Domínio Delimitado, Ódio Exposto: O Uso de *Prompts* para Identificação de Discurso de Ódio *Online* com LLMs*

Laryssa Paiva[1], Gabriel Assis[1], Annie Amorim[1], Luiz Gustavo Dias[1,2], Aline Paes[1], Daniel de Oliveira[1]

[1]Universidade Federal Fluminense, Niterói, RJ, Brasil

[2]Instituto Federal Goiano, Goiânia, GO, Brasil

{laryssapaiva,assisgabriel,annieamorim,lgdias}@id.uff.br,
{alinepaes,danielcmo}@ic.uff.br

Abstract. *With the expansion of social media in recent decades, online hate speech has emerged as a relevant social and technological issue. Automatically identifying this type of content is challenging due to the inherent subjectivity of the messages, as well as cultural and individual variations. Large Language Models (LLMs) have emerged as promising alternatives for this task, but their performance is sensitive to prompt formulation. This paper investigates the impact of event contextualization in prompts on the task of hate speech detection. Three progressively structured prompt versions were developed and evaluated on the Command-A, GPT-4o, and DeepSeek-V3 models. Using real user comments collected around the release of the film Emilia Pérez, the results show that including context about the event in the prompt positively influences model performance, with improvements of up to 16% in the F1 score.*

Resumo. *Com a expansão das redes sociais nas últimas décadas, o discurso de ódio online emergiu como um problema social e tecnológico relevante. A identificação automática desse tipo de conteúdo é desafiadora devido à subjetividade inerente às mensagens e às variações culturais e individuais. Modelos de linguagem de grande escala (LLMs) surgem como alternativas promissoras para essa tarefa, mas seu desempenho é sensível à formulação dos prompts. Este artigo investiga o impacto da contextualização de eventos nos prompts na tarefa de detecção de discurso de ódio. Foram desenvolvidas três versões progressivas de prompts e avaliadas nos modelos Command-A, GPT-4o e DeepSeek-V3. Utilizando comentários reais coletados sobre o filme Emilia Pérez, os resultados mostram que a inclusão de contexto sobre o evento no prompt impacta positivamente o desempenho dos modelos, com ganhos registrados de até 16% na métrica F1.*

1. Introdução

O uso de redes sociais tem crescido de maneira exponencial nas últimas décadas em escala global [Nguyen et al. 2025]. Tais plataformas transformaram como os indivíduos interagem e constroem relações interpessoais, possibilitando, em especial, a conexão entre indivíduos geograficamente distantes. No entanto, apesar desses benefícios, as redes sociais também

*__Nota dos autores:__ Este artigo contém exemplos de conteúdo ofensivo explícito e de discurso de ódio, utilizados exclusivamente para fins de ilustração e análise dos problemas discutidos. Esses exemplos não refletem, em hipótese alguma, as opiniões ou posicionamentos pessoais dos autores e suas instituições.

passaram a constituir um ambiente propício à disseminação de conteúdos ofensivos, discriminatórios e prejudiciais, especialmente direcionados a grupos historicamente marginalizados ou sub-representados [Kim et al. 2021]. Esse tipo de manifestação tem sido denominado de *discurso de ódio online* [Vargas et al. 2021, Aluru et al. 2020].

O discurso de ódio *online* é um fenômeno que envolve não apenas a dimensão social, mas também a dimensão tecnológica, configurando-se como um problema multifacetado e de difícil resolução. Apesar da crescente atenção dedicada tanto pela comunidade científica [Albladi et al. 2025], por órgãos como as Nações Unidas[1], quanto pelos próprios proprietários das plataformas de redes sociais, sua detecção continua sendo um desafio em aberto. Um primeiro problema se encontra no volume de conteúdo gerado continuamente nessas plataformas, o que torna inviável a detecção manual (além da mesma estar sujeita a vieses individuais dos anotadores). Mesmo com o avanço de soluções automatizadas [Saraiva et al. 2021] para a identificação do discurso de ódio *online*, a tarefa ainda é desafiadora por conta de questões como a subjetividade das mensagens (*e.g.*, o uso de figuras de linguagem, ironia, emojis, *etc*.) e pelas variações culturais e linguísticas.

Além dos desafios supracitados, outro problema associado à detecção automática de discurso de ódio *online* é como utilizar informações complementares e acessórias de suporte referentes ao domínio do texto postado. Tomemos como exemplo o *tweet* apresentado na Figura 1. A versão à esquerda poderia, em um primeiro momento, ser classificada como uma mensagem ofensiva, por conter linguagem de baixo calão, uma vez que está direcionada "somente" a uma emissora de televisão (*i.e.*, Rede Globo). No entanto, ao se considerar a *hashtag* que acompanha o conteúdo na versão da direita (que está associada a participação da cantora Pabllo Vittar no programa Encontro[2]), a mesma mensagem é passível de ser reinterpretada como discurso de ódio, por estar direcionada à artista, que é abertamente homossexual e se apresenta como *drag queen*. Esse exemplo evidencia como a ausência de informações de suporte pode comprometer a detecção de discurso de ódio e reforça a importância de considerar dados complementares para um melhor julgamento das mensagens.

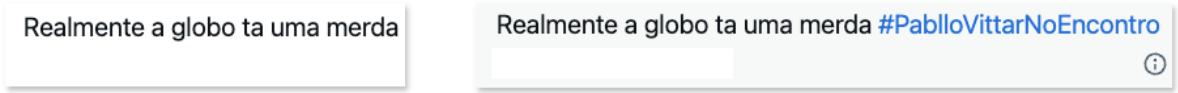

Figura 1. *Tweet* sobre a participação da cantora Pabllo Vittar no programa Encontro.

Nos últimos anos, os modelos de linguagem de grande escala (do inglês *Large Language Model* - LLMs) [Zhao et al. 2023] têm se consolidado como uma alternativa para viabilizar diversas tarefas de classificação sem a necessidade de treinamento de modelos específicos [Dong et al. 2024b], incluindo a detecção de discurso de ódio [Albladi et al. 2025]. Apesar dos avanços, os LLMs ainda se mostram sensíveis à formulação dos *prompts* [Anagnostidis and Bulian 2024], o que representa um desafio adicional, especialmente em tarefas como a detecção de discurso de ódio *online*, nas quais a ausência de contexto explícito pode comprometer a acurácia das respostas.

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

[2] https://www.correiobraziliense.com.br/app/noticia/diversao-e-arte/2017/08/09/interna_diversao_arte,616440/pabllo-vittar-participa-do-encontro-e-toma-conta-da-internet.shtml

A literatura recente tem explorado estratégias variadas de engenharia de *prompts* [Zhou et al. 2023, Li and Liang 2021, Lester et al. 2021], como a inserção de exemplos representativos no enunciado das tarefas [Dong et al. 2024a] e a incorporação de personas para guiar o comportamento do modelo [Choi and Li 2024]. Contudo, ainda há uma lacuna de estudos que investiguem o impacto da inclusão de informações de suporte específicas sobre o domínio ao qual pertencem os textos analisados. Considerando que a interpretação de um comentário que propaga ódio *online* tem potencial de depender de conhecimento prévio sobre eventos, pessoas ou situações mencionadas, torna-se importante compreender de que forma a presença (ou a ausência) de contexto explícito afeta o desempenho dos LLMs em tarefas sensíveis como esta.

Neste artigo, propomos avaliar o impacto da inserção de informações de suporte nos *prompts* aplicados aos LLMs na tarefa de detecção de discurso de ódio *online*. Para isso, foram desenvolvidas três versões progressivas de *prompts*, com diferentes níveis de informação acessórias, aplicadas a três modelos distintos: (i) Command-A, (ii) GPT-4o e (iii) DeepSeek V3. A avaliação foi conduzida com base em um conjunto de comentários reais extraídos de vídeos do YouTube, relacionados à repercussão do filme Emilia Pérez[3]. Os resultados, tanto quantitativos quanto qualitativos, indicam que a inclusão de informações de suporte nos *prompts* contribui para uma melhor interpretação das mensagens e para a redução de erros, especialmente na identificação de discursos de ódio *online* com linguagem disfarçada, como os associados à transfobia. Esses achados evidenciam a relevância da contextualização como estratégia de engenharia de *prompt* em LLMs, com implicações diretas para o desenvolvimento de sistemas mais justos e sensíveis às nuances sociais.

O artigo se encontra organizado em cinco seções, além da Introdução. A Seção 2 apresenta o referencial teórico necessário à compreensão do estudo. A Seção 3 discute os trabalhos relacionados. A Seção 4 detalha a metodologia adotada. A Seção 5 analisa os resultados obtidos, tanto sob uma perspectiva qualitativa quanto quantitativa. Por fim, a Seção 6 apresenta as conclusões.

2. Modelos de Linguagem de Grande Escala e Engenharia de *Prompt*

Esta seção apresenta e contextualiza conceitos-chave relacionados aos LLMs, que fundamentam as discussões e experimentos desenvolvidos neste artigo. Desde a introdução da arquitetura *Transformer* [Vaswani et al. 2017], os modelos de linguagem baseados em redes neurais têm alcançado avanços significativos. Dentre esses, os LLMs destacam-se por sua capacidade de interpretar e gerar textos com elevada coerência semântica. Arquiteturalmente, esses modelos derivam do componente decodificador do *Transformer* original e se diferenciam não apenas pela escala, com bilhões de parâmetros, mas também por apresentarem habilidades emergentes, comportamentos complexos que não se manifestam em modelos de menor porte [Paes et al. 2024].

O desenvolvimento de modelos como o GPT-3 e o GPT-3.5 [Brown et al. 2020], utilizados como base nas primeiras versões do ChatGPT[4], evidenciou o potencial desses sistemas para executar uma ampla gama de tarefas, incluindo classificação, tradução e sumarização, mesmo na ausência de treinamento supervisionado específico. Essa versatilidade está associada ao chamado Aprendizado por Contexto (*in-context learning*) [Dong et al. 2024a],

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

no qual a tarefa é apresentada ao modelo por meio de um *prompt* escrito em linguagem natural, eventualmente complementado com exemplos. Nessa abordagem, o *prompt* atua como uma instrução textual que orienta a geração da resposta pelo LLM. Embora sua estrutura seja flexível, o conteúdo do *prompt* tem influência direta no comportamento do modelo, uma vez que a geração textual se fundamenta nos princípios da semântica distribucional [Paes et al. 2024]. Dessa forma, a formulação do *prompt* exerce impacto direto sobre a qualidade e adequação da resposta gerada.

Essa constatação foi um dos principais catalisadores para o surgimento da área atualmente conhecida como *engenharia de prompts*, cujo objetivo central é investigar e desenvolver estratégias de estruturação textual que otimizem o desempenho dos LLMs em tarefas específicas [Liu et al. 2023]. Diversas técnicas têm sido propostas nesse campo, incluindo a inserção de exemplos representativos no enunciado das tarefas [Dong et al. 2024a], a incorporação de personas para guiar o comportamento do modelo [Choi and Li 2024], e a decomposição do raciocínio em etapas sucessivas, como ocorre na abordagem conhecida como *chain-of-thought prompting* (CoT) [Wei et al. 2022]. Tais estratégias têm demonstrado eficácia em diversos contextos e aplicações. No entanto, a seleção da estratégia mais apropriada para cada cenário, bem como a sua implementação, permanece um desafio em aberto e constitui uma área ativa de investigação [Liu et al. 2023]. Dessa forma, ainda que os LLMs apresentem um potencial para o aprimoramento de diversas tarefas, a sua aplicação demanda um uso criterioso e consciente, sobretudo em domínios sensíveis, como a detecção de discurso de ódio *online*. Essa tarefa, em particular, envolve não apenas elementos de subjetividade individual, mas também aspectos contextuais e culturais não triviais, os quais exigem especial atenção durante o processo de modelagem e avaliação.

3. Trabalhos Relacionados

Nos últimos anos, os LLMs têm sido empregados na tarefa de detecção e classificação de discurso de ódio *online*, consolidando-se como uma ferramenta promissora nesse domínio [Albladi et al. 2025]. A principal razão para essa popularidade reside na capacidade desses modelos de capturar e interpretar o conteúdo das entradas textuais, o que lhes dá uma vantagem em relação às abordagens tradicionais. Embora boa parte das soluções existentes sejam voltadas à língua inglesa, observam-se avanços recentes em outros idiomas, como o espanhol [Pérez et al. 2025] e o português [Assis et al. 2024b].

[Chiu et al. 2022] apresentam um estudo utilizando o modelo GPT-3 para a identificação de conteúdos com teor racista e sexista. Nesse trabalho, os autores exploram abordagens baseadas em *prompts* contendo exemplos ilustrativos, com o intuito de orientar o modelo na tarefa de detecção desses discursos ofensivos. Entretanto, [Chiu et al. 2022] não exploram informações de suporte nos *prompts* gerados. Também utilizando os modelos da família GPT, [Li et al. 2024] apresentam um estudo sobre o desempenho do ChatGPT na análise de comentários publicados em plataformas digitais, considerando os textos sob a seguinte taxonomia: discurso de ódio (*Hate*), conteúdo ofensivo (*Offensive*) e linguagem tóxica (*Toxic*), agrupadas sob a sigla HOT. O estudo avalia cinco variações distintas de *prompt*, nas quais o modelo é instruído a indicar, de forma binária, a presença ou ausência de uma das categorias HOT para cada comentário. As variações consideradas envolvem diferentes graus de complexidade, incluindo ou não justificativas geradas pelo modelo, estimativas de probabilidade e níveis variados de contextualização. Os resultados obtidos revelam que o desempenho do modelo é influenciado pela formulação do *prompt*, indicando que a escolha da instrução

mais apropriada deve levar em conta os objetivos específicos da tarefa em questão. Entretanto, [Li et al. 2024] se limitaram a utilizar o ChatGPT, sem explorar outros LLMs.

Similarmente ao trabalho de [Li et al. 2024], [Guo et al. 2024] apresentam um estudo no qual o ChatGPT foi aplicado a diversos contextos sensíveis, como a detecção de sexismo, manifestações xenofóbicas e conteúdos publicados durante períodos eleitorais. Nesse trabalho, foram comparadas quatro estratégias distintas de *prompt*: (i) instruções simples, (ii) inclusão de definições conceituais, (iii) exemplos contextualizados e (iv) a técnica de raciocínio encadeado (*chain-of-thought*). Os resultados reforçam a conclusão de que o desempenho do modelo depende de maneira substancial da estrutura do *prompt*, sendo particularmente eficazes os formatos que fornecem exemplos ilustrativos e instruções que induzem uma sequência lógica de raciocínio por parte do modelo.

Por sua vez, [Oliveira et al. 2023] avaliam duas configurações distintas de *prompt* aplicadas ao modelo GPT-3.5 na tarefa de detecção de textos tóxicos. Os resultados indicam que *prompts* mais completos e explicativos tendem a produzir um desempenho superior, corroborando a hipótese de que a riqueza das instruções fornecidas ao modelo impactam na sua eficácia. Achados semelhantes são relatados por [Oliveira et al. 2024a, Assis et al. 2024b], os quais conduzem uma comparação entre o desempenho do GPT e de um *chatbot* brasileiro, o Maritalk[5]. Especificamente, [Assis et al. 2024b] investigam o impacto do uso de exemplos ilustrativos e da inserção de palavras-chave contextuais nos *prompts*, sugerindo que tais elementos contribuem para um desempenho mais equilibrado do modelo. Além disso, esses estudos realizam comparações entre os LLMs e classificadores tradicionais especificamente treinados, destacando a competitividade dos modelos ativados via *prompting*, mesmo na ausência de ajustes supervisionados específicos para a tarefa em questão.

Adicionalmente, embora os modelos da família GPT sejam bastante utilizados em estudos voltados à detecção de discurso de ódio, LLMs abertos também têm sido objeto de investigação. Em particular, [Assis et al. 2024a] e [Oliveira et al. 2024b] exploram variantes baseadas na arquitetura LLaMA [Touvron et al. 2023], realizando ajustes explícitos nos pesos dos modelos por meio de técnicas de *fine-tuning*. Contudo, as abordagens adotadas por esses estudos diferem do presente artigo e entre si em aspectos metodológicos relevantes. [Assis et al. 2024a] adaptam os modelos diretamente para atuarem como classificadores supervisionados, seguindo a lógica tradicional de treinamento com rótulos anotados. Já em [Oliveira et al. 2024b], o processo é precedido por uma etapa de engenharia de *prompt*, na qual diferentes estratégias são testadas para identificar aquelas que induzem o melhor desempenho inicial. Apenas após essa etapa exploratória é realizado o ajuste fino do modelo para a tarefa de classificação de discurso de ódio, o que evidencia uma integração entre estratégias de *prompt* e aprendizado supervisionado.

Os trabalhos anteriormente discutidos nos mostram que a formulação do *prompt* exerce influência direta sobre o desempenho de LLMs em tarefas relacionadas à detecção e classificação de discurso de ódio. Com base nesse cenário, o presente estudo se propõe a investigar o impacto da inclusão de informações adicionais nos *prompts*, com ênfase na incorporação de informações de contexto dos eventos associados às postagens analisadas. A hipótese central é que tais informações podem auxiliar os modelos a realizar inferências mais precisas, especialmente em casos ambíguos ou dependentes de conhecimento de mundo. Para isso, são avaliados não apenas modelos amplamente utilizados, como o GPT-

[5]`https://chat.maritaca.ai/`

4o [OpenAI 2024], mas também LLMs menos recorrentes em estudos sobre discurso de ódio, como o DeepSeek-V3 [DeepSeek-AI 2025] e o Command-A [Cohere 2025]. Os experimentos foram conduzidos com base em um conjunto de dados desenvolvido especificamente para este estudo, coletado, revisado e anotado manualmente, o que assegura um controle da qualidade e representatividade do *dataset*. Dessa forma, os resultados apresentados oferecem novas perspectivas para a aplicação de LLMs em contextos sensíveis, apoiando-se em dados cuidadosamente curados e em estratégias de *prompting* orientadas por informação sobre os eventos relacionados.

4. Metodologia

A metodologia seguida neste artigo foi delineada com o propósito de investigar o impacto de diferentes níveis de informação de suporte incorporadas aos *prompts* submetidos aos LLMs na tarefa de detecção de discurso de ódio *online*. Com esse objetivo, a metodologia foi estruturada em quatro etapas principais e interdependentes: (i) coleta e preparação de um conjunto de dados composto por postagens reais, representativas de situações potencialmente ofensivas ou discriminatórias; (ii) elaboração de versões progressivas de *prompts*, variando em complexidade e grau de informação de suporte, de modo a permitir a avaliação comparativa de seu efeito sobre o desempenho dos modelos; (iii) anotação manual das postagens por avaliadores humanos, com o intuito de construir um conjunto de referência sensível às nuances do cenário analisado; e (iv) aplicação dos modelos selecionados e avaliação de seus resultados por meio de análises quantitativas e qualitativas.

4.1. Coleta e Preparação dos Dados

Os comentários analisados neste estudo foram extraídos de um único vídeo hospedado no YouTube, com o objetivo de limitar o contexto ao mesmo tempo em que se procurava uma compreensão mais aprofundada dos textos submetidos à avaliação. A seleção do vídeo seguiu critérios específicos, com ênfase na relevância e na densidade de interações dos usuários. Inicialmente, realizou-se uma busca utilizando a *query* `karla sofia in:english`, a fim de localizar conteúdos relacionados à atriz Karla Sofía Gascón que apresentassem potencial para gerar reações variadas do público. A referida atriz foi escolhida uma vez que estava no epicentro de ataques durante a campanha do Oscar do filme *Emilia Pérez* por conta de antigos *tweets* que revelaram inúmeras opiniões preconceituosas. Entre os resultados, identificaram-se vídeos de entrevistas, premiações e reportagens relacionadas a controvérsias envolvendo a atriz na rede social X[6].

O critério final para a escolha do vídeo considerou tanto o número de visualizações quanto a quantidade de comentários disponíveis, resultando na seleção de uma entrevista concedida por Karla Sofía Gascón ao programa *The Tonight Show Starring Jimmy Fallon*, intitulada "*Karla Sofía Gascón Fangirled Over Harrison Ford and Mark Hamill, Talks Emilia Pérez Transformation*"[7]. O vídeo, com duração de 10 minutos e 21 segundos, foi publicado em 14 de janeiro de 2025. Na data da coleta dos comentários, realizada em 14 de março de 2025, o vídeo acumulava 243.394 visualizações e 1.425 comentários. A coleta dos dados foi realizada por meio da API do YouTube[8], e resultou em uma amostra aleatória de 851 comentários, que foram posteriormente submetidos às etapas de anotação e análise.

[6]https://oglobo.globo.com/cultura/noticia/2025/01/30/internet-resgata-tweets-antigos-de-karla-sofia-gascon-criticando-isla-george-floyd-e-diversidade-no-oscar.ghtml

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

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

Após a etapa de coleta, foi realizada uma filtragem manual dos comentários com o intuito de garantir a qualidade necessária para a análise. Inicialmente, foram excluídos todos os comentários redigidos em idiomas diferentes do inglês, incluindo comentários que apresentavam trechos ou expressões em múltiplos idiomas. Comentários compostos exclusivamente por emojis também foram descartados. Ao término da filtragem, o conjunto final passou a ser composto por 604 comentários. Dentre esses textos, três comentários foram selecionados para compor os exemplos utilizados nos *prompts*, sendo um exemplo por classe — neste estudo, seguindo [Assis et al. 2024b]: neutro, ofensivo e discurso de ódio —, de modo a preservar a proximidade semântica e temática com o conjunto analisado. A seleção foi realizada de forma aleatória, até que se identificassem textos representativos para cada uma das classes consideradas. Esses comentários foram então removidos do conjunto principal, com o objetivo de evitar possíveis vieses nos resultados do experimento. A escolha dos exemplos foi conduzida por um anotador externo à equipe principal de anotação (ver Seção 4.3), e posteriormente validada por todos os demais anotadores quanto à adequação dos textos às respectivas categorias.

4.2. Elaboração dos *Prompts*

Com o objetivo de investigar o impacto da formulação de *prompts* na tarefa de detecção de discurso de ódio *online* por meio de LLMs, foram desenvolvidas três versões progressivas de *prompts*, todas em inglês, em consonância com o idioma dos comentários extraídos do YouTube. Cada versão introduz um incremento em termos de informação de suporte, permitindo avaliar o efeito dessa variação sobre o desempenho dos modelos. A primeira versão do *prompt* (P1) consiste na descrição da tarefa e na apresentação das definições das classes, instruindo o LLM da seguinte maneira:

Prompt base

> 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.

A segunda versão do *prompt* (P2) preservou a estrutura básica da primeira versão (P1), composta pela descrição da tarefa e pelas definições das classes, mas introduziu, adicionalmente, exemplos representativos para cada uma das classes avaliadas. A inclusão desses exemplos teve como objetivo fornecer aos modelos instruções mais concretas, com exemplos das classes esperadas, favorecendo, assim, uma compreensão mais acurada da tarefa de classificação. Os exemplos selecionados foram os seguintes: (i) para a classe *discurso de ódio* foi selecionado o texto *"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) para a classe *ofensivo* foi selecionado o texto *"This guys are nuts"*; e (iii) para a classe *neutro* foi selecionado o texto *"I'm not sure he's concerned with your opinion. The audience seemed to enjoy it"*.

A terceira versão do *prompt* (P3), por sua vez, ampliou mais o nível de detalhamento fornecido ao LLM, ao incorporar informações de suporte adicionais àquelas já presentes nas versões anteriores. Essa versão incluiu uma descrição sucinta da narrativa central do filme *Emilia Pérez*, informações sobre o elenco principal e esclarecimentos quanto à origem dos

comentários analisados, incluindo o vídeo do qual os textos foram extraídos. O objetivo dessa formulação com mais informações de suporte foi permitir que os LLMs acessassem informações complementares relevantes, que auxiliassem na interpretação dos comentários. O trecho fornecido foi o seguinte: "*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. Anotação dos Dados

O processo de anotação dos dados foi conduzido em duas etapas. A primeira etapa correspondeu à fase de anotação inicial, na qual cada comentário foi rotulado de forma independente por, no mínimo, dois anotadores. Para garantir a consistência do procedimento, todos os anotadores tiveram acesso ao mesmo conjunto de instruções fornecidas aos LLMs, incluindo as definições das três classes, os exemplos ilustrativos e as informações de suporte disponibilizadas na versão mais completa do *prompt* (P3). No entanto, cada anotador visualizou exclusivamente o subconjunto de textos que lhe foram atribuídos para rotulação, sem acesso prévio às anotações dos demais.

Todos os anotadores assistiram previamente ao filme *Emilia Pérez*, bem como ao vídeo específico do qual os comentários foram extraídos. A equipe de anotação foi composta por seis participantes, sendo três mulheres e três homens, com idades entre 20 e 35 anos. Todos são estudantes de graduação ou pós-graduação, e a maioria possuía experiência prévia em atividades de pesquisa relacionadas à análise de discurso de ódio em ambientes digitais. Considerando a natureza sensível dos temas abordados tanto no filme quanto nos comentários, a equipe foi deliberadamente composta de forma diversa, incluindo participantes que se identificam como membros da comunidade LGBTQIAPN+, contribuindo para uma análise que considera as nuances socioculturais envolvidas.

Após a anotação individual, as respostas foram comparadas. Dos 601 textos anotados, houve discordância em 127 casos. Assim, a segunda etapa da anotação consistiu em uma adjudicação realizada por um juiz. O juiz comparou as respostas e os comentários fornecidos pelos primeiros anotadores e atribuiu o rótulo final para cada comentário em disputa. O juiz estava limitado a escolher entre um dos rótulos propostos pelos anotadores iniciais. Vale ressaltar que o juiz também se identifica abertamente como parte da comunidade LGBTQIAPN+. Ao final do processo de anotação, dos textos anotados, 310 foram classificados como neutros, 129 como ofensivos e 162 como discurso de ódio.

4.4. Aplicação dos Modelos

Para a aplicação e posterior avaliação das variações de *prompts* desenvolvidas neste artigo, foram empregados três LLMs: (i) o Command-A [Cohere 2025], (ii) o GPT-4o [OpenAI 2024], e (iii) o DeepSeek V3 [DeepSeek-AI 2025]. Em virtude das limitações impostas pelos custos de acesso às APIs dos modelos GPT-4o e DeepSeek V3, a submissão dos *prompts* foi realizada por meio das respectivas interfaces públicas de *chat*, disponíveis nas plataformas das empresas desenvolvedoras. Nesses dois casos, por fins práticos, sete comentários distintos foram submetidos em cada interação para cada uma das versões de *prompt* avaliadas.

No caso do LLM Command-A, a interação foi conduzida por meio da API oficial da Cohere, viabilizada pelo uso de créditos concedidos para fins acadêmicos. Essa forma de acesso possibilitou o ajuste de parâmetros de inferência, especificamente, a configuração da `temperature`, fixada em 0,3. Esse valor foi adotado com base em estudos anteriores [Assis et al. 2024b], por apresentar uma boa relação entre redução da variabilidade das respostas e aumento da consistência com as classes-alvo da tarefa, a saber: *hate speech*, *offensive* e *neutral*. Todos os LLMs foram submetidos às três versões progressivas de *prompts* (P1, P2 e P3), de modo a permitir a análise comparativa do impacto das variações no grau de detalhamento das instruções sobre o desempenho classificatório dos modelos. Tal abordagem possibilita uma avaliação da influência do conteúdo das instruções presentes nos *prompts* sobre a sensibilidade e a precisão dos LLMs no reconhecimento de conteúdos sensíveis.

5. Avaliação Experimental

A avaliação experimental teve como objetivo examinar o desempenho dos LLMs diante das diferentes versões de *prompts* elaboradas neste artigo. Para isso, foram conduzidas análises quantitativas e qualitativas, de forma complementar, a fim de analisar os efeitos das variações no conteúdo de suporte dos *prompts* na capacidade dos LLMs de identificar corretamente os comentários conforme as classes avaliadas, nominalmente: *hate speech*, *offensive* e *neutral*.

5.1. Resultados Quantitativos

A avaliação quantitativa foi conduzida com base em quatro métricas tradicionais de desempenho em tarefas de classificação: (i) acurácia, (ii) precisão, (iii) sensibilidade (*recall*) e (iv) medida F1. Adicionalmente, calculou-se a variação percentual da medida F1 em relação à versão mais simples do *prompt* (P1), utilizada como *baseline*. Dada a distribuição desbalanceada entre as classes no conjunto de dados, a medida F1 foi adotada como principal indicador de desempenho. A Tabela 1 sintetiza os resultados obtidos por cada um dos LLMs avaliados, nas três versões progressivas de *prompt* (P1, P2 e P3). De modo geral, os resultados evidenciam que o *prompt* P3, que incorpora, de forma cumulativa, definições das classes, exemplos representativos e informações de suporte sobre o evento, proporcionou o melhor desempenho para todos os LLMs testados. Esse padrão sugere que a inclusão de informações de suporte pertinentes no *prompt* contribui para aprimorar a capacidade dos LLMs em diferenciar entre conteúdos neutros, ofensivos e de ódio.

Especificamente, no caso do Command-A, observou-se uma evolução consistente na medida F1, que passou de 0,582 com o *prompt* P1 para 0,676 com o *prompt* P3, o que representa um ganho percentual de 16,15%. Por outro lado, o modelo GPT-4o apresentou um comportamento não linear: a introdução isolada de exemplos no *prompt* (P2) resultou em uma redução de desempenho, com queda de 4,25% na F1 em comparação à versão base. No entanto, ao se incorporar também o contexto (P3), o modelo alcançou um aumento de 11,57% em relação à P1, atingindo a maior F1 geral entre todos os cenários (0,761). Esse resultado reforça a hipótese de que a introduçao de informações de suporte exerce papel-chave na eficácia da classificação.

Dada a complexidade inerente à detecção de discurso de ódio *online*, foi realizada uma análise complementar com foco específico na medida F1 da classe de discurso de ódio. Essa análise visa verificar em que medida as diferentes versões de *prompt* impactam a capacidade dos LLMs em identificar corretamente conteúdos associados a essa categoria, frequentemente caracterizada por nuances semânticas, ambiguidade e subjetividade. Os resultados

Tabela 1. Resultados de avaliação para os LLMs com variação dos *prompts*.

Modelo	Prompt	Acurácia	Precisão	Sensibilidade	F1	Δ F1 (%)	F1 Hate
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**

obtidos reafirmam o padrão observado na avaliação geral, evidenciando ganhos de desempenho com a utilização do *prompt* mais completo (P3). No caso do modelo Command-A, o valor de F1 para a classe de discurso de ódio evoluiu de 0,397 com o *prompt* P1 para 0,620 com o *prompt* P3, representando um aumento na capacidade de detecção dessa classe. O modelo DeepSeek V3 seguiu tendência similar, com crescimento de 0,584 para 0,733. Já o GPT-4o apresentou um aumento de 0,604 para 0,776, reafirmando o impacto positivo da informação de suporte sobre o evento no desempenho do modelo, especialmente em tarefas como a classificação de discurso de ódio.

Esses achados sustentam a hipótese de que a presença de informações de suporte explícitas nos *prompts* favorece a interpretação mais acurada das intenções comunicativas subjacentes aos comentários analisados. De maneira geral, os dados indicam que, além de melhorar o desempenho global, as informações adicionais nos *prompts* são especialmente benéficas para a classe mais sensível e difícil de rotular, *i.e.*, o discurso de ódio, contribuindo para uma classificação mais robusta e socialmente responsável. Em termos comparativos, o melhor resultado foi novamente registrado com o modelo GPT-4o utilizando o *prompt* P3, que atingiu a maior medida F1 entre todos os cenários analisados (0,761).

Além dos resultados na medida F1, verificou-se uma melhora consistente nas demais métricas avaliadas com a utilização do *prompt* P3. Em termos de acurácia, os três LLMs alcançaram seus desempenhos máximos com essa versão: o Command-A obteve 0,742, o GPT-4o atingiu 0,782, e o DeepSeek V3 registrou 0,722. Esses resultados indicam uma redução na taxa global de erros de classificação, evidenciando que a inserção de informação de suporte não apenas favorece classes específicas, mas contribui para um ganho geral na confiabilidade dos modelos. De maneira semelhante, observou-se um incremento nas métricas de precisão e sensibilidade, o que demonstra que os modelos se tornaram simultaneamente mais capazes de identificar corretamente os textos pertencentes a categorias sensíveis (menor incidência de falsos negativos) e de evitar classificações indevidas (menor incidência de falsos positivos). Por exemplo, a sensibilidade do modelo GPT-4o aumentou de 0,687 com o *prompt* P1 para 0,775 com o *prompt* P3, enquanto sua precisão evoluiu de 0,705 para 0,765.

5.2. Resultados Qualitativos

A análise qualitativa concentrou-se nos comentários rotulados como discurso de ódio (HS) que foram corretamente classificados pelos modelos quando submetidos somente ao *prompt* P3. Os exemplos apresentados, destacados nos comentários 1 a 6 da Tabela 2, evidenciam os

desafios enfrentados pelas LLMs na detecção de manifestações de transfobia, especialmente em textos breves e sutilmente hostis. Nessas instâncias, termos aparentemente neutros como *"he"*, *"man"* e *"actor"* podem ser empregados com intenção pejorativa, deslegitimando a identidade de gênero da pessoa mencionada. Tais expressões, quando avaliadas fora de contexto, tendem a ser erroneamente interpretadas pelos modelos como neutros (NEU) ou, no máximo, ofensivos (OFF). No entanto, os anotadores humanos, com base na compreensão ampliada do cenário sociocultural em que os comentários foram produzidos, atribuíram a classificação de discurso de ódio, reconhecendo os elementos discriminatórios implícitos.

A incorporação de informações de suporte no *prompt* P3 mostrou-se importante para que os LLMs replicassem esse julgamento mais sensível. Ao fornecer detalhes sobre a obra cinematográfica, o perfil da atriz e a origem dos comentários analisados, os LLMs foram capazes de captar nuances antes negligenciadas, resultando em classificações mais precisas. Em contraste, nos casos em que os mesmos comentários foram analisados com os *prompts* P1 ou P2 — que apresentam menos informações de suporte —, os modelos frequentemente atribuíram classes equivocadas, majoritariamente classificando os textos como neutros.

Tabela 2. Classificações dos comentários pelas LLMs segundo as diferentes versões de *prompts*, em comparação com o rótulo atribuído ao texto.

Comentário		Command-A			GPT-4o			DeepSeek V3			Rótulo
ID	**Texto**	**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

Outro aspecto relevante refere-se à classificação equivocada de conteúdos neutros como discurso de ódio, o que poderia resultar em censura indevida de opiniões legítimas. Essa preocupação é ilustrada pelos comentários 7 e 8 apresentados na Tabela 2, nos quais o modelo Command-A, quando usando os *prompts* P1 e P2, atribuiu incorretamente a categoria HS a textos neutros. Apenas com a utilização da versão P3 foi possível alcançar a interpretação correta. Um comportamento semelhante foi observado nos comentários 9 e 10, analisados pelo modelo DeepSeek V3, evidenciando a recorrência desse tipo de erro quando os LLMs operam sem acesso a informações que permitam desambiguar o tom e o conteúdo das mensagens. Nessas situações, a ausência de informações de suporte leva os LLMs a superestimar a hostilidade, confundindo críticas legítimas, direcionadas à qualidade cinematográfica da obra, com manifestações discriminatórias. Assim, o fornecimento de informações de contexto auxilia em duas frentes complementares: revela agressões implícitas que escapariam sem contexto e, ao mesmo tempo, preserva manifestações críticas legítimas, evitando sua indevida rotulação como discurso de ódio.

De modo geral, considerando somente os resultados obtidos com o uso do *prompt* P3, o Command-A apresentou o maior número de erros em comparação aos demais, classificando incorretamente 27 textos que foram corretamente classificados pelos outros LLMs. A maioria desses casos foi marcada como ofensiva, como ilustrado pelo comentário 11. Em contraste, os modelos GPT-4o e DeepSeek V3 cometeram apenas um erro cada, nos comentários 12 e 13, respectivamente. No comentário 12, a presença sutil de transfobia, sugerida pelo uso do pronome *"him"*, dificultou a identificação correta sem a informação de suporte. Já o comentário 13 apresenta uma crítica implícita à questão de gênero na indústria cinematográfica, exigindo uma leitura mais sensível para ser reconhecida como discurso de ódio, algo que apenas foi alcançado com a informação de suporte. Tais resultados reforçam que, mesmo com *prompts* mais completos, a detecção de discurso de ódio continua desafiadora, especialmente quando as mensagens carregam preconceito de forma implícita. A inclusão de informações de suporte sobre os eventos se mostra chave para melhorar a interpretação das LLMs, ao mesmo tempo em que contribui para reduzir erros tanto na detecção de conteúdo odioso quanto na prevenção de classificações indevidas de críticas legítimas.

6. Conclusões

Este artigo investigou o impacto da inclusão de informações de suporte nos *prompts* utilizados por LLMs na tarefa de detecção de discurso de ódio *online*. Foram desenvolvidas três versões progressivas de *prompts*, variando desde instruções básicas da tarefa (P1), passando pela adição de exemplos ilustrativos (P2), até a incorporação de informações detalhadas sobre o evento relacionado aos comentários analisados (P3). Os experimentos foram conduzidos com três LLMs: (i) Command-A, (ii) GPT-4o e (iii) DeepSeek V3. Foi utilizado um conjunto de comentários reais extraídos do YouTube, relacionados ao filme *Emilia Pérez*. A análise quantitativa revelou que a versão do *prompt* que utiliza informações de suporte (P3) resultou em melhorias em todas as métricas avaliadas. Destaca-se o aumento de até 16,15% na F1 em relação ao *prompt* base, além de ganhos em acurácia, precisão e sensibilidade. Esses resultados indicam que a presença de informações de suporte auxilia os LLMs não apenas na detecção mais precisa de discursos de ódio, mas também na redução de falsos positivos. A análise qualitativa complementa esses achados, reforçando que considerar informações de suporte sobre os eventos permite aos LLMs interpretar corretamente manifestações mais sutis de ódio, especialmente aquelas associadas à transfobia, e distinguir críticas legítimas de ataques discriminatórios.

Entretanto, algumas limitações devem ser consideradas. O uso das interfaces públicas de *chat* para os modelos GPT-4o e DeepSeek V3 impôs restrições operacionais, como a impossibilidade de explorar diferentes configurações de hiperparâmetros (*e.g.*, temperatura), que podem afetar os resultados. Além disso, apesar de fluentes em inglês, os anotadores têm o português como idioma materno, o que pode ter influenciado a interpretação de construções ambíguas, especialmente em um domínio sensível como o de identidade de gênero. Essas limitações não comprometem os resultados do estudo, mas apontam para a necessidade de experimentos complementares. Como trabalhos futuros, propõe-se a ampliação da análise para outros domínios de discurso sensível, bem como a aplicação da metodologia a outros idiomas, com destaque para o português, a fim de verificar sua eficácia em diferentes contextos culturais. Pretende-se ainda avaliar a abordagem em múltiplas plataformas digitais.

Agradecimentos

Esta pesquisa foi financiada pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), processos 311898/2021-1 e 307088/2023-5, pela Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), processos SEI-260003/013660/2024 (E-26/204.238/2024), SEI-260003/002930/2024 e SEI-260003/000614/2023, e pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001. Este trabalho também contou com o apoio de créditos computacionais concedidos por uma Bolsa de Pesquisa da Cohere Labs.

Referências

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 (2025). Deepseek-V3 Technical Report.

Dong, Q. et al. (2024a). 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.

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

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 Zong, C., Xia, F., Li, W., and Navigli, R., editors, *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 4582–4597, Online. Association for Computational Linguistics.

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.

Paes, A., Vianna, D., and Rodrigues, J. (2024). Modelos de linguagem. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 17. BPLN, 3 edition.

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.