Beyond Toxicity: Evaluating Language Models for Ableism Detection in Online Social Media

Annie Amorim[1‡], Gabriel Assis[1‡], Daniele Pimenta[1], Débora Pina[2], Aline Paes[1], Daniel de Oliveira[1]

[1] Universidade Federal Fluminense, Niterói, Brasil
{annieamorim,assisgabriel,danielepimenta}@id.uff.br, {alinepaes,danielcmo}@ic.uff.br
[2] Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil
dbpina@cos.ufrj.br

Abstract. Ableism is a form of discrimination directed at people with disabilities that can appear through explicit offenses, but also through subtle, indirect, and context-dependent expressions that are difficult to identify automatically. This challenge remains underexplored in content moderation, particularly in Brazilian Portuguese. This paper investigates how different levels of prompt information affect the performance of Large Language Models (LLMs) in detecting ableist comments on social media. To this end, we build a dataset from YouTube comments related to disability-related content, followed by manual annotation. Four LLMs are evaluated using five progressively detailed prompts, ranging from a simple classification instruction to prompts that include definitions, examples, contextual information, and the video transcript. Results show that more informative prompts tend to improve classification performance, although the effect varies across models. Qualitative inspection further reveals that LLMs often struggle to distinguish ableist discourse from criticism of ableism, as well as to separate ableism from other forms of offensive or discriminatory language. These findings highlight the importance of contextualized prompting for sensitive classification tasks and point to the need for further research on ableism detection in Portuguese.

CCS Concepts: • **Computing methodologies** → **Natural language processing**.

Keywords: ableism detection, large language models, prompt engineering, social media, text classification

1. INTRODUÇÃO

As redes sociais se consolidaram como espaços centrais de circulação de discursos públicos, nos quais, todavia, diferentes formas de violência podem ser produzidas, reproduzidas e amplificadas. Entre essas formas, o capacitismo se manifesta por meio de expressões, piadas, comentários e avaliações que inferiorizam pessoas com deficiência ou associam determinadas condições corporais, cognitivas ou sensoriais a estigmas sociais. Embora esse tipo de discurso possa ocorrer de maneira explícita, muitas manifestações capacitistas são sutis, dependem de contexto e exigem sensibilidade interpretativa para serem identificadas [Phutane et al. 2025; Rizvi et al. 2025].

Nos últimos anos, modelos de linguagem de grande porte (LLMs, do inglês *Large Language Models*), têm sido amplamente empregados em tarefas de classificação textual, incluindo a detecção de toxicidade, discurso de ódio e conteúdos ofensivos [Li et al. 2024; Assis et al. 2024a]. No entanto, a identificação de capacitismo ainda representa um desafio particular, uma vez que nem todo comentário capacitista é necessariamente reconhecido como tóxico ou ofensivo por sistemas automáticos [Phutane et al. 2025]. Além disso, a interpretação desse conteúdo pode depender de informações adicionais,

Os autores notificam que este artigo inclui exemplos de conteúdo nocivo, utilizados unicamente como ilustrações dos problemas discutidos, e que não expressam de forma alguma as suas opiniões pessoais.
Além disso, agradecem ao financiamento do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), bolsa 307088/2023-5, da Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), processos SEI-260003/002930/2024, SEI-260003/000614/2023, e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Código Financeiro 001. ‡ Igual contribuição.

como o tema discutido, o grupo social afetado e o conteúdo que motivou a interação [Li et al. 2024].

Nesse contexto, este trabalho investiga o uso de LLMs na classificação de comentários capacitistas em português brasileiro. Para isso, analisamos comentários extraídos de um vídeo curto do YouTube relacionado a uma fala humorística envolvendo crianças com hidrocefalia. A partir dos comentários coletados, foi construída uma amostra anotada manualmente por sete avaliadores, considerando as categorias '*capacitista*' e '*não capacitista*'. Essa etapa permitiu produzir um conjunto de referência para avaliar o desempenho dos modelos na tarefa.

Além disso, o estudo examina como diferentes níveis de informação fornecidos nos *prompts* afetam a classificação automática [Paiva et al. 2025]. Para isso, foram desenvolvidas cinco versões progressivas de *prompt*, partindo de uma instrução simples e avançando para versões com definições, exemplos e informações sobre o vídeo analisado. Quatro LLMs foram avaliados nessa configuração, buscando observar tanto diferenças entre modelos quanto o impacto da inclusão gradual de contexto na identificação de manifestações capacitistas. Os resultados, apoiados por análises quantitativas e qualitativas, indicam que, de modo geral, *prompts* com mais informações de suporte tendem a produzir melhores resultados. No entanto, o efeito de cada informação adicional varia entre os modelos, sugerindo que diferentes LLMs respondem de maneira distinta à inclusão de informação de suporte adicional.

Assim, este artigo contribui para a discussão sobre o uso de LLMs em tarefas sensíveis de classificação textual, com foco em uma forma de discriminação ainda pouco explorada em português brasileiro. Ao articular anotação humana, variação controlada de *prompts* e avaliação de modelos acessíveis, buscamos analisar em que medida LLMs conseguem lidar com manifestações sutis e contextualmente dependentes de capacitismo em comentários de redes sociais.

2. TRABALHOS RELACIONADOS

Trabalhos recentes têm investigado como LLMs percebem e classificam manifestações de capacitismo. Phutane et al. [2025] comparam avaliações produzidas por sistemas de IA e por pessoas com deficiência, a partir de comentários direcionados a esse grupo. Os autores mostram que os modelos tendem a subestimar a toxicidade percebida por pessoas com deficiência, além de produzir avaliações instáveis sobre capacitismo e justificativas frequentemente pouco sensíveis ao contexto. Esses resultados indicam que a detecção de capacitismo não pode ser reduzida à identificação de toxicidade explícita, pois envolve experiência social, interpretação contextual e formas sutis de violência discursiva. De modo complementar, Li et al. [2024] analisam vieses capacitistas em LLMs a partir de uma perspectiva interseccional. Os autores mostram que a manifestação e a percepção do capacitismo nos resultados dos modelos são afetadas pela combinação entre deficiência, gênero e classe, evidenciando que esses vieses não operam de forma isolada. O estudo também destaca a necessidade de estratégias de mitigação mais inclusivas e sensíveis a nuances.

Mais próximo ao nosso estudo, mas em inglês, Rizvi et al. [2025] propõem o AUTALIC, um conjunto de dados para detecção de linguagem capacitista anti-autista em sentenças do Reddit. O trabalho utiliza anotadores com formação relacionada à neurodiversidade e avalia LLMs de menor porte, reportando baixo nível de concordância, com Fleiss' Kappa médio de 0,25, reforçando a subjetividade da tarefa. O estudo também explora cenários *zero-shot* e *few-shot* com LLMs, sem investigar progressivamente diferentes níveis de informação contextual nos *prompts*. Os resultados indicam que a detecção de linguagem capacitista permanece desafiadora para LLMs.

A influência da formulação dos *prompts* tem sido discutida em tarefas de detecção de discurso de ódio. Nesse contexto, Li et al. [2024] e Guo et al. [2024] avaliam o LLMs em contextos sensíveis, como sexismo, xenofobia e conteúdos eleitorais, comparando instruções simples, definições, exemplos e raciocínio encadeado. Os resultados indicam que o desempenho depende da estrutura do *prompt*, especialmente quando são fornecidos exemplos e instruções mais detalhadas. Em português, estudos como [Oliveira et al. 2024; Assis et al. 2024a] também mostram que *prompts* mais completos, com

exemplos e palavras-chave de suporte, tendem a melhorar a classificação de toxicidade e discurso de ódio. Alinhados à nossa configuração experimental, Paiva et al. [2025] avaliam a inserção progressiva de informações de suporte em *prompts*, partindo de instruções simples e avançando para definições, exemplos e contexto de vídeo do YouTube, embora com foco na classificação geral de discurso de ódio.

Diferentemente desses trabalhos, este estudo investiga a classificação de manifestações capacitistas em português brasileiro, tomando como caso comentários associados a um vídeo do Youtube sobre hidrocefalia. A partir de evidências anteriores sobre a relevância do contexto e da estrutura dos *prompts* [Paiva et al. 2025], avaliamos quatro LLMs com o objetivo de analisar como esses modelos lidam com formas sutis e contextualmente dependentes de capacitismo.

3. MATERIAIS & MÉTODOS

Esta seção apresenta os procedimentos de coleta, preparação, seleção e anotação dos dados, bem como a construção dos *prompts* e a classificação automática dos comentários. O objetivo é construir uma base para identificar manifestações capacitistas em redes sociais, combinando análise humana e LLMs.

3.1 Coleta e Preparação dos Dados

Os textos analisados neste estudo consistem em comentários extraídos de um único vídeo do YouTube. Essa escolha metodológica permitiu delimitar o escopo da análise, reduzir a variabilidade temática dos dados e favorecer uma interpretação mais aprofundada das manifestações submetidas à avaliação, uma vez que todos os comentários foram produzidos em resposta ao mesmo conteúdo. Foram coletados mais de 8 mil comentários do vídeo "*Absurdo! Léo Lins debochando de crianças com hidrocefalia*"[1], publicado no formato YouTube Shorts e com 23 segundos de duração.

Após a coleta inicial, realizou-se uma etapa de limpeza e filtragem dos dados. Foram removidos comentários compostos apenas por *emojis*, risadas, pontuações isoladas ou outros conteúdos sem informação textual relevante. Também foram excluídas respostas a outros comentários, a fim de manter apenas comentários principais e reduzir dependências conversacionais entre os textos. Ao final, obteve-se um conjunto de 3.549 comentários em português válidos para análise.

3.2 Seleção e Redução dos Dados

Foi aplicada uma estratégia de amostragem para viabilizar a anotação manual dos comentários com os recursos humanos disponíveis (§3.4). O objetivo foi selecionar um subconjunto menor de textos, preservando diferentes níveis de potencial ofensivo nos dados analisados. Para isso, utilizou-se o modelo ShieldGemma-2B [Zeng et al. 2024], desenvolvido para a detecção de conteúdo potencialmente nocivo e capaz de atribuir probabilidades associadas à presença de conteúdos abusivos, ofensivos ou inseguros em textos. Embora o foco deste estudo seja a identificação de manifestações capacitistas, empregou-se, nesta etapa preliminar, a configuração padrão do modelo, baseada na distinção entre discurso de ódio e não discurso de ódio. O ShieldGemma-2B foi utilizado exclusivamente para orientar a seleção da amostra, uma vez que não foi treinado especificamente para a detecção de capacitismo, e seus resultados não foram utilizados como rótulos de referência na avaliação dos LLMs. Partiu-se da premissa de que comentários com maior probabilidade de discurso de ódio poderiam concentrar casos potencialmente capacitistas, enquanto comentários com menor probabilidade deveriam ser preservados para garantir contraste, diversidade e equilíbrio na análise posterior.

A seleção da amostra foi definida a partir das probabilidades atribuídas pelo modelo. Entre os 3.549 comentários válidos, 128 apresentaram probabilidade igual ou superior a 0,5 e foram, portanto, classificados como discurso de ódio pela configuração adotada. Devido à baixa frequência desses

[1] `https://www.youtube.com/shorts/nOAvOIrsm-s`

casos e à sua relevância potencial para a identificação de manifestações capacitistas, todos foram mantidos na amostra. Para os comentários restantes, com probabilidade inferior a 0,5, adotou-se uma estratificação por faixas de probabilidade. O intervalo entre 0 e 0,5 foi dividido em quatro faixas: $[0, 0{,}125)$, $[0{,}125, 0{,}25)$, $[0{,}25, 0{,}375)$ e $[0{,}375, 0{,}5)$. De cada faixa, foram selecionados 50 comentários. Essa escolha buscou preservar comentários com diferentes graus de potencial ofensivo, incluindo casos de menor probabilidade, que ainda poderiam conter manifestações sutis ou indiretas de capacitismo. Além disso, incorporou-se um critério de engajamento, selecionando 45 comentários com mais de 50 curtidas que não haviam sido contemplados pelos critérios anteriores. Essa decisão considera que comentários com maior engajamento tendem a ter maior visibilidade na plataforma e podem ser relevantes para a análise independentemente da probabilidade atribuída pelo modelo. Ao final, foram selecionados 373 comentários para anotação manual. Assim, a estratégia combinou três objetivos: preservar os casos considerados mais críticos pelo modelo, garantir variação ao longo do espectro de probabilidade e incluir comentários com maior repercussão.

3.3 Elaboração dos *Prompts*

Alinhado ao objetivo de investigar como diferentes níveis de informação de suporte afetam o desempenho dos modelos e com trabalhos anteriores [Paiva et al. 2025], foram desenvolvidas cinco versões progressivas de *prompt*, denominadas P1, P2, P3, P4 e P5. Todas foram redigidas em português, em consonância com o idioma dos comentários analisados. Essa organização permitiu comparar o comportamento dos modelos diante de instruções gradualmente mais informativas, partindo de uma formulação mínima da tarefa e avançando para versões com definições, exemplos e contexto adicional.

A primeira versão, P1, apresentou apenas a tarefa de classificar os comentários nas categorias '*capacitista*" e '*não capacitista*", sem definições adicionais. Com isso, buscou-se observar se os modelos seriam capazes de interpretar os rótulos a partir de seu conhecimento prévio. O P2 manteve a mesma formulação, mas acrescentou descrições conceituais das categorias, permitindo avaliar o efeito da explicitação dos rótulos na classificação, conforme apresentado no Quadro 1. Na sequência, o P3 incorporou exemplos, de modo a verificar se exemplos mínimos auxiliariam os modelos a calibrar suas decisões. Esses exemplos foram selecionados aleatoriamente entre comentários que não integravam a amostra destinada à anotação manual. Como exemplo de comentário não capacitista, utilizou-se: '*Não sei o que é mais nojento, quem fez a piada ou quem riu da piada*'; e, como capacitista: '*Ué, essa piada já é antiga no show dele, pq esse mimimi só agora? Cambada de enjoado... se tem defeito tem que zuar mesmo kkk todo mundo zoa mas ninguém tem coragem de falar*'.

Quadro 1 - Estrutura do *Prompt* P2

Você receberá um comentário em português extraído do YouTube, o qual deverá ser classificado em uma das seguintes categorias: "CAPACITISTA" ou "NÃO CAPACITISTA".
Para realizar essa tarefa de classificação, considere atentamente as definições fornecidas a seguir:
CAPACITISTA: Engloba comentários que ridicularizam, inferiorizam, zombam, degradam, desumanizam, estereotipam, estigmatizam, discriminam ou promovem ódio contra pessoas com deficiência, condições médicas, síndromes, neurodivergências ou impedimentos físicos, intelectuais, cognitivos ou sensoriais. Inclui piadas depreciativas, manifestações de desprezo, exclusão, humilhação, insultos ou referências ofensivas a condições relacionadas à deficiência, à saúde, ao corpo, à cognição ou ao desenvolvimento.
NÃO CAPACITISTA: Corresponde a comentários que não apresentam ridicularização, inferiorização, zombaria, degradação, desumanização, estereótipos, estigmatização, discriminação ou promoção de ódio contra pessoas com deficiência, condições médicas, síndromes, neurodivergências ou impedimentos físicos, intelectuais, cognitivos ou sensoriais. Inclui comentários neutros, críticas ao humorista ou ao vídeo, manifestações de indignação, discussões sobre humor ou liberdade de expressão, bem como menções educativas, jornalísticas, médicas, informativas ou de defesa de direitos relacionadas à deficiência, desde que não reproduzam discurso discriminatório ou referências ofensivas contra pessoas com deficiência.

Considerando maiores níveis de informações, o P4 acrescentou elementos contextuais sobre a origem dos comentários, indicando que eles foram extraídos de um vídeo curto publicado no YouTube relacionado a uma fala humorística envolvendo crianças com hidrocefalia. Também foi incluída uma breve descrição da condição, destacando que a hidrocefalia está associada ao acúmulo de líquido no

cérebro e pode envolver necessidades específicas de cuidado, além de estigmas sociais. Por fim, o P5 reuniu todos os elementos anteriores e acrescentou informações mais detalhadas sobre o vídeo analisado, incluindo a transcrição integral do conteúdo audiovisual. A transcrição fornecida foi a seguinte: *"Eu acho muito legal o Teleton, porque eles ajudam crianças com vários tipos de problemas. Eu vi o vídeo de um garoto no interior do Ceará com hidrocefalia. O lado bom é que o único lugar na cidade onde tem água é a cabeça dele. A família nem mandou tirar, instalou um poço. Agora o pai puxa água do filho, todos felizes tomando banho."*. A comparação entre P4 e P5 permitiu investigar se o acesso ao conteúdo completo que motivou os comentários contribuiria para a classificação.

3.4 Anotação Manual dos Dados

Na etapa de rotulação manual, sete anotadores foram instruídos a classificar os 373 comentários como '*capacitista*' ou '*não capacitista*'. Para isso, receberam o *prompt* P5, com orientações detalhadas, e assistiram ao vídeo de origem dos dados, a fim de considerar o panorama geral dos comentários. Cada item foi avaliado por três anotadores independentes, e o rótulo final foi definido por maioria simples. O grupo de anotadores foi composto por estudantes e pesquisadores de diferentes níveis de formação, incluindo graduação, pós-graduação, mestrado, doutorado e pós-doutorado. Buscou-se formar um grupo diverso em termos de gênero, idade, raça e relação com o tema, incluindo pessoas com deficiência e avaliadores com vivência, formação ou proximidade com debates sobre deficiência e capacitismo.

Ao final, 46 comentários foram classificados como capacitistas e 327 como não capacitistas, resultando em uma distribuição desbalanceada entre as classes. A concordância entre os avaliadores, mensurada pelo coeficiente Fleiss' Kappa, foi de 0,478. Esse resultado indica concordância moderada e reforça a subjetividade da tarefa, na qual a identificação de manifestações capacitistas pode depender de contexto, interpretação e sensibilidade dos anotadores.

3.5 Aplicação dos Modelos

A avaliação das variações de *prompt* foi conduzida com quatro LLMs: (i) Command R7B[2] [Cohere 2025], (ii) DeepSeek-R1 14B [DeepSeek 2025], (iii) Gemma-4 26B-A4B [Google DeepMind 2026] e (iv) GPT OSS 20B [OpenAI 2025]. Buscou-se selecionar modelos de famílias distintas, permitindo observar possíveis diferenças de comportamento diante da tarefa e das cinco versões de *prompt* (P1–P5). Além disso, a escolha por modelos desse porte está alinhada aos recursos computacionais disponíveis, favorecendo a execução e a reprodução dos experimentos.

Os modelos possuem pesos abertos e foram acessados por meio do Ollama[3], o que favorece a padronização do ambiente de execução e a reprodução dos experimentos. Além disso, foi adotada uma configuração fixa de `temperature`=0,3 para todos os modelos, seguindo evidências de trabalhos anteriores [Assis et al. 2024b] que indicam maior consistência das respostas em tarefas de classificação.

4. AVALIAÇÃO EXPERIMENTAL

Esta seção avalia os modelos na classificação de comentários capacitistas, considerando diferentes versões de *prompt* e combinando métricas quantitativas com exemplos qualitativos de acertos e erros.

4.1 Resultados Quantitativos

Considerando o desbalanceamento entre as classes, a acurácia deve ser interpretada com cautela. Por esse motivo, o F1-macro é utilizado como principal métrica para a avaliação. Os resultados

[2] `https://docs.cohere.com/docs/command-r7b`
[3] `https://ollama.com/`

da Tabela I indicam diferenças relevantes entre os modelos e entre as versões de *prompt*. De modo geral, a inclusão progressiva de informações tende a melhorar o desempenho, sobretudo quando são adicionados definições, exemplos e informações de suporte sobre o vídeo. No entanto, esse efeito não é uniforme, sugerindo que cada modelo responde de forma distinta ao aumento de informações.

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

Modelo	Prompt	Acurácia	Precisão	Sensibilidade	F1-macro	Δ F1-macro (%)
Command R7B	P1	0,306	0,461	0,426	0,287	-
	P2	0,252	0,441	0,405	0,244	-14,97%
	P3	0,271	0,471	0,453	0,264	-8,25%
	P4	0,252	0,456	0,433	0,246	-14,17%
	P5	**0,324**	**0,491**	**0,484**	**0,308**	**+7,31%**
DeepSeek-R1 14B	P1	0,748	0,544	0,567	0,545	-
	P2	0,711	0,521	0,536	0,514	-5,68%
	P3	0,780	0,546	0,557	0,549	+0,68%
	P4	0,769	0,545	0,560	0,548	+0,51%
	P5	**0,845**	**0,648**	**0,659**	**0,653**	**+19,78%**
Gemma-4 26B-A4B	P1	0,802	0,381	0,386	0,383	-
	P2	0,853	**0,620**	**0,580**	**0,592**	**+54,77%**
	P3	0,853	0,428	0,405	0,414	+8,20%
	P4	0,863	0,456	0,428	0,439	+14,82%
	P5	**0,885**	0,518	0,455	0,478	+25,01%
GPT OSS 20B	P1	0,716	0,349	0,347	0,342	-
	P2	0,812	0,551	0,547	0,549	+60,37%
	P3	0,804	0,558	0,561	0,560	+63,46%
	P4	0,810	0,588	**0,602**	0,594	+73,34%
	P5	**0,839**	**0,613**	0,600	**0,606**	**+76,90%**

O Command R7B apresentou o desempenho mais baixo. Seu melhor resultado ocorreu com o P5, com acurácia de 0,324 e F1-macro de 0,308, um ganho de 7,31% em relação ao P1. Ainda assim, os valores permaneceram inferiores aos dos demais modelos. Além disso, P2, P3 e P4 reduziram o F1-macro em relação ao P1, indicando que a adição parcial de informação não foi suficiente para melhorar sua classificação. O DeepSeek-R1 14B apresentou desempenho mais estável e obteve seu melhor resultado também com o P5, alcançando acurácia de $0,845$ e F1-macro de $0,653$, ganho de $19,78\%$ sobre o P1. Esse foi o maior F1-macro absoluto entre todos os modelos.

O Gemma-4 26B-A4B apresentou um comportamento distinto dos demais modelos. Seu melhor F1-macro foi obtido com o P2, atingindo 0,592, o que representa um ganho de 54,77% em relação ao P1. Embora o P5 tenha alcançado a maior acurácia do modelo, 0,885, seu F1-macro foi inferior, sugerindo que o aumento de acertos gerais ocorreu sem melhora equivalente no equilíbrio entre as classes. Para o Gemma-4 26B-A4B, portanto, a explicitação das definições das classes parece ter sido mais efetiva do que a inclusão posterior de exemplos e informação adicional.

Já o modelo GPT OSS 20B apresentou os maiores ganhos relativos de F1-macro. O desempenho cresceu de forma consistente a partir do P1, com ganhos superiores a 60% em P2, P3, P4 e P5. Seu melhor resultado ocorreu com o P5, com acurácia de 0,839 e F1-macro de 0,606, representando aumento de 76,90% em relação ao P1. Esse comportamento sugere que o modelo se beneficiou fortemente da contextualização progressiva, especialmente com a inclusão da transcrição completa do vídeo.

No geral, os resultados mostram que o desempenho dos modelos é influenciado pela formulação dos *prompts*, mas não de maneira linear. Para três dos quatro modelos, o melhor F1-macro foi obtido com versões mais informativas, especialmente o P5, que inclui definições, exemplos, contexto do vídeo e transcrição. Isso sugere que informações de suporte podem auxiliar os LLMs a interpretar manifestações capacitistas mais dependentes de contexto. No entanto, o caso do Gemma-4 26B-A4B indica que mais informação nem sempre resulta em melhor equilíbrio entre as classes, uma vez que o P2 foi suficiente para produzir seu melhor F1-macro. Assim, os resultados reforçam que a variação de *prompt* é um fator central na tarefa, mas que o efeito de cada elemento adicional pode depender do modelo avaliado. Novamente, apesar das tendências observadas, essas conclusões devem ser tratadas como indicativas, dado o caráter delimitado do estudo e o número reduzido de instâncias da classe-alvo.

4.2 Resultados Qualitativos

A inspeção qualitativa dos comentários evidencia que uma das principais dificuldades dos modelos está na identificação do alvo discursivo. Nos exemplos 1 a 3 da Tabela II, os comentários apresentam linguagem negativa, mas direcionada ao humorista, à piada ou às pessoas que riram do conteúdo, e não a pessoas com deficiência. Por isso, foram rotulados como não capacitistas. Ainda assim, algumas versões de *prompt*, sobretudo as menos contextualizadas, classificaram esses casos como capacitistas, sugerindo uma tendência a associar agressividade linguística à presença de discriminação. Esse padrão foi mais evidente no Command R7B, que classificou corretamente esses exemplos apenas com o P5, indicando que o contexto e a transcrição do vídeo ajudaram o modelo a diferenciar crítica ao capacitismo de reprodução de discurso capacitista.

Outro padrão aparece nos exemplos 4 e 5, em que os comentários criticam o humorista ou seu público, mas utilizam termos historicamente associados ao capacitismo, como "retardado". Por esse motivo, foram rotulados como capacitistas. Esses casos mostram que a tarefa não consiste apenas em identificar se o comentário apoia ou condena a piada, mas também em verificar se, ao fazê-lo, reproduz vocabulário discriminatório. Termos como "retardado", "imbecil" ou "maluco" podem ampliar a ambiguidade, pois nem sempre são dirigidos diretamente a pessoas com deficiência, mas carregam sentidos ligados à estigmatização de condições cognitivas, intelectuais ou psicossociais.

Os exemplos 6, 7 e 10 representam casos em que o capacitismo aparece de forma mais diretamente vinculada ao conteúdo da piada original. O exemplo 6 associa a hidrocefalia à ideia de água em uma construção depreciativa, além de combinar essa referência com outro marcador discriminatório (origem africana). O exemplo 7 retoma a mesma associação, acrescentando um esteriótipo regional. Já o exemplo 10 reproduz diretamente a lógica da fala do vídeo. Nesses casos, os *prompts* mais informativos tendem a favorecer a classificação correta, pois fornecem aos modelos o contexto necessário para reconhecer formulações que dependem da piada original para produzir sentido capacitista.

Tabela II. Comparação dos comentários pelas LLMs segundo as diferentes versões de *prompts* com o rótulo atribuído.

Comentário		**Rót.**	**Command**					**DeepSeek**					**Gemma**					**GPT OSS**				
ID	**Texto**		**P1**	**P2**	**P3**	**P4**	**P5**	**P1**	**P2**	**P3**	**P4**	**P5**	**P1**	**P2**	**P3**	**P4**	**P5**	**P1**	**P2**	**P3**	**P4**	**P5**
1	*Precisa ser muito hiena pra rir de uma idiotice dessas.*	**N**	C	C	C	C	N	C	N	N	N	N	N	N	N	N	N	N	N	N	N	N
2	*Que horror!!! Parabéns ao SBT por ter demitido esse idiota!!!*	**N**	C	C	C	C	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
3	*Absurdo msm/Não sabe o que um pai ou uma mãe passa com um filho com hidrocefalia*	**N**	C	C	C	N	N	C	N	N	N	N	N	N	N	N	N	C	N	N	N	N
4	*Piadas sem graça véi. E um monte de retardados pagam pra ouvir e ver um idiota assim.*	**C**	C	C	C	C	C	N	C	C	C	C	C	C	C	C	C	N	C	C	C	C
5	*o melhor e as pessoas defendendo um retardado que nem esse, quero ver se fosse com algum familiar dele que tivesse essa doença e ficassem debochando se ele ia gostar*	**C**	C	C	C	C	C	N	C	C	C	C	C	C	C	C	C	N	C	C	C	C
6	*Sabe qual a semelhança de um muleque com hidrocefalia e eu africano/ Os dois só pensa em água*	**C**	C	C	C	C	C	N	C	C	C	C	C	C	C	C	C	C	C	C	C	C
7	*mais legal é que o cara é cearense, então equivale a uma caixa d'água*	**C**	C	N	C	C	C	C	N	C	N	C	N	N	N	C	C	N	N	N	C	C
8	*Qual semelhança de uma mulher negra gravida e um carro quebrado? Os 2 esperam um macaco*	**N**	C	C	C	C	C	N	C	C	C	C	N	C	N	N	N	N	N	N	N	N
9	*Eu tinha um amigo negro, eu dava comida, cobertor, água, ele só reclamava do tamanho da gaiola. kkkkklll*	**N**	C	C	C	C	C	N	C	C	C	C	N	C	N	N	N	C	N	N	N	N
10	*Puxar água do filho é foda kkkkkkkkkkkkkkkkkkkk*	**C**	N	C	C	C	C	N	N	N	N	C	N	N	N	C	C	N	N	N	N	C

Nota: C = Capacitista; N = Não capacitista.

Por fim, os exemplos 8 e 9 mostram a confusão entre capacitismo e outras formas de discurso de ódio. Embora contenham conteúdo racista e ofensivo, esses comentários não se dirigem a pessoas com deficiência, condições médicas ou impedimentos físicos, intelectuais, cognitivos ou sensoriais; por isso, foram rotulados como não capacitistas. A classificação incorreta desses casos como capacitistas por

alguns modelos, especialmente Command R7B e DeepSeek-R1 14B, sugere dificuldade em delimitar a categoria-alvo. Assim, a inspeção reforça que detectar capacitismo exige mais do que reconhecer toxicidade: é necessário identificar o alvo da ofensa, o contexto discursivo, o uso de léxico capacitista e a distinção entre diferentes formas de discriminação.

5. CONCLUSÕES E DISCUSSÕES

Este trabalho avaliou como diferentes níveis de detalhamento em *prompts* afetam o desempenho de LLMs na detecção de capacitismo em comentários de redes sociais. Para isso, foi construída uma base em português a partir de comentários de um vídeo do YouTube sobre uma piada envolvendo crianças com hidrocefalia, seguida de seleção amostral e anotação manual por sete avaliadores. Os resultados indicam que a inclusão de definições, exemplos e informações contextuais tende a melhorar o desempenho dos modelos, embora esse efeito varie conforme o LLM avaliado. O *prompt* com maior nível de informação (P5) obteve o melhor F1-macro para três dos quatro modelos avaliados.

A inspeção qualitativa reforçou que a detecção de capacitismo exige mais do que o reconhecimento de linguagem ofensiva. Os modelos apresentaram dificuldades para distinguir críticas ao conteúdo de manifestações capacitistas, bem como para separar capacitismo de outras formas de discurso discriminatório. Esses achados indicam que estratégias de *prompting* podem apoiar tarefas sensíveis de classificação, desde que combinadas com categorias bem definidas e informações de suporte relevantes. A utilização de comentários de um único vídeo permitiu controlar melhor o contexto dos *prompts*, mas limita a generalização dos resultados para outros contextos de capacitismo. Como trabalhos futuros, pretende-se ampliar a base para outros contextos, plataformas e tipos de deficiência, métodos de amostragem alternativos, além de avaliar estratégias multirrótulo e modelos ajustados especificamente para o português, o que pode auxiliar na interpretação de nuances linguísticas e culturais.

REFERÊNCIAS

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

Assis, G. et al. 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*. ACL, Santiago de Compostela, 2024b.

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

DeepSeek. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature* 645 (8081), 2025.

Google DeepMind. Gemma 4: Our most capable open models to date. Google Blog, 2026. Accessed: 2026-05-07.

Guo, K. et al. An Investigation of Large Language Models for Real-World Hate Speech Detection, 2024.

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

Li, R., Kamaraj, A., Ma, J., and Ebling, S. Decoding ableism in large language models: An intersectional approach. In *Proceedings of the Third Workshop on NLP for Positive Impact*. Association for Computational Linguistics, Miami, Florida, USA, pp. 232–249, 2024.

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

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

Paiva, L., Assis, G., Amorim, A., Dias, L. G., Paes, A., and de Oliveira, D. 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*. SBC, Porto Alegre, RS, Brasil, pp. 493–506, 2025.

Phutane, M., Seelam, A., and Vashistha, A. "cold, calculated, and condescending"': How AI identifies and explains ableism compared to disabled people. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '25. Association for Computing Machinery, New York, NY, USA, pp. 1927–1941, 2025.

Rizvi, N. et al. AUTALIC: A dataset for anti-AUTistic ableist language in context. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Vienna, Austria, pp. 20999–21015, 2025.

Zeng, W. et al. ShieldGemma: Generative AI content moderation based on Gemma, 2024.