Journal of Information and Data Management, 2026, 17:1
ISSN: 2178-7107 • doi: 10.5753/jidm.2026.7421
This work is licensed under a Creative Commons Attribution 4.0 International License.

When Prompts Know the Story: Improving LLM Detection of Online Hate Speech

Annie Amorim [Universidade Federal Fluminense | *annieamorim@id.uff.br*]
Gabriel Assis [Universidade Federal Fluminense | *assisgabriel@id.uff.br*]
Laryssa Paiva [Universidade Federal Fluminense | *laryssapaiva@id.uff.br*]
Luiz Gustavo Dias [Universidade Federal Fluminense/Instituto Federal Goiano | *lgdias@id.uff.br*]
Aline Paes [Universidade Federal Fluminense | *alinepaes@ic.uff.br*]
Daniel de Oliveira [Universidade Federal Fluminense |*danielcmo@ic.uff.br*]

Universidade Federal Fluminense (UFF), Niterói, RJ, Brazil – Institute of Computing

Abstract. Recent advances in Large Language Models (LLMs) have created new possibilities for automated hate speech detection in online platforms. This manuscript investigates how prompt design choices, *e.g.*, contextualization and domain delimitation, affect LLMs' ability to accurately identify online hateful content. We evaluate a range of prompt formulations, varying the presence of explicit instructions and illustrative examples, to analyze their influence on classification performance across both state-of-the-art commercial models and smaller open-weight alternatives. The results demonstrate that constraining the thematic scope of hate speech through domain delimitation improves detection accuracy by minimizing semantic ambiguity and reducing false positives. Taken together, these findings highlight the importance of carefully designed prompts for improving hate speech detection with LLMs in online environments.

Keywords: LLM, Prompt engineering, hate speech detection, Social Networks

Received: 19 January 2026 • **Published:** 27 November 2026

1 Introduction

The use of social networks has grown exponentially over the last two decades on a global scale [Nguyen *et al.*, 2025]. These platforms have transformed how individuals interact and build relationships, enabling, in particular, connections between people who are geographically distant. However, despite these benefits, social networks have also become environments that ease the dissemination of offensive, discriminatory, and harmful content, especially targeting historically marginalized or underrepresented groups [Kim *et al.*, 2021]. This type of manifestation has been referred to as *online hate speech* [Vargas *et al.*, 2021; Aluru *et al.*, 2020].

Online hate speech involves not only a social but also a technological dimension, making it a multifaceted and challenging problem to address. Despite the increasing attention it has received from the scientific community [Albladi *et al.*, 2025], from organizations such as the United Nations[1], and from some owners of social media platforms themselves, its detection remains an open challenge. The major difficulty lies in the sheer volume of content continuously generated on these platforms, which makes manual detection unfeasible (and subject to bias). Even with the advancement of automated solutions [Saraiva *et al.*, 2021] for identifying online hate speech, the task remains challenging due to factors such as the subjectivity of messages (*e.g.*, figures of speech, irony, emojis) and cultural or linguistic variations.

In addition to the aforementioned challenges, another problem associated with the automatic detection of online hate speech is how to incorporate complementary and supporting information related to the domain of the posted text. Let us consider the comments presented in Figure 1, all of which refer to the movie Emilia Pérez[2]. The texts shown in panels (a) and (b) can initially be interpreted as expressions of opinion. In panel (a), the comments show negative evaluations of the movie, referring to its quality and, possibly, to how Mexico is portrayed in the screenplay; under this initial reading, they can be classified as critical remarks or personal opinions.On the other hand, the examples in panel (b) are more difficult to interpret in isolation, since without additional contextual information, it is unclear who or what they refer to, making their correct classification challenging when relying solely on textual content. However, once it is clear that all four comments concern the film Emilia Pérez, starring the Spanish transgender actress Karla Sofía Gascón[3], the interpretative context changes: the texts in panel (a) remain identifiable as opinions about the movie, its artistic quality, and its relationship to the country in which the story is set, whereas the texts in panel (b), when re-evaluated in light of this context, can be classified as hate speech, as they deny the gender identity of the actress, thereby reproducing a transphobic discourse. The examples also highlight the importance of incorporating complementary data to enable more accurate judgments, particularly when hostility is implicit and highly context-dependent.

In recent years, Large Language Models (LLMs) [Zhao *et al.*, 2023] have emerged as a viable alternative for enabling various classification tasks without the need for training specific models [Dong *et al.*, 2024a], including hate speech detection [Albladi *et al.*, 2025]. Despite their advancements, LLMs remain sensitive to prompt formulation [Anagnostidis and Bulian, 2024], posing an additional challenge, particularly in tasks such as online hate speech detection, where the absence of explicit context can compromise accuracy.

[1] https://brasil.un.org/pt-br/249816-como-combater-o-discurso-de-%C3%B3dio-nas-redes-sociais

[2] https://www.imdb.com/pt/title/tt20221436/

[3] https://www.bbc.com/portuguese/articles/cvg9wlz1v39o

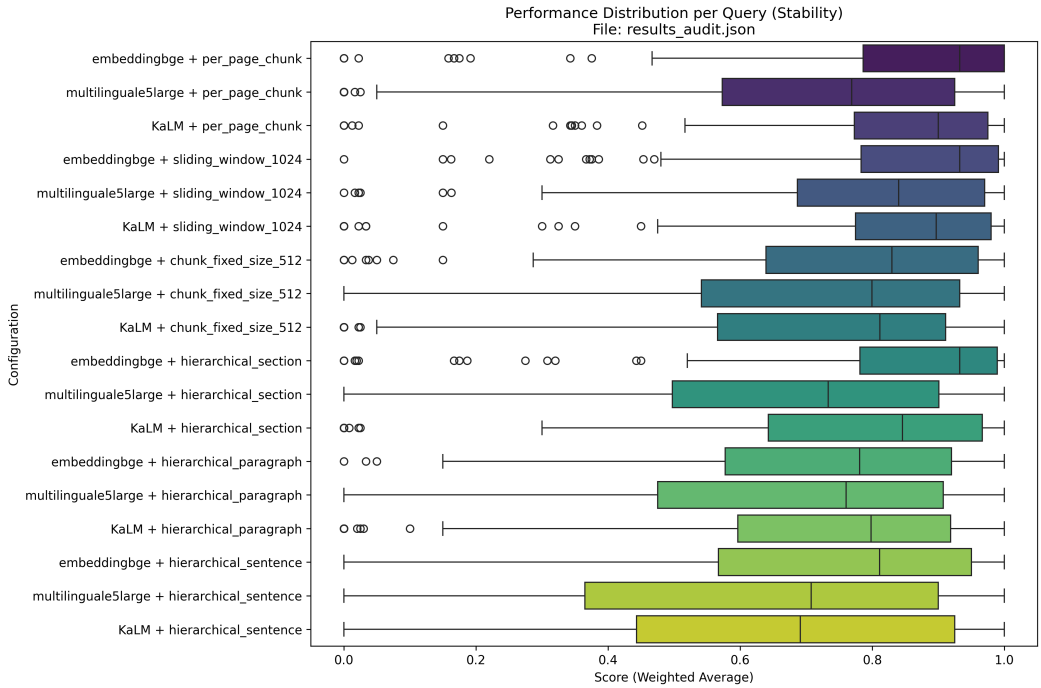

Figure 1. Texts collected from YouTube illustrating the impact of contextual information on message interpretation.

As a result, previous works have explored various prompt engineering strategies [Zhou *et al.*, 2023; Li and Liang, 2021; Lester *et al.*, 2021], such as inserting representative examples in task descriptions [Dong *et al.*, 2024b] and incorporating personas to guide model behavior [Choi and Li, 2024]. However, there remains a gap in studies examining the impact of incorporating domain-specific supporting information related to the analyzed texts. Given that the interpretation of an online hate message may depend on prior knowledge of events, people, or situations mentioned, it is essential to understand how the presence (or absence) of explicit context affects LLM performance on sensitive tasks such as this one.

In this manuscript, we propose evaluating the impact of inserting supporting information into prompts used with LLMs for the task of detecting online hate speech. To this end, we developed three progressive versions of prompts, each with different levels of additional information, and applied them across three distinct LLM families: (i) Cohere Command, (ii) OpenAI GPT, and (iii) DeepSeek. Our selection spans both smaller open-weight or commercially accessible models and larger proprietary ones. Specifically, we evaluated Command-A and Command-R7B from Cohere, GPT-4o and GPT-OSS-20B from OpenAI, and DeepSeek-V3 and DeepSeek-R1-Distill-14B from DeepSeek. This diversity of architectures and scales allows us to comprehensively evaluate how prompt design influences model behavior across different model families and capacity tiers. The evaluation was conducted using a set of real comments extracted from YouTube videos related to the public reception of the movie Emilia Pérez[4].

The results, both quantitative and qualitative, indicate that including supporting information in the prompts contributes to better interpretation of messages and reduces errors, particularly in identifying hate speech with disguised language, such as that associated with transphobia. These findings highlight the importance of contextualization as a prompt engineering strategy for LLMs, with implications for fairer and more socially aware systems.

This manuscript[5] is an extended version of the work presented in [Paiva *et al.*, 2025]. More specifically, it broadens the previous analysis by examining the effect of inserting supporting information into prompts used with LLMs for smaller open-weight models, allowing an assessment of whether the findings previously observed for larger and commercial models also generalize to this setting. The remainder of this paper is organized as follows: Section 2 presents the theoretical framework necessary to understand the study. Section 3 discusses related work. Section 4 details the adopted methodology. Section 5 analyzes the results obtained from both qualitative and quantitative perspectives. Finally, Section 6 presents the conclusions.

2 LLMs and Prompt Engineering

In this section, we introduce and contextualize key concepts underlying the discussions and experiments in this manuscript. Since the introduction of the Transformer architecture [Vaswani *et al.*, 2017], neural network-based language models have achieved remarkable advances. Among them, LLMs stand out for their ability to interpret and generate text with high semantic coherence. Architecturally, these models derive from the decoder component of the original Transformer and differ not only in scale — with billions of parameters — but also by exhibiting emergent abilities, that

[4] `https://www.imdb.com/pt/title/tt20221436/`

[5] **Author's Note:** This manuscript contains examples of explicit offensive content and hate speech, used exclusively for the purposes of illustration and analysis of the issues discussed. These examples do not, under any circumstances, reflect the personal opinions or positions of the authors or their institutions.

is, complex behaviors that do not appear in smaller models [Paes *et al*., 2024].

The development of models such as GPT-3 and GPT-3.5 [Brown *et al*., 2020], which served as the foundation for the first versions of ChatGPT[6], highlighted the potential of these systems to perform a wide range of tasks, including classification, translation, and summarization, even without task-specific supervised training. This versatility is associated with the phenomenon known as *in-context learning* [Dong *et al*., 2024b], in which the task is presented to the model through a natural-language prompt accompanied by examples. In this approach, the prompt serves as a textual instruction guiding the model's output generation. Although its structure is flexible, the content of the prompt directly affects model behavior, since text generation is grounded in the principles of distributional semantics [Paes *et al*., 2024]. Consequently, prompt formulation directly affects the quality and adequacy of the generated response.

This realization was one of the main catalysts for the emergence of the field now known as *prompt engineering*, whose central goal is to investigate and develop textual structuring strategies that optimize LLM performance on specific tasks [Liu *et al*., 2023]. Various techniques have been proposed in this domain, including the insertion of representative examples within task descriptions [Dong *et al*., 2024b], the incorporation of personas to guide model behavior [Choi and Li, 2024], and the decomposition of reasoning into successive steps, as in the chain-of-thought prompting (CoT) approach [Wei *et al*., 2022].

Such strategies have proven effective in several contexts and applications. However, selecting and implementing the most appropriate approach for each scenario remains an open challenge and an active area of research [Liu *et al*., 2023]. Thus, although LLMs hold great potential for improving various tasks, their application requires careful, conscious use, particularly in sensitive domains such as online hate speech detection. This task, in particular, involves not only individual subjectivity but also nontrivial contextual and cultural aspects that demand special attention during modeling and evaluation.

3 Related Work

LLMs have been increasingly employed for detecting and classifying online hate speech, establishing themselves as a promising tool in this domain [Albladi *et al*., 2025]. One of the key reasons for this popularity lies in their ability to capture and interpret textual input content, providing an advantage over traditional approaches. Although most existing solutions have focused on English, recent advances have also been reported for other languages, such as Spanish [Pérez *et al*., 2025] and Portuguese [Assis *et al*., 2024b].

Chiu *et al*. [2022] presents a study using GPT-3 to identify content with racist and sexist undertones. In that work, the authors explore prompt-based approaches containing illustrative examples to guide the model in detecting offensive discourse. However, Chiu *et al*. [2022] does not explore the use of supporting information within the generated prompts. Also leveraging GPT-family models, Li *et al*. [2024] examines ChatGPT's performance in analyzing comments posted on digital platforms and classifying texts under the following taxonomy: hate speech (Hate), offensive Content (offensive), and Toxic Language (Toxic), grouped under the acronym HOT. The study evaluates five distinct prompt variations, in which the model is instructed to indicate, in binary form, the presence or absence of each HOT category for a given comment. The variations differ in complexity, including or excluding model-generated justifications, probability estimates, and varying degrees of contextualization. The results reveal that model performance is influenced by prompt formulation, indicating that the choice of an appropriate instruction should consider the task's specific goals. However, Li *et al*. [2024] limited their investigation to ChatGPT, without exploring other LLMs.

Similarly, Guo *et al*. [2024] utilized ChatGPT in various sensitive contexts, including detecting sexism, xenophobic expressions, and political discourse during election periods. The authors compared four distinct prompt strategies: (i) simple instructions, (ii) inclusion of conceptual definitions, (iii) contextualized examples, and (iv) the chain-of-thought reasoning technique. Their results reinforce the idea that model performance strongly depends on the structure of the prompt, with the most effective formats providing illustrative examples and inducing logical reasoning.

Oliveira *et al*. [2023] evaluated two different prompt configurations applied to GPT-3.5 for toxic text detection. Their findings indicate that more detailed and explanatory prompts tend to yield superior performance, supporting the hypothesis that richer instructions improve model effectiveness. Reports similar findings Oliveira *et al*. [2024a]; Assis *et al*. [2024b], who compared the performance of GPT and a Brazilian chatbot, Maritalk[7]. Specifically, Assis *et al*. [2024b] investigated the impact of including illustrative examples and contextual keywords in prompts, finding that these elements contribute to more balanced performance. Moreover, these studies compared LLMs with traditionally trained classifiers, highlighting the competitiveness of prompt-based models even without supervised fine-tuning for the specific task.

In addition to the GPT-family models, open-source LLMs have also been the focus of investigation. In particular, Assis *et al*. [2024a] and Oliveira *et al*. [2024b] explored variants of the LLaMA architecture [Touvron *et al*., 2023], applying explicit weight adjustments via fine-tuning. However, the approaches adopted in these studies differ from each other and from the present work in relevant methodological aspects. Assis *et al*. [2024a] adapted models directly to act as supervised classifiers, following the traditional training paradigm with annotated labels. In contrast, Oliveira *et al*. [2024b] introduced an initial stage of prompt engineering, testing different strategies to identify those that induced the best preliminary performance. Only after this exploratory phase was fine-tuning performed for hate speech classification, thereby illustrating the integration of prompt-based and supervised learning strategies.

The studies discussed above demonstrate that prompt formulation directly influences LLM performance in hate

[6]https://chatgpt.com/

[7]https://chat.maritaca.ai/

speech detection and classification tasks. Based on this context, we investigate the impact of including additional contextual information in prompts, with a focus on incorporating event-related information linked to the analyzed posts. The central hypothesis is that such information can help models make more accurate inferences, particularly in ambiguous cases or those requiring world knowledge. To this end, we evaluate not only widely used models such as GPT-4o [OpenAI, 2024], but also less commonly explored LLMs in hate speech research, such as DeepSeek-V3 [DeepSeek-AI, 2025b] and Command-A [Cohere, 2025]. Beyond evaluating the larger versions of these models, we also included, within each family, a smaller open-weight variant, allowing us to investigate the impact of the prompting approach across different setups and available computational resources. The experiments we conducted utilized a dataset developed explicitly for this study, which was collected, reviewed, and manually annotated to ensure data quality and representativeness. Thus, the results presented here offer new insights into the application of LLMs in sensitive contexts, grounded in carefully curated data and prompting strategies informed by event-related information.

4 Methodology

The methodology we adopted examined how varying levels of supporting information included in prompts influence LLM performance on online hate speech detection. To this end, we structured the methodological framework into four main stages: (i) collecting and preparing a dataset of real posts representative of potentially offensive or discriminatory situations; (ii) developing progressively refined prompt versions that vary in the amount of supporting information, enabling a comparative evaluation of their effects on model performance; (iii) manually annotating the posts to build a reference set sensitive to the nuances of the analyzed context; and (iv) applying the selected models and assessing their outputs through both quantitative and qualitative analyses.

4.1 Data Collection and Preparation

We extracted comments from a single YouTube video to constrain context while enabling a deeper understanding of the texts submitted for evaluation. The selection of the video was based on specific criteria that emphasized both its relevance and the density of user interactions. Initially, we conducted a search using the query "`karla sofia in:english`" to locate content related to the actress Karla Sofía Gascón that could elicit diverse audience reactions. The actress was chosen because she became the target of online attacks during the Oscar campaign for the film Emilia Pérez, following the resurfacing of past tweets containing prejudiced opinions. Among the search results, several videos were identified, including interviews, award ceremonies, and news reports discussing controversies involving the actress on the social network X[8].

The final criterion for selecting the video considered both the number of views and the volume of available comments. The chosen video was an interview given by Karla Sofía Gascón on The Tonight Show Starring Jimmy Fallon, titled "Karla Sofía Gascón Fangirled Over Harrison Ford and Mark Hamill, Talks Emilia Pérez Transformation"[9]. The video was published on January 14, 2025, and has a runtime of 10 minutes and 21 seconds. By the time we collected the comments — on March 14, 2025 — the video had accumulated 243,394 views and 1,425 comments. We performed data collection using the YouTube API[10], resulting in a random sample of 851 comments, which we subsequently annotated and analyzed.

After collection, we manually filtered the data to ensure data quality. We excluded all comments written in languages other than English, including those containing mixed-language segments or expressions. We also discarded comments composed solely of emojis. After filtering, the final dataset comprised 604 comments. Among these texts, three comments we selected to serve as examples in the prompts — one per class (neutral - NEU, offensive - OFF, and hate speech - HS) — following the taxonomy adopted by [Assis *et al.*, 2024b]. The goal was to preserve semantic and thematic proximity with the analyzed corpus. Selection was performed randomly until representative texts were identified for each class. These example comments were then removed from the primary dataset to prevent potential bias in the experimental results. The selection of examples was conducted by an annotator external to the primary annotation team (Section 4.3) and subsequently validated by all other annotators for category suitability.

4.2 Prompt Construction

We investigated how prompt formulation affects LLM-based hate speech detection by developing three progressively refined prompt versions, all crafted in English to match the language of the comments we collected. Each version introduced an incremental level of supporting information, enabling an assessment of how such variation affects model performance. The first version (P1) consisted of a brief task description followed by the class definitions, instructing the LLM as follows:

Base Prompt

You will receive a list of YouTube comments to classify into one of the categories: 'hate speech', 'offensive', or 'neutral'. Definitions are provided below. Please carefully consider all elements during classification.
Hate Speech: Abusive or threatening expressions that convey prejudice against individuals or groups based on Health, Ideology, LGBTQphobia, Lifestyle, Physical Appearance, Race (Racism), Religion (Religious Intolerance), Gender (Sexism), or National Origin (Xenophobia).
Offensive: Contains profanity, obscene language, or insults, but does not explicitly express prejudice.
Neutral: Does not contain hate speech or offensive language.

[8] `https://www.cnnbrasil.com.br/entretenimento/karla-sofia-gascon-desativa-conta-no-x-antigo-twitter-apos-polemicas/`

[9] `https://www.youtube.com/watch?v=l-akpQZD5pk`

[10] `https://github.com/onlyphantom/youtube_api_python`

The second version (P2) retained the same structure as P1, comprising the task description and class definitions, but also introduced representative examples for each class under evaluation. The inclusion of these examples provided the models with more concrete guidance by presenting instances of each expected category, thereby promoting a more accurate understanding of the classification task. The selected examples were as follows:

(i) **Hate speech**: "*It is not need for a nuclear war, not need for AI to take control of humanity or a virus to kill us, the fact that a man pretends to be a woman, it will be the end of humanity. No reproduction that is the key*".
(ii) **Offensive**: "*This guys are nuts*".
(iii) **Neutral**: "*I'm not sure he's concerned with your opinion. The audience seemed to enjoy it*".

The third version (P3) further expanded the level of detail provided to the LLM by incorporating additional supporting information beyond that in previous versions. In this version, we included a concise description of the central plot of the film Emilia Pérez, information about the main cast, and clarifications regarding the origin of the analyzed comments, including the video from which they were extracted. The goal of this more contextually enriched prompt was to provide the LLMs with relevant background information to help them interpret the comments. The supporting text was as follows: "*The comments refer to YouTube videos discussing the film Emilia Perez, directed by Jacques Audiard and starring Karla Sofía Gascón, Selena Gomez, and Zoe Saldaña. Karla Sofía Gascón is a transgender actress. The film tells the story of a Mexican cartel leader who undergoes gender reassignment surgery to escape his criminal past and reinvent himself. The comments analyzed were taken from the video interview entitled 'Karla Sofía Gascón Fangirled Over Harrison Ford and Mark Hamill, Talks Emilia Pérez Transformation', available on YouTube (https://www.youtube.com/watch?v=l-akpQZD5pk)*".

4.3 Data Annotation

The data annotation process was carried out in two stages. The first stage corresponded to the initial annotation phase, during which at least two annotators independently labeled each comment. To ensure consistency, all annotators had access to the same set of instructions provided to the LLMs, including definitions of the three classes, illustrative examples, and supporting information from the most detailed version of the prompt (P3). However, each annotator viewed only their assigned subset of texts, without access to the labels assigned by others.

All annotators had previously watched the film Emilia Pérez as well as the specific video from which the comments were extracted. The annotation team consisted of six participants, three women and three men, aged between 20 and 35. All participants were undergraduate or graduate students, and most had prior research experience in analyzing hate speech in digital environments. Given the sensitive nature of the themes addressed in both the film and the comments, the team was intentionally diverse, including participants who identify as LGBTQIAP+, to ensure that sociocultural nuances were appropriately considered.

The results were compared after individual annotation. Among the 601 annotated texts, disagreements were found in 127 cases. Thus, the second stage of the annotation process involved adjudication by a judge, who compared the initial annotators' labels and comments before assigning the final label for each disputed comment. The judge was limited to selecting one of the labels initially proposed by the annotators. The judge also openly identifies as a member of the LGBTQIAPN+ community. At the end of the annotation process, 310 texts were labeled as NEU, 129 as OFF, and 162 as HS.

The reliability of the annotation process was further evaluated using Cohen's Kappa, a widely adopted metric for nominal classification that accounts for the agreement expected by chance. Although the two independent annotations were not necessarily produced by the same fixed pair of individuals, but rather by annotators distributed across two distinct annotation streams, the use of Kappa remains methodologically appropriate, since its application requires only that each item receive two independent labels, a condition that is fully satisfied in this study. The initial agreement between the two annotation streams yielded a Kappa of 0.6651, indicating substantial concordance and reflecting consistent alignment between annotators, even in the presence of inherently ambiguous cases common in hate speech classification. We also measured agreement between each annotation stream and the final adjudicated label, yielding Kappa values of 0.9050 for group 1 and 0.7535 for group 2, both indicating high concordance. Taken together, these results demonstrate that both annotation streams contributed meaningfully to the overall quality of the dataset, and that the adjudication process played a central role in consolidating reliable final labels.

4.4 Model Application

The proposed prompting strategies were evaluated across models that can be divided into two distinct groups. The first group comprises three large, state-of-the-art LLMs: (i) Command-A [Cohere, 2025], (ii) GPT-4o [OpenAI, 2024], and (iii) DeepSeek-V3 [DeepSeek-AI, 2025b]. The second group, conversely, utilizes smaller, open-weight variants derived from similar model families, namely: (iv) Command R7B[11] [Cohere, 2025], (v) GPT OSS 20B [OpenAI, 2025], and (vi) DeepSeek-R1 14B [DeepSeek-AI, 2025a]. This selection allows us to evaluate the impact of the prompting strategies on high-performance models, such as those in the first group, and also assess whether their impact generalizes to smaller variants, that is, the second group. Therefore, we aim to investigate not only the intrinsic effect of prompting strategies but also their robustness across heterogeneous computational and architectural settings, a factor crucial for implementation and deployment decisions.

Due to cost constraints on accessing the GPT-4o and DeepSeek V3 APIs, we submitted prompts through the respective public chat interfaces on the developers' platforms. In both cases, for practical reasons, seven distinct comments were submitted per interaction for each evaluated prompt version. For the Command-A LLM, the interaction was conducted via Cohere's official API, made possible through aca-

[11] `https://docs.cohere.com/docs/command-r7b`

demic research credits. This access enabled configuration of inference parameters, specifically the `temperature`, which was set to 0.3. This value was chosen based on previous studies [Assis *et al.*, 2024b], as it offers a good balance between reducing response variability and increasing consistency with the task's target classes. All LLMs were tested with three progressively detailed *prompt* versions (P1, P2, and P3), enabling a comparative analysis of how variations in instruction detail affect the models' classification performance. This approach enables an evaluation of the extent to which the informational content of the prompts influences the sensitivity and accuracy of LLMs in recognizing sensitive content.

On the other hand, the three smaller models were run locally using LlamaIndex integrated with the Ollama[12] ecosystem, within a dedicated L40S GPU environment. To ensure reproducibility and consistency across evaluations, all models were configured with a fixed set of inference parameters — the same used for Command-A — including a `temperature` of 0.3. This setup enables us to evaluate the feasibility of applying the proposed prompting strategies in resource-constrained settings, a feature particularly relevant for scenarios that rely on open-weight models rather than commercial APIs. In this way, the locally conducted experiments complement those performed with large proprietary models, extending our understanding of how prompt design generalizes across architectures and deployment contexts.

5 Experimental Evaluation

The experimental evaluation examined how the LLMs performed when exposed to the different prompt versions developed in this study. Both quantitative and qualitative analyses were conducted in a complementary manner to assess how variations in the prompts' supporting content affected the models' ability to classify comments into the evaluated categories — again, hate speech (HS), offensive (OFF), and neutral (NEU).

5.1 Quantitative Results

We used four traditional performance metrics for classification tasks in the quantitative evaluation: accuracy, precision, recall, and F1-score. Additionally, the percentage variation in F1 was calculated relative to the simplest prompt version (P1), which served as the baseline. Given the dataset's class imbalance, the F1-score was adopted as the primary performance indicator. Tables 1 and 2 summarize the results obtained for each of the evaluated LLMs — larger models and smaller variants, respectively — across the three progressive prompt versions (P1, P2, and P3). Overall, in both cases, the results indicate that the P3, which cumulatively incorporates class definitions, representative examples, and contextual support information about the event, yielded the best performance across all tested LLMs. This pattern suggests that including relevant supporting information within the prompt enhances the models' ability to distinguish between neutral, offensive, and hate-related content.

For the large-scale models presented in Table 1, specifically for the Command-A model, a consistent improvement in F1 was observed, rising from 0.582 with P1 to 0.676 with P3, representing a 16.15% relative gain. In contrast, GPT-4o exhibited a nonlinear trend: the isolated inclusion of examples in P2 led to a 4.25% drop in F1 performance compared to the baseline. However, when contextual information was also added (P3), the model achieved an 11.57% increase relative to P1, resulting in the highest overall F1 score (0.761) among all evaluated scenarios. This finding reinforces the hypothesis that the inclusion of contextual support information plays a key role in classification effectiveness.

Table 1. Evaluation results for the LLMs (larger variants).

Model	Prompt	Accuracy	Precision	Recall	F1	Δ F1 (%)	Hate-F1
Command-A	P1	0.677	0.635	0.587	0.582	-	0.397
	P2	0.681	0.645	0.583	0.589	+1.20%	0.446
	P3	**0.742**	**0.724**	**0.664**	**0.676**	**+16.15%**	**0.620**
GPT-4o	P1	0.727	0.705	0.687	0.682	-	0.604
	P2	0.705	0.679	0.659	0.653	-4.25%	0.552
	P3	**0.782**	**0.765**	**0.775**	**0.761**	**+11.57%**	**0.776**
DeepSeek V3	P1	0.649	0.637	0.639	0.622	-	0.584
	P2	0.687	0.677	0.682	0.662	+6.43%	0.621
	P3	**0.722**	**0.717**	**0.727**	**0.705**	**+13.34%**	**0.733**

Examining Table 2, which presents the results of the smaller models, a pattern similar to that observed in the large-scale models becomes evident: prompt P3 consistently yields the best performance across all systems. For Command R7B, the use of P3 resulted in a substantial improvement of 53.04% over P1, increasing the F1 score from 0.347 to 0.531 — a pronounced gain derived from the inclusion of more complete and contextualized instructions. In the case of GPT OSS 20B, the initial performance achieved with P1 (F1 = 0.571) already surpassed the best configuration of Command R7B. However, mirroring the behavior of its larger counterpart, GPT-4o, prompt P2 resulted in a substantial 21.74% drop relative to P1. This outcome reinforces the observation that introducing examples without sufficient supporting information can introduce ambiguity and impair model performance, depending on the model. Beyond this, it highlights a recurrent and stable behavioral pattern associated with our prompt design. When contextual information was incorporated in P3, the model achieved an F1 score of 0.694, corresponding to a 21.56% increase over P1.

DeepSeek-R1 14B displayed modest improvements, with gains of 4.75% in P2 (F1 = 0.552) and 9.68% in P3 (F1 = 0.578). Nevertheless, across all smaller models — as with the large-scale ones — prompt P3 consistently produced the highest F1 scores, further strengthening our hypothesis that prompts enriched with contextual information play a decisive role in enhancing classification performance.

Given the inherent complexity of detecting hate speech, we conducted a complementary analysis focusing specifically on the F1 for the hate speech class. This analysis aimed to assess the extent to which the different prompt versions affected the LLMs' ability to identify content associated with this category correctly — often characterized by semantic nuances, ambiguity, and subjectivity. The results confirmed the general pattern observed in the overall evaluation, showing performance gains with the most comprehensive prompt (P3). For the Command-A model, the F1

[12] `https://developers.llamaindex.ai/python/framework-api-reference/llms/ollama/`

Table 2. Evaluation results for the LLMs (smaller variants).

Model	Prompt	Accuracy	Precision	Recall	F1	Δ F1 (%)	Hate-F1
Command R7B	P1	0.544	0.352	0.346	0.347	-	0.372
	P2	0.564	0.410	0.414	0.403	+16.16%	**0.467**
	P3	**0.570**	**0.549**	**0.546**	**0.531**	**+53.04%**	0.431
GPT OSS 20B	P1	0.652	0.608	0.570	0.571	-	0.401
	P2	0.653	0.456	0.450	0.447	-21.74%	0.440
	P3	**0.735**	**0.711**	**0.692**	**0.694**	**+21.56%**	**0.638**
DeepSeek-R1 14B	P1	0.622	0.552	0.555	0.527	-	0.265
	P2	0.629	0.559	0.567	0.552	+4.75%	0.334
	P3	**0.655**	**0.614**	**0.595**	**0.578**	**+9.68%**	**0.395**

score for the hate speech class increased from 0.397 with P1 to 0.620 with P3, indicating improved detection capability. The DeepSeek V3 model followed a similar trend, increasing from 0.584 to 0.733. GPT-4o also exhibited a significant improvement, rising from 0.604 to 0.776, reaffirming the positive impact of contextual support information on model performance, particularly in hate speech classification tasks. Among the smaller models, we observe a pattern largely consistent with the larger ones: in GPT OSS 20B, the hate speech F1 increased from 0.401 (P1) to 0.638 (P3), while in DeepSeek-R1 14B it rose from 0.265 (P1) to 0.395 (P3). In both cases, P3 produced the highest performance, indicating that prompts enriched with contextual information systematically enhance hate-speech detection. The only exception within this group was Command R7B, whose best hate speech F1 value emerged with P2 (0.467), outperforming both P1 (0.372) and P3 (0.431). Nevertheless, even for this smaller-capacity model, using the more comprehensive prompt (P3) resulted in a substantial improvement over the baseline configuration (P1), suggesting that contextual support information also benefits compact models, even though their optimal level of instruction richness may differ due to architectural constraints.

These findings support the hypothesis that including explicit supporting information in prompts enhances the accurate interpretation of the communicative intent underlying the analyzed comments. Overall, the results indicate that, beyond improving global performance, adding contextual information to prompts is especially beneficial for the most sensitive and challenging class — i.e., hate speech —, contributing to a more robust and socially responsible classification. Compared with large variant models, the best overall result was again achieved by GPT-4o using the P3, yielding the highest F1 across all evaluated scenarios (0.761). Within the more minor model variants, GPT OSS 20B also reached its best performance with prompt P3, obtaining an F1 score of 0.694. Although GPT OSS 20B is allegedly substantially smaller than GPT-4o, the performance difference between the two models in P3 is only 0.067, whereas in P1 it was 0.111. This narrowing gap suggests that, despite capacity limitations, well-designed prompts can help smaller models approach the performance of frontier-scale systems — a particularly relevant observation in contexts where the high computational and financial cost of larger models restricts their practical use.

In addition to improvements in F1, consistent gains were also observed across the other evaluation metrics when using P3. In terms of accuracy, all six LLMs achieved their best performance with this version: Command-A reached 0.742, GPT-4o achieved 0.782, DeepSeek V3 recorded 0.722, Command R7B reached 0.570, GPT OSS 20B obtained 0.735, and DeepSeek-R1 14B achieved 0.655. These results indicate a reduction in overall misclassification rates, showing that the inclusion of contextual information benefits not only specific classes but also enhances the general reliability of the models. Similarly, both precision and recall improved, demonstrating that the models became simultaneously better at correctly identifying sensitive-category texts (with fewer false negatives) and at avoiding incorrect classifications (with fewer false positives). For instance, GPT-4o's recall increased from 0.687 with P1 to 0.775 with P3, while its precision rose from 0.705 to 0.765. Similarly, in GPT OSS 20B, recall increased from 0.570 with P1 to 0.692 with P3, while precision rose from 0.608 to 0.711.

5.2 Qualitative Results

We conducted a qualitative analysis of the classifications produced by the LLMs to gain a deeper understanding of how prompts influence model behavior beyond quantitative results. Given the substantial differences in scale, architecture, and representational capacity between the larger models and their smaller open-weight counterparts, the results are presented separately. This separation enables the identification of patterns in error types, sensitivity to contextual cues, and reliance on supporting information across the families of models.

5.2.1 Larger Variants LLMs

The qualitative analysis focused on comments labeled as hate speech that the models correctly classified only when using P3. The examples shown — highlighted in comments 1 to 6 of Table 3 — illustrate the challenges faced by LLMs in detecting manifestations of transphobia, particularly in short and subtly hostile texts. In such cases, seemingly neutral terms like 'he', 'man', and 'actor' may be employed with derogatory intent, undermining the gender identity of the individual mentioned. When evaluated out of context, these expressions are often misinterpreted by the models as neutral or, at most, OFF. However, human annotators, drawing on a broader understanding of the sociocultural context in which the comments were produced, classified them as HS, recognizing the implicit discriminatory elements in the texts.

Another relevant aspect concerns the misclassification of neutral content as HS, which could result in unjust censorship of legitimate opinions. This issue is illustrated by comments 7 and 8 in Table 3, where the Command-A model, when using P1 and P2, incorrectly assigned the hate speech label to neutral texts. The correct interpretation was achieved only with version P3. A similar pattern was observed in comments 9 and 10, analyzed by DeepSeek V3, showing the recurrence of such errors when LLMs operate without access to contextual cues that help disambiguate tone and intent. In these situations, the absence of supporting information leads LLMs to overestimate hostility, confusing legitimate criticism—directed at the artistic or cinematic quality of the work—with discriminatory speech. Thus, providing contextual information assists on two complementary fronts: it reveals implicit aggression that would otherwise go unnoticed and, at the same time, preserves legitimate critical discourse by preventing it from being mislabeled as HS.

Table 3. Classifications produced by larger LLM variants across prompts, compared to the gold label of each text. NEU denotes neutral, OFF offensive, and HS hate speech.

Comment		Command-A			GPT-4o			DeepSeek V3			Label
ID	**Text**	**P1**	**P2**	**P3**	**P1**	**P2**	**P3**	**P1**	**P2**	**P3**	
1	**He*	NEU	NEU	HS	NEU	NEU	HS	NEU	NEU	HS	HS
2	*Actor not actress.*	NEU	NEU	HS	NEU	NEU	HS	NEU	NEU	HS	HS
3	*It's a man.*	NEU	NEU	HS	OFF	NEU	HS	NEU	NEU	HS	HS
4	*Carlos Gascon is a man*	NEU	NEU	HS	OFF	NEU	HS	NEU	NEU	HS	HS
5	*It's a man in a dress speaking Spanish*	OFF	NEU	HS	OFF	NEU	HS	OFF	NEU	HS	HS
6	*Left, not a woman; right, not a bunny.*	NEU	NEU	HS	NEU	NEU	HS	NEU	NEU	HS	HS
7	*Emilia Perez is an insult not only for México, but for all Latin America.*	HS	HS	NEU	OFF	OFF	OFF	OFF	OFF	OFF	NEU
8	*Absolutely hate Emilia Perez, an insult to Mexico.*	HS	HS	NEU	HS	OFF	OFF	OFF	OFF	OFF	NEU
9	*@USER The criticism of this film has not much to do with Latino representation. As about trans...this movie does a horrible job for them. Sorry. Most trans hate this movie.*	NEU	NEU	NEU	NEU	NEU	NEU	HS	HS	NEU	NEU
10	*Is there any Mexican who actually liked that movie?*	NEU	NEU	NEU	NEU	OFF	OFF	HS	HS	NEU	NEU
11	*Another drag queen*	NEU	NEU	OFF	OFF	OFF	HS	HS	HS	HS	HS
12	*Hate from Spain to him*	HS	HS	HS	HS	NEU	OFF	HS	HS	HS	HS
13	*A woman should have won that award,have they learnt nothing? Oh that's right it's Hollywood.*	OFF	OFF	HS	HS	NEU	HS	OFF	OFF	OFF	HS

Overall, across only P3 results, the Command-A model had the highest number of errors, misclassifying 27 texts correctly labeled by the remaining LLMs. Most of these cases were marked as OFF, as illustrated by comment 11. In contrast, GPT-4o and DeepSeek V3 each made only one error, in comments 12 and 13, respectively. In comment 12, the subtle presence of transphobia — suggested by the use of the pronoun 'him' — made correct identification difficult without contextual information. Comment 13, in turn, contained an implicit critique related to gender issues in the movie industry, requiring a more nuanced interpretation to be recognized as hate speech — something achieved only when supporting information was provided. These results reinforce that even with more detailed prompts, hate speech detection remains challenging, particularly when prejudice is expressed implicitly. Including contextual support information about the events proves crucial not only for improving LLM interpretation but also for reducing both false negatives (missed HS) and false positives (misclassified legitimate commentaries)

Including supporting information in P3 enabled the LLMs to make more context-aware judgments. By providing details about the film, the actress's profile, and the origin of the analyzed comments, the models were able to capture previously overlooked nuances, leading to more accurate classifications. In contrast, when the same comments were analyzed with P1 or P2 — both of which contained less supporting information — the models frequently produced incorrect labels, most often classifying the texts as NEU.

5.2.2 Smaller Variants LLMs

Once again, some texts in the hate speech category were correctly classified by the models exclusively when using prompt P3, while at least one of the other prompts failed in these same cases, as shown in Table 4, for example, Comments 1 to 5. In both instances, Command R7B with P1 was unable to identify the hate speech content. In the case of Comment 1 — "*They're giving it awards... because... trans.*" — both Command R7B and DeepSeek-R1 14B labeled the comment as neutral. Although the token "*trans*" in isolation does not necessarily function as an explicit marker of hostility, the utterance as a whole conveys a depreciative critique directed at a vulnerable group, implying that the award was granted solely because of the actress's gender identity. The ellipses reinforce a sarcastic and delegitimizing tone, features commonly associated with implicit forms of transphobia. This type of nuance, which requires recognizing indirect discursive strategies, is often overlooked by models when supporting information is not provided. Without additional context, LLMs may struggle to detect the underlying discriminatory intent, classifying the comment as neutral or merely offensive rather than correctly identifying it as hate speech. Another particularly concerning example is Comment 3, in which most models using prompts P1 and P2 misclassified a comment clearly labeled as hate speech as merely OFF. The utterance contains multiple explicit markers of transphobia, including the denial of the individual's gender identity ("sounds like a male", "That's a man") and the use of the derogatory term "Freak". In this case, only prompt P3 enabled all models to correctly identify the content as hate speech.

Another relevant aspect observed in comments 6 to 11 concerns cases in which models incorrectly labeled neutral texts as hate speech. While offensive content may, in some instances, share semantic boundaries with HS, neutral texts typically reflect subjective evaluations, non-discriminatory criticism, or ordinary conversational remarks. In such contexts, the absence of explicit hostility makes false positives especially undesirable. A clear example is Comment 10, which offers a harmless aesthetic opinion about the film ("Didn't care for the singing, but the movie was great!"). Despite its evidently neutral nature, DeepSeek-R1 14B with P2 classified the comment as HS, illustrating an excessive sensitivity to potentially harmful phrasing. Command R7B exhibited the highest number of such cases, assigning the hate-speech label to comments that, according to the gold standard, were neutral in both P1 and P2. These errors highlight the importance of more informative prompts, such as

Table 4. Classifications produced by smaller LLM variants across prompts, compared to the gold label of each text. NEU denotes neutral, OFF offensive, and HS hate speech.

Comment		Command R7B			GPT OSS 20B			DeepSeek-R1 14B			Label
ID	**Text**	**P1**	**P2**	**P3**	**P1**	**P2**	**P3**	**P1**	**P2**	**P3**	
1	*They're giving it awards....because....trans.*	NEU	HS	HS	HS	HS	HS	NEU	HS	HS	HS
2	*This is a Gay guy please stop saying he a woman my kids watch TV and we a Christian Family!*	OFF	HS	HS	HS	HS	HS	HS	HS	HS	HS
3	*Jesus! Even in his language he sounds like a male! That's a man! Freak!*	OFF	HS	HS	OFF	OFF	HS	OFF	OFF	HS	HS
4	*she looks post op but the surgery was done in a back alley in tijuana*	HS	HS	HS	OFF	NEU	HS	OFF	HS	HS	HS
5	*As a Mexican I'm disgusted by the fact that a Spanish guy dressed as a woman is playing a role in a "Mexican" movie. I wanna vomit.*	OFF	HS	HS	OFF	HS	HS	OFF	HS	HS	HS
6	*You know what else is disrespectful to Mexican culture ? THE ACTUAL DRUG CARTELS , THE DIRTY AND CORRUPT POLITICIANS!! Keep the eye on the ball*	OFF	HS	NEU	NEU	NEU	NEU	OFF	OFF	NEU	NEU
7	*Gato/a, when used as slang in Puerto Rican Spanish (not referring to the animal), is a term used to describe individuals within a group, such as 'cuatro gatos/gatas.*	HS	HS	NEU	NEU	NEU	NEU	OFF	NEU	NEU	NEU
8	*Which is the other*	HS	HS	NEU	NEU	NEU	NEU	NEU	NEU	NEU	NEU
9	*she looks post op but the surgery was done in a back alley in tijuana*	HS	HS	NEU	NEU	NEU	NEU	NEU	NEU	NEU	NEU
10	*Didn't care for the singing but the movie was great!*	NEU	NEU	NEU	NEU	NEU	NEU	NEU	HS	NEU	NEU
11	*There are tons of video talking that but they are in spanish*	NEU	NEU	NEU	NEU	NEU	NEU	NEU	HS	NEU	NEU
12	*We Spaniards do not hate Mexicans, on the contrary, they are the ones who hate us*	HS	HS	HS	HS	NEU	NEU	HS	NEU	HS	NEU
13	*Us as mexicans can't relate to her awful movie, it is insulting, out of touch and this woman has always been a racist towards mexicans*	HS	HS	HS	HS	HS	HS	HS	HS	HS	NEU
14	*She has been very mean with Mexicans on X*	HS	HS	HS	HS	HS	NEU	HS	HS	HS	NEU
15	*Please stop promoting this awful film. Is an insult to Mexico and the Spanish language*	HS	OFF	HS	HS	HS	HS	OFF	OFF	OFF	NEU
16	*Emilia Perez is an insult not only for México, but for all latinamerica*	HS	HS	HS	OFF	HS	HS	HS	OFF	OFF	NEU
17	**He*	NEU	HS	NEU	NEU	NEU	NEU	NEU	HS	NEU	NEU
18	*He, is a man*	NEU	NEU	NEU	NEU	NEU	NEU	NEU	NEU	OFF	NEU
19	*man*	NEU	HS	NEU	NEU	NEU	NEU	NEU	NEU	NEU	NEU
20	*Is she trans?*	NEU	HS	NEU	NEU	NEU	NEU	NEU	OFF	OFF	NEU

P3, which help models better distinguish legitimate criticism from discriminatory attacks, thereby reducing false positives and avoiding interpretations that could lead to undue censorship of users.

Although prompt P3 achieved the best overall results, important classification limitations were still observed, particularly among the smaller models. In these variants, there were more concerning cases in which P3 incorrectly labeled neutral comments as hate speech. In these examples, not only did P1 and P2 fail, as expected given the lack of supporting information, but P3 also misclassified them, indicating that certain types of errors persist regardless of the level of prompt detail. A common feature among these texts is the presence of terms related to Mexico, Mexicans, or Latin Americans, suggesting that the models tend to automatically interpret such references as potential markers of xenophobia, even when the surrounding context does not convey targeted hostility. This bias directly contributes to the misidentification of HS, as illustrated by comments 12-16. A notable contrast with the larger models is that the smaller variants were unable to identify very short texts as potential cases of hate speech. As shown in comments 17 to 20, expressions such as "*He", "man", or even "Is she trans?" were frequently classified as neutral or OFF, despite containing implicit forms of transphobia reflected in the gold label. Thus, even with P3, these smaller models continue to exhibit structural limitations in detecting implicit hostility.

6 Conclusions

We investigated the impact of incorporating supporting information into LLM prompts for detecting online hate speech. Three progressively refined prompt versions were developed, ranging from basic task instructions (P1) to the addition of illustrative examples (P2) and finally, to the inclusion of detailed contextual information about the event related to the analyzed comments (P3). The experiments were carried out using six language models organized into two distinct groups. The first group comprises the larger-scale variants: (i) Command-A, (ii) GPT-4o and (iii) DeepSeek-V3, while the second group includes the smaller open-weight models: (iv) Command R7B, (v) GPT OSS 20B and (vi) DeepSeek-R1 14B. The dataset consisted of real YouTube comments related to the film Emilia Pérez, collected and annotated with careful manual curation.

The quantitative analysis revealed that the prompt version incorporating supporting information (P3) consistently improved all evaluated metrics. Notably, F1 increased by up to 16.15% relative to the baseline prompt (P1) for the larger models, and by up to 53.04% for the smaller variants. This result suggests that models with more limited parameter capacity rely even more heavily on detailed instructions and explicit contextualization to perform accurate inference, therefore benefiting substantially from more structured prompts. The qualitative analysis complemented these findings, reinforcing that contextual information about events enables LLMs to correctly interpret subtler forms of hate, particularly those related to transphobia, and to distinguish legitimate criticism from discriminatory attacks. Larger models demonstrated a greater ability to identify hate speech, particularly in cases of implicit hostility. Nevertheless, given the practical difficulties of operating such models — whether due to computational cost, restricted access, or usage limitations — the results obtained by the smaller variants when us-

ing prompt P3 reveal a substantial improvement, especially in the case of GPT OSS 20B, which achieved performance that was closest to that of the larger models.

Nevertheless, some limitations should be acknowledged. The use of public chat interfaces for GPT-4o and DeepSeek V3 imposed operational constraints, such as the inability to explore different hyperparameter configurations, which may have influenced the outcomes. Moreover, although the annotators are fluent in English, their native language is Portuguese, which may have influenced their interpretation of ambiguous constructions—particularly in sensitive domains such as gender identity. These limitations do not compromise the validity of the results but highlight the need for complementary experiments. As future work, we propose extending the analysis to other domains of sensitive discourse and applying the methodology to additional languages, particularly Portuguese, to assess its effectiveness across diverse cultural contexts. Furthermore, future research will include evaluating the proposed approach across multiple digital platforms.

Declarations

Acknowledgements

This research was funded by the National Council for Scientific and Technological Development (CNPq), grants 311898/2021-1 and 307088/2023-5; by the Carlos Chagas Filho Foundation for Research Support of the State of Rio de Janeiro (FAPERJ), grants SEI-260003/013660/2024 (E-26/204.238/2024), SEI-260003/002930/2024, and SEI-260003/000614/2023; and by the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) – Funding Code 001. This work also benefited from computational credits provided through Cohere Labs.

References

Albladi, A. *et al*. (2025). Hate Speech Detection Using Large Language Models: A Comprehensive Review. *IEEE Access*, 13:20871–20892.

Aluru, S. S., Mathew, B., Saha, P., and Mukherjee, A. (2020). Deep learning models for multilingual hate speech detection.

Anagnostidis, S. and Bulian, J. (2024). How susceptible are llms to influence in prompts?

Assis, G. *et al*. (2024a). Explorando técnicas de aprendizado em modelos de linguagem para classificação de discurso de Ódio e ofensivo em português. *Linguamática*, 16(2):91–113.

Assis, G. *et al*. (2024b). Exploring Portuguese Hate Speech Detection in Low-Resource Settings: Lightly Tuning Encoder Models or In-context Learning of Large Models? In *Proc. of 16th PROPOR*, pages 301–311, Santiago de Compostela. ACL.

Brown, T. B. *et al*. (2020). Language Models are Few-Shot Learners.

Chiu, K.-L., Collins, A., and Alexander, R. (2022). Detecting Hate Speech with GPT-3.

Choi, H. K. and Li, Y. (2024). PICLe: Eliciting Diverse Behaviors from Large Language Models with Persona In-Context Learning. In *Proc. of the 41st PMLR*, volume 235 of *Proceedings of Machine Learning Research*, pages 8722–8739. PMLR.

Cohere (2025). Command A: An Enterprise-Ready Large Language Model.

DeepSeek-AI (2025a). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

DeepSeek-AI (2025b). Deepseek-V3 Technical Report.

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., *et al*. (2024a). A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128.

Dong, Q. *et al*. (2024b). A Survey on In-context Learning. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Guo, K., Hu, A., Mu, J., Shi, Z., Zhao, Z., Vishwamitra, N., and Hu, H. (2024). An Investigation of Large Language Models for Real-World Hate Speech Detection.

Kim, J. W., Guess, A., Nyhan, B., and Reifler, J. (2021). The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication*, 71(6):922–946.

Lester, B. *et al*. (2021). The power of scale for parameter-efficient prompt tuning. In *Proc. of the 2021 EMNLP*, pages 3045–3059, Punta Cana. ACL.

Li, L., Fan, L., Atreja, S., and Hemphill, L. (2024). "HOT" ChatGPT: The Promise of ChatGPT in Detecting and Discriminating Hateful, Offensive, and Toxic Comments on Social Media. *ACM Trans. Web*, 18(2).

Li, X. L. and Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4582–4597. ACL.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.*, 55(9).

Nguyen, N. D., Truong, N., Dao, P. Q., and Nguyen, H. H. (2025). Can online behaviors be linked to mental health? active versus passive social network usage on depression via envy and self-esteem. *Comput. Hum. Behav.*, 162:108455.

Oliveira, A. *et al*. (2023). How good is ChatGPT for detecting Hate Speech in Portuguese? In *Anais do XIV STIL*, pages 94–103, Porto Alegre, RS, Brasil. SBC.

Oliveira, A. *et al*. (2024a). Toxic Speech Detection in Portuguese: A Comparative Study of Large Language Models. In *Proc.s of the 16th PROPOR*, pages 108–116, Santiago de Compostela. ACL.

Oliveira, A. *et al*. (2024b). Toxic Text Classification in Portuguese: Is LLaMA 3.1 8B All You Need? In *Anais do XV STIL*, pages 57–66, Porto Alegre, RS, Brasil. SBC.

OpenAI (2024). GPT-4o System Card.

OpenAI (2025). gpt-oss-120b and gpt-oss-20b model card.

Paes, A., Vianna, D., and Rodrigues, J. (2024). Modelos

de linguagem. In *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 17. BPLN, 3 edition.

Paiva, L., Assis, G., Amorim, A., Dias, L. G., Paes, A., and de Oliveira, D. (2025). Domínio delimitado, Ódio exposto: O uso de prompts para identificação de discurso de Ódio online com llms. In *Anais do XL Simpósio Brasileiro de Bancos de Dados*, pages 493–506, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbbd.2025.247279.

Pérez, J. M. *et al*. (2025). Exploring Large Language Models for Hate Speech Detection in Rioplatense Spanish. In *NAACL 2025*, pages 7174–7187, Albuquerque, New Mexico. ACL.

Saraiva, G. D. *et al*. (2021). A semi-supervised approach to detect toxic comments. In *Proc. of the RANLP 2021*, pages 1261–1267, Online. INCOMA Ltd.

Touvron, H. *et al*. (2023). LLaMA: Open and Efficient Foundation Language Models.

Vargas, F. *et al*. (2021). Contextual-lexicon approach for abusive language detection. In *Proc. of the RANLP 2021*, pages 1438–1447, Online. INCOMA Ltd.

Vaswani, A. *et al*. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wei, J. *et al*. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Zhao, W. X. *et al*. (2023). A survey of large language models. *arXiv:2303.18223*, 1(2).

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., and Ba, J. (2023). Large language models are human-level prompt engineers.