Assignment 2

Battistoni Leandro, Ponzuoli Giovanni and **Rimondi Simone**
Master's Degree in Artificial Intelligence, University of Bologna
{ leandro.battistoni, giovanni.ponzuoli, simone.rimondi5 }@studio.unibo.it

Abstract

This work investigates the use of large language models (LLMs) for multi-class text classification through prompt-based learning, focusing on zero-shot and few-shot prompting without model fine-tuning. Two open-weight models are compared to analyze how different prompt designs affect their behavior and performance. The study emphasizes prompt engineering, examining how choices such as example selection, ordering, and instruction detail influence model outputs. The results highlight that prompt design is a critical factor in prompt-based classification and that well-structured prompts can yield substantial performance gains.

1 Introduction

In this study, we address EDOS Task B, which consists of classifying a given text into one of five categories: not-sexist, threats, derogation, animosity, or prejudiced discussion.

Traditional approaches to sexism detection typically rely on supervised learning with task-specific classifiers, often based on fine-tuned transformer models. While effective, these methods require large labeled datasets and may struggle to generalize across domains or evolving language use. Recently, large language models (LLMs) have demonstrated strong zero-shot and few-shot capabilities, enabling task adaptation through prompt engineering without additional training.

In this work, we explore prompt-based inference with open-weight LLMs as an alternative approach. We investigate zero-shot and few-shot prompting strategies, analyzing how prompt structure, example selection and ordering, and reasoning styles—such as Chain-of-Thought prompting—affect model performance. This design enables a systematic analysis of model behavior under controlled prompt variations, rather than optimizing performance through extensive fine-tuning.

2 System description

Our experimental setup evaluates two state-of-the-art open-weight, decoder-only large language models—LLaMA-3.1 (Meta AI, 2024) and Mistral-0.3 (Mistral AI, 2024)—on the EDOS Task B dataset under multiple prompting configurations. Performance is evaluated using F1 Score and Fail Ratio, capturing both classification quality and adherence to the required output format.

The system is implemented as a prompt-based inference pipeline, with models downloaded from the Hugging Face Model Hub (Hugging Face) and used exclusively in inference mode. To accommodate hardware constraints, quantization techniques are applied during model loading and execution, enabling the evaluation of larger models while maintaining manageable memory usage and computational costs.

Given the autoregressive nature of the evaluated models, prompt engineering serves as the primary mechanism for guiding behavior. The experimental design therefore emphasizes controlled prompt variations over stochastic sampling. Few-shot demonstrations are selected deterministically according to predefined criteria, ensuring reproducibility across experiments.

The pipeline is modular, with distinct stages for prompt construction, response generation, output decoding, and metric computation. This structure supports flexible prompt manipulation and systematic ablation studies, allowing the isolation of the effects of example selection and prompt formatting strategies—an important consideration given the subtle distinctions between closely related sexism categories.

3 Experimental setup and results

In the following Table 1, we summarize the performance of best models across different prompting strategies. Zero-shot relies solely on the task

description without any examples. Few-shot introduces a limited set of demonstration examples to guide the model's predictions. Mixed and length-constrained few-shot strategies manipulate the order and size of examples while Chain-of-Thought (CoT) prompts encourage the model to reason step by step, with optional modifications to template length. These strategies represent incremental refinements designed to assess how prompt design and example presentation influence both F1 Score and Fail Ratio (FR).

Table 1: Model Performaces Comparison

Strategy	**LLaMA-v3.1**		**Mistral-v0.3**	
	F1	**FR**	**F1**	**FR**
Zero-shot	0.476	0.080	0.396	0.007
Few-shot	0.490	**0.017**	0.527	0.010
Few-shot mixed	0.502	0.050	0.532	0.107
Few-shot mixed length	**0.523**	0.033	**0.555**	0.037
CoT	0.410	0.140	0.544	0.003
CoT prompt length	0.467	0.227	0.540	**0.000**

4 Discussion

Quantitative Results. The experimental results, based on the best-performing configuration for each strategy, highlight how well-structured prompts combined with concise few-shot examples can substantially influence model behavior relative to the zero-shot baseline.

For LLaMA-v3.1, the inclusion of few-shot examples yields modest but consistent improvements in macro F1 Score, indicating that the model already captures the task reasonably well from the prompt alone and benefits incrementally from successive refinement strategies. In contrast, Mistral-v0.3 shows a stronger response to few-shot prompting, surpassing LLaMA's performance with a simple few-shot setup, which suggests a greater reliance on example-driven learning rather than on the task description alone.

Chain-of-Thought (CoT) prompting affects the two models differently: Mistral exhibits a small decrease of approximately 1% in F1 while reducing the Fail Ratio to zero, whereas LLaMA experiences a degradation in both metrics.

Error Analysis and Limitations. Confusion matrix analysis frequently highlights bias toward certain classes that are overpredicted, degrading model performance. This also reflects the inherent human difficulty of class subdivision, as some text examples are challenging to assign to a single category, without given guidelines, especially between sexist classes.

The analysis shows that Mistral consistently generalizes better as the number of examples per class increases, whereas LLaMA's best-performing configurations typically rely on at most four examples per class. This confirms that LLaMA can already generalize effectively with limited context in zero-shot settings, while the introduction of Chain-of-Thought prompting tends to exaggerate this behavior, often to the detriment of performance.

Some methodological choices provide important context. Deterministic selection of examples, rather than random sampling, enabled controlled comparisons of different manipulation strategies and ensured reproducible results. While selecting an optimal "best" set of examples might further improve performance, our focus was on analyzing the effects of various prompting strategies. Similarly, assigning the *not-sexist* label to failed responses simplified the analysis, though it introduces slight noise, particularly affecting the F1 score.

5 Conclusion

This study investigated prompt-based sexism detection using two different open-weight LLMs under zero-shot and few-shot strategies. The results highlight that the design of the prompt, the selection and ordering of examples, and the adoption of reasoning strategies, such as Chain-of-Thought, can have a substantial impact on model performance. In particular, Mistral-v0.3 shows a strong benefit from few-shot demonstrations, quickly surpassing LLaMA even with a limited number of examples, whereas LLaMA performs relatively well even with minimal context, showing more gradual improvements as additional examples are provided. These findings demonstrate that careful prompt structuring and strategic use of demonstrations can significantly enhance classification accuracy, especially in tasks involving closely related categories. Looking ahead, potential extensions include dynamic selection of demonstrations tailored to each input, exploration of more refined and detailed prompt templates, and multi-step classification strategies that may further improve model performance on challenging examples.

6 Links to external resources

The source code is available at this repository.

References

Hugging Face. Hugging face model hub: Platform for pretrained machine learning models. https://huggingface.co.

Meta AI. 2024. LLaMA-v3.1 (8B Instruct): Instruction Tuned Large Language Model by Meta AI. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.

Mistral AI. 2024. Mistral-v0.3 (7B Instruct) : Instruction Tuned Large Language Model by Mistral AI. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3.