Each prompt corresponds to a specific protocol which either belongs to the task comprehension protocols (see section 3.3.1 and Table 3 for the results) or the language sensitivity protocols (see section 3.3.2 and Table 4 for the results).

C List Of Models

Table 6 provides the complete list of Large Languages Models which were included in our experiment. We report for each the provider, the release date, the number of parameters and the number of downloads last month (provided on Hugging Face).

D Robustness Checks

We perform multiple robustness checks on our results. First, to test whether there is variability due to the non-deterministic nature of the models, we run the baseline protocol three times and find that there is no significant difference in our evaluation metrics (see Figure 3). Second, in subsection D.1, we test the robustness of our methodological choice to evaluate general-purpose models rather than models fine-tuned specifically for content moderation, by running the evaluation with Meta’s Llama Guard model, a model fine-tuned for content safety. Third, in subsections D.2 and D.3, we provide a metric to test for inter-model agreement and a qualitative error analysis, to investigate why including the guide-

Table 5: The list of prompts used for the task comprehension protocols and language sensitivity protocols.

Protocol name	Classes	Prompt
Baseline Eng. prompt	yes, no	Is the following Tweet {insert tweet} sexist?
Eng. Annotation Guidelines + Eng. prompt	yes, no	Annotation Guideline in English (see section A.1) + Taking into account these annotation instructions, is the following Tweet {insert tweet} sexist?
Fr. Annotation Guidelines + Fr. prompt	oui,non	Annotation Guideline in French (see section A.2) + En tenant compte des consignes d’annotation, le Tweet suivant {insert tweet} est-il sexiste ?
Eng. Annotation Guidelines + Eng. $\neg$ prompt	yes, no	Annotation Guideline in English (see section A.1) + Taking into account these annotation instructions, is the following Tweet {insert tweet} non-sexist?
Fr. Annotation Guidelines + Fr. $\neg$ prompt	oui,non	Annotation Guideline in French (see section A.2) + En tenant compte des consignes d’annotation, le Tweet suivant {insert tweet} est-il non-sexiste ?
Fr. prompt + Fr. Labels	oui,non	Le Tweet suivant {insert tweet} est-il sexiste?
Fr. prompt + Eng. labels	yes,no	Le Tweet suivant {insert tweet} est-il sexiste?
Eng. prompt + French Labels	oui,non	Is the following Tweet {insert tweet} sexist?

Table 6: The list of Large Language Models used for our evaluation task, including release dates, parameter counts, and monthly downloads (collected on December 23, 2025). Models are marked with symbols indicating their category in terms of number of parameters: ♠ big, ♣ average, ◇ small.

Model	Provider	Release Date	Params	Downloads
◇ gemma-3-1b-it	google	Mar 2025	1.0B	1,698,178
◇ DeepSeek-R1-Distill-Qwen-1.5B	deepseek-ai	Jan 2025	2.0B	2,666,104
◇ Qwen2.5-3B-Instruct	Qwen	Sep 2024	3.1B	9,161,762
◇ Llama-3.2-3B-Instruct	meta-llama	Sep 2024	3.2B	1,970,705
◇ Qwen3-4B	Qwen	Apr 2025	4.0B	3,695,155
♣ Mistral-7B-Instruct-v0.3	mistralai	May 2024	7.0B	1,152,149
♣ Qwen2.5-7B-Instruct	Qwen	Sep 2024	7.6B	6,889,798
♣ Llama-3.1-8B-Instruct	meta-llama	Jul 2024	8.0B	10,642,549
♣ aya-23-8B	CohereLabs	May 2024	8.0B	8,507
♣ Ministral-8B-Instruct-2410	mistralai	Oct 2024	8.0B	250,209
♣ DeepSeek-R1-Distill-Qwen-7B	deepseek-ai	Jan 2025	8.0B	1,413,417
♣ DeepSeek-R1-Distill-Llama-8B	deepseek-ai	Jan 2025	8.0B	911,402
♣ gemma-3-12b-it	google	Mar 2025	12.0B	1,437,880
♠ gemma-3-27b-it	google	Mar 2025	27.0B	1,639,410
♠ Qwen3-30B-A3B-Instruct-2507	Qwen	Jul 2025	29.9B	619,223
♠ aya-expense-32b	CohereForAI	Oct 2024	32.0B	11,062
♠ aya-23-35B	CohereLabs	May 2024	35.0B	6,804
♠ Phi-3.5-MoE-instruct	microsoft	Aug 2024	42.0B	109,377
♠ Mixtral-8x7B-Instruct-v0.1	mistralai	Dec 2023	47.0B	524,888
♠ Llama-3.3-70B-Instruct	meta-llama	Dec 2024	70.0B	381,671

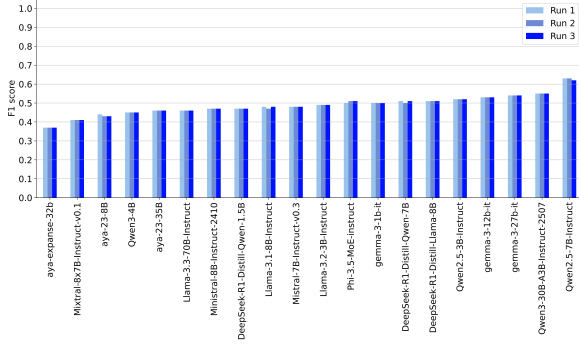


Figure 3: Three runs of the baseline protocol, with no significant difference.

lines did not significantly improve the performance of models.

D.1 Evaluation Task With a Fine-Tuned Model

For the purposes of our evaluation task, recall that we made a methodological choice to select LLMs that were not fine-tuned for moderation tasks on social media. This is because, sexist hate speech has a dynamic aspect, where new waves can emerge in relation to ongoing events. Hence, to test the robustness of our results, we evaluate the performance of two Llama Guard models (1B and 8B), which are fine-tuned for content safety classification. As shown in Tables 7 and 8, the 1B and 8B models report a Macro F_1 score of 0.49 and 0.54 respectively, placing them respectively in the 14th and 6th position among our evaluated models. This behavior can be linked to the broader content safety taxonomy¹⁰ which was used to train the models, suggesting that for specific moderation tasks where multiple nuances can exist, general purpose LLMs may perform better.

D.2 Inter-Annotator Agreement

To evaluate the robustness of model predictions, we measured inter-annotator agreement across all LLMs using Fleiss’ kappa for both the baseline protocol and the protocol including the annotation guidelines. The baseline protocol returned a Fleiss’ Kappa of 0.020, whereas the protocol augmented with the guidelines resulted in 0.013. Both values are near zero, indicating that the models widely disagree across many tweets regardless of the variation introduced in the prompt. Furthermore, a lower value for the protocol augmented with guide-

¹⁰Examples of categories include hate, violent crimes, sexual content and elections.

lines can indicate that the guidelines restricted the models and that they may have been able to classify correctly some instances by relying on their general knowledge. We complement this analysis in the following section (see Section D.3) by inspecting closely a sample of predictions from our top-performing model.

D.3 Error Analysis

To inspect how models can shift between the baseline protocol and the protocol augmented with the guidelines, we conduct an error analysis with a balanced random sample¹¹ of 30 non-sexist and 30 sexist Tweets¹² classified by Qwen2.5-7B-Instruct for both protocols, our top-performing model in terms of Macro F_1 .

Table 9 summarizes the agreement and misclassification counts across both protocols, including cases where their predictions diverged. The main focus of our qualitative analysis is about borderline cases, where a content reporting sexism may be confused with sexist or non-sexist content. For the remaining cases, annotation by professionals acquainted with the linguistic nuances related to sexist speech, is needed and this task is left for future research.

Among the 39 identical predictions, 35 were correctly classified in both protocols. Only 4 instances were misclassified by both. Among those, two tweets corresponded to content which reported sexism in a very subtle way via indirect and non-aggressive critique.

Furthermore, we find that 10 instances that were misclassified in the baseline protocol, became correctly classified when the annotation guidelines were included in the prompt. Among those instances, only two out of 10 were originally classified as non-sexist under the baseline protocol, and a manual inspection shows that both are tweets reporting sexism. Hence, the guidelines helped the model classify tweets reporting sexism as sexist, since they explicitly state that such tweets should be annotated as sexist.

However, we find that 11 instances that were correctly classified under the baseline protocol, were misclassified when the annotation guidelines were provided to the model. These numbers suggest

¹¹Seed 2026 was used to ensure reproducibility.

¹²Among the 30 sexist tweets, 15 correspond to direct sexism and 15 criticise or report sexist content / sexism in general, which belong to the positive class (sexist) following the annotation guidelines, see section A.1.

Table 7: Macro F_1 of the **task comprehension protocols**

model	Baseline Eng. prompt	Eng. Annotation Guidelines + Eng. prompt	Fr. Annotation Guidelines + Fr. prompt	Eng. Annotation Guidelines + Eng. $\neg$ prompt	Fr. Annotation Guidelines + Fr. $\neg$ prompt
Llama-Guard-3-8B	0.54	0.56	0.58	0.52	0.51
Llama-Guard-3-1B	0.49	0.47	0.42	0.50	0.48

Table 8: Macro F_1 of the **language sensitivity protocols**.

model	Eng. prompt Eng. Labels	Fr. prompt Fr. Labels	Fr. prompt Eng. labels	Eng. prompt French Labels
hline Llama-Guard-3-8B	0.54	0.51	0.43	0.51
Llama-Guard-3-1B	0.49	0.37	0.36	0.48

Table 9: Error Analysis of a random sample of 60 predictions by Qwen2.5-7B-Instruct for the baseline protocol and the protocol augmented with annotation guidelines.

Category	Instances
Correct classification for both protocols	35
Misclassification for both protocols	4
Correctly classified with guidelines	10
Become misclassified with guidelines	11

that guidelines do help the models classify a few instances correctly, but equally confuse them on others, leading to new errors.

A manual inspection shows that the only two instances out of the 11, that were correctly classified as non-sexist in the baseline protocol and classified as sexist when guidelines were provided, actually correspond to tweets that report sexism, possibly indicating errors committed by human annotators. We also find three cases where the baseline protocol correctly classified the tweets as being sexist (positive class) and in fact they were reporting sexism. But when the annotation guidelines were introduced, the predictions became non-sexist. Two out of three of these instances did report sexism in the form of an ironic tweet, possibly indicating that the model failed to grasp the meaning. The third instance included a broad critique of sexism and the hashtag #moiaussi (the French equivalent of #Metoo), but did not have a reporting verb or quotation marks as indicated in the annotation guidelines, suggesting that the model may have restricted the understanding of broader sexism to follow the guidelines.