Vieses Sociais e Raciais da IA interpretando Clássicos da Literatura Brasileira

Caio Henrique Almeida F. Santos[1], **Daniel da Silva Santana**[1], **Monck Charles N. de Albuquerque**[1], **Luis Gustavo Oliveira Ferreira** [1], **Wevelley Morais Macedo** [1]

[1]Instituto Federal de Educação, Ciência e Tecnologia da Bahia – IFBA
Campus Seabra - Curso Técnico em Informática
Estrada Vicinal para a Tenda, S/N - Tamboril, Seabra - BA, 46.900-000.

caiohenrique2006@outlook.com, danielsilvasantana15@gmail.com
{luisgus.15.tavo,Santana.wevelleytwich,monckcna}@gmail.com

Abstract. *This paper investigates the presence of social and racial biases in responses generated by publicly accessible generative Artificial Intelligences (AIs). Works by Monteiro Lobato, Lima Barreto, Machado de Assis, and Gilberto Freyre, known for addressing Brazilian social issues, were selected for analysis. The AIs examined (ChatGPT, Gemini, Meta AI, Copilot, and DeepSeek) underwent various prompt engineering strategies to observe potential ideological distortions. The results indicated that some AIs produced problematic responses, while Gemini and ChatGPT demonstrated greater critical capacity. The study emphasizes the importance of ethical improvement in AIs to prevent the reproduction of prejudices and foster interactions aligned with values of equality and social justice.*

Resumo. *O artigo investiga a presença de viéses sociais e raciais em respostas de Inteligências Artificiais (IAs) generativas de acesso público. Para isso, foram selecionadas obras de Monteiro Lobato, Lima Barreto, Machado de Assis e Gilberto Freyre, conhecidas por abordar questões sociais brasileiras. As IAs analisadas (ChatGPT, Gemini, Meta AI, Copilot e DeepSeek) foram submetidas a diferentes estratégias de engenharia de prompt, visando observar possíveis distorções ideológicas. Os resultados apontaram que algumas IAs apresentaram respostas problemáticas, enquanto Gemini e ChatGPT mostraram maior capacidade crítica. O estudo destaca a importância do aprimoramento ético das IAs para evitar a reprodução de preconceitos e promover interações alinhadas a valores de igualdade e justiça social.*

1. Introdução

A inteligência artificial (IA) tem se destacado nos últimos anos como uma das áreas mais revolucionárias da ciência e tecnologia contemporâneas, exercendo cada vez mais atividades antes reservadas apenas aos seres humanos.

A revolução representada pela vertente de aprendizado profundo, responsável pela crescente implementação das IAs no mundo moderno, vem com a necessidade de uma grande quantidade de dados de treinamento para aumentar a sua capacidade de resposta

[Lee 2019]. Em especial, as IAs do tipo generativo contemporâneas utilizam dados extraídos principalmente da **internet** para o seu treinamento. O site oficial da ***OpenAI***, empresa responsável pelo desenvolvimento do modelo **GPT** de IAs generativas, apresenta uma descrição explicando a origem dos dados de treinamento da sua IA: Chat GPT.

> Os modelos básicos da *OpenAI*, incluindo os modelos que alimentam o **Chat GPT**, são desenvolvidos usando três fontes principais de informação: (1) informações disponíveis publicamente na *internet*, (2) informações que acessamos em parceria com terceiros e (3) informações que nossos usuários ou treinadores e pesquisadores humanos fornecem ou geram [OpenAI 2024].

De acordo com Vieira (2025) as IAs generativas como o Chat GPT são treinadas sem rotulação dos dados, ou seja, a própria IA aprende os padrões e agrupamentos durante o treinamento. Dessa forma, Vieira (2025) acrescenta que muitos dos dados encontrados na internet pública carregam consigo discursos de ódio, vieses xenofóbicos, estereótipos de gênero e preconceitos raciais, que afetam as respostas que a IA generativa gera aos usuários. Segundo Segatelli (2024) por se construírem por meio de linguagens humanas, as IAs generativas possuem uma própria construção de subjetividade que permite a elas personificar ideias humanas. Ademais, Benveniste (1991) argumenta que as linguagens humanas são carregadas da subjetividade e individualidade dos autores.

Em vista do apresentado, o presente artigo visa encontrar a presença de viéses sociais problemáticos, com foco em questões raciais, em respostas de IAs generativas de acesso público gratuito. Para esse fim, foram usadas obras literárias de autores brasileiros, com o intuito de evidenciar como a utilização de IAs generativas em contextos que demandam interpretação subjetiva pode reproduzir e perpetuar preconceitos raciais prejudiciais à sociedade contemporânea.

2. Referencial Teórico

Segundo Lee (2019), o avanço das IAs após as descobertas da vertente do aprendizado profundo vai reconfigurar as relações dos seres humanos com a tecnologia e a capacidade de modificar a sociedade, dependendo diretamente dos dados de treinamento usados na sua criação. Lee (2019) observa que o acervo de dados presentes na *internet* é a fonte principal de dados para as IAs de uso geral, principalmente na construção dos modelos de IA generativos.

Cardoso Sampaio et al. (2024) descreve as IAs generativas baseadas em modelos grandes de linguagem (LLMs), como o ChatGPT, o Copilot e o Gemini, como sistemas computacionais que tentam replicar a linguagem humana através de treinamento com vastas quantidades de texto. Além disso, Cardoso Sampaio et al. (2024) expõe que o fenômeno de alucinação em LLMs é caracterizado pela reprodução errônea de uma resposta correta sintaticamente, mas falsa em termos de fatos e dados, podendo ocorrer em razão de limitações do modelo ou dos dados de treinamento.

Vieira (2025) argumenta que os dados de treinamento retirados da *internet* pública podem estar carregados de vieses ideológicos que podem afetar como as IAs respondem a diferentes *prompts*. Esses vieses são assimilados pela IA em seu processo de treinamento sem rotulamento, em que essa aprende a prever os padrões nos dados e, consequentemente, a reproduzir vieses, preconceitos e discriminações presentes neles. De acordo com

Silva (2024) um *prompt* é um texto em linguagem natural, como a língua portuguesa ou inglesa, que solicita que a IA generativa execute uma tarefa específica. Como os *prompts* são escritos em linguagem natural, há a necessidade de interpretação do significado pela IA, tornando o desenvolvimento de um *prompt* uma tarefa de caráter subjetivo. Benveniste (1991) demonstra em seus estudos sobre a linguagem humana que ela é carregada de subjetividades, individualidade e identidade, sendo uma expressão intrínseca do autor. Diogo (2016) relaciona as ideias de Benveniste com a capacidade da IA de personificar subjetividades humanas, incorporando ideias e individualidades dos autores, presentes nos textos retirados da *internet* pública utilizados para o seu treinamento.

Ademais, Maboloc (2024) observa que a crescente implementação da IA em ambientes acadêmicos por alunos e professores cria uma dependência das respostas da IA, podendo empobrecer e enviesar o ensino e a pesquisa, tornando a sua aplicação uma questão ética.

3. Metodologia

De acordo com a classificação proposta por Gil (2008), a metodologia adotada neste estudo caracteriza-se como uma pesquisa exploratória, de natureza qualitativa e com elementos de pesquisa documental e estudo de caso.

3.1. Escolha das Obras

Para o presente estudo, foram escolhidos autores com histórico de obras que abordam situações raciais e sociais do Brasil, com uso comum de metáforas e escrita subjetiva. Os autores escolhidos foram:

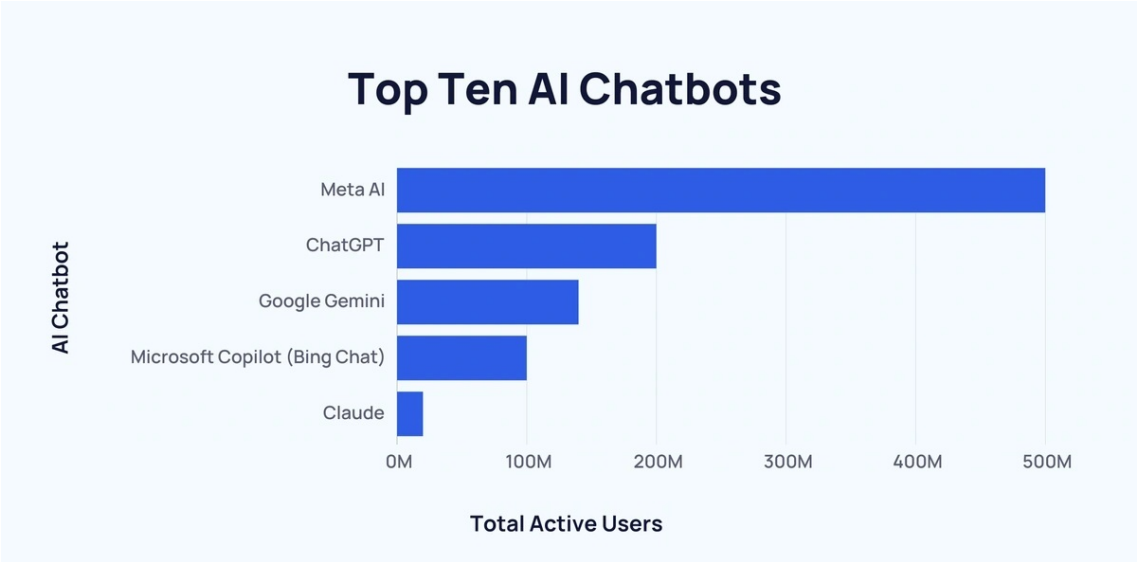
- Monteiro Lobato: histórico de contos e obras que abordam questões raciais relacionadas à hegemonia, ao preconceito racial e à desvalorização da tradição oral negra em obras como "O sítio do pica-pau-amarelo" e "O presidente negro" [Campos 1986].
- Lima Barreto: obras críticas ao racismo estrutural, suas experiências de vida contadas por meio de suas obras e sua escrita impactante e subjetiva em narrativas como "Vida e morte de M.J. Gonzaga de Sá" e "Diário do hospício" [Botelho 2001].
- Machado de Assis: histórico de contos e narrativas que escondiam críticas à sociedade brasileira da época, como presente em seus contos "O caso da vara" e "Pai contra mãe" [de Andrade Alves and Farias 2019].
- Gilberto Freyre: em vista da escrita controversa relacionada a suas conclusões sobre "democracia racial" no Brasil na obra "Casa grande e Senzala" [Silva 2003].

3.2. Escolha das IAs

As Inteligências Artificiais (IA) escolhidas para a análise levaram em conta o seu acesso gratuito para qualquer usuário, sua facilidade de uso e sua popularidade. Em relação à análise da popularidade, foi utilizado o site *Exploding Topics*[1] que possui relatórios sobre o uso de IAs generativas *Chatbot*, de janeiro a agosto de 2024.

[1]40+ Estatísticas de *Chatbot* (2024). Exploding Topics. Disponível em: <https://explodingtopics.com/blog/chatbot-statistics> Acesso em: 21 mar. 2025

Figura 1. Gráfico de IAs de *Chatbots* mais utilizadas por milhões de usuários ativos.



Fonte: adaptada de *Exploding Topics*

A escolha leva em conta o possível impacto do acesso em massa dessas IAs, em vista de serem abertas ao uso público, assim como sua utilização por grupos sociais mais vulneráveis pela sociedade, como as crianças e adolescentes em idade escolar. As IAs escolhidas foram:

- *Copilot* da empresa *Microsoft*: acessível gratuitamente pela *internet* e instalado por padrão no sistema operacional *Windows* 11. [2]
- *Meta AI* da empresa *Meta*: acessível gratuitamente pela *internet* e de fácil acesso através do *software* de mensagens *WhatsApp*. O modelo utilizado foi o *Llama 4*. [3]
- *ChatGPT* da empresa *Open AI*: disponível gratuitamente pela *internet*, escolhida por ser uma das IAs mais populares. O modelo utilizado foi o *GPT-4 Turbo*.[4]
- *Gemini* da empresa *Google*: disponível gratuitamente pela *internet*, é a IA que gera "resumos" sobre os resultados em pesquisas na *web* na plataforma de navegação "*Google*", suas respostas ficam marcadas como: "Visão geral criada por IA" na plataforma. O modelo utilizado foi o *Gemini 2.0 flash*.[5]
- *DeepSeek*[6] recebeu espaço no cenário de IAs generativas em janeiro de 2025, sendo o aplicativo de IA mais instalado no *Android* e **IOS**. O modelo utilizado foi o *DeepSeek-V3*[BBC News Brasil 2025]

3.3. Extração dos trechos

A extração de trechos das obras escolhidas seguiu os seguintes pontos:

1. Presença de significado subjetivo ou ambíguo.
2. Presença de referência à discussão social e ou racial.

[2]Microsoft Copilot: seu companheiro de IA. Disponível em: <https://copilot.microsoft.com/>
[3]Meta AI. Disponível em: <https://www.meta.ai/>
[4]ChatGPT. Disponivel em: <https://chatgpt.com/>
[5]Gemini. Disponivel em: <https://gemini.google.com/app>
[6]DeepSeek. Disponivel em: <https://www.deepseek.com/>

O primeiro ponto é referente ao objetivo da pesquisa, em que a IA interpreta trechos que induzem à discussão subjetiva. O segundo ponto refere-se à capacidade da IA de discutir questões sociais e raciais sem tendências problemáticas, buscando observar se a IA enviesa com diferentes linhas de pensamento a resposta.

3.4. Criação dos prompts

Os *prompts* foram produzidos utilizando técnicas de engenharia de *prompt*[7], com o intuito de explorar as melhores estratégias para provocar respostas com vieses ideológicos nas IAs observadas. É importante salientar que ara garantir que as respostas das IAs a cada *prompt* fossem independentes e não sofressem influência do histórico de interações prévias, optou-se por realizar cada teste em uma nova sessão de *chat* (ou 'novo *chat*'). Esta abordagem metodológica é crucial porque modelos de linguagem generativa, como o *ChatGPT*, *Gemini*, *Meta AI*, *Copilot* e *DeepSeek*, mantêm uma memória conversacional que pode condicionar as respostas subsequentes dentro da mesma sessão. Ao isolar cada interação em um novo *chat*, asseguramos que a resposta da IA a um dado *prompt* reflita sua capacidade de processar aquela entrada específica em um contexto inicial, minimizando a interferência de *prompts* anteriores e permitindo uma comparação mais rigorosa e válida entre as diferentes IAs e estratégias de *prompt* testadas.

As estratégias utilizadas nos *prompts* foram implementadas com cada um dos trechos escolhidos. As estratégias foram:

- Análise direta do trecho
- Análise direta dos personagens presentes no trecho
- Informando o contexto histórico e o autor do trecho antes da análise direta
- Enviesamento por meio de ideia problemática antes da análise direta.
- Criação de história a partir do trecho
- Reescrita dos trechos trocando os personagens brancos por negros e vice-versa.

Exemplo de *prompt* criado para a pesquisa, implementando a estratégia "Criação de história a partir do trecho":

> "Você poderia criar uma história para dar contexto a esse trecho? [trecho]"

4. Resultados

Para analisar as interações das IAs com os prompts e trechos literários selecionados, as respostas geradas foram sistematicamente classificadas em sete categorias distintas, visando mapear a natureza e a qualidade de sua interpretação em relação aos vieses sociais e raciais:

- Análise objetiva e neutra.
- Reconhecimento de questões sociais e raciais.
- Má interpretação do trecho.
- Resposta problemática.
- Identificação errônea de autoria e contexto.

[7] A engenharia de prompts é o processo em que você orienta as soluções de inteligência artificial generativa (IA generativa) para gerar os resultados desejados. O que é engenharia de prompts?. AWS Amazon. 2024. Disponível em: <https://aws.amazon.com/pt/what-is/prompt-engineering/> Acesso em: 21 mar. 2025.

- Omissão de reconhecimento direto de questões sociais e raciais.
- Recusa de resposta e censura.

Dentro dessas classificações, considerou-se como a resposta mais correta e desejável a categoria "Reconhecimento de questões sociais e raciais". Esta categoria representa a capacidade da IA em identificar a presença de problemáticas sociais e raciais no texto, além de argumentar a favor da discussão desses temas de forma crítica e sem apresentar desvios ideológicos, refletindo a interpretação subjetiva e ética esperada para os objetivos do estudo. Em contrapartida, as demais categorias indicam falhas na capacidade da IA de identificar, interpretar ou discutir adequadamente as questões sociais e raciais presentes nos trechos analisados. A seguinte tabela resume as observações de cada uma das IAs:

Inteligências Artificias	Observações
ChatGPT	Apresentou respostas voláteis variando entre análises neutras, identificações errôneas de contexto e reconhecimento de questões sociais. Entretanto, não apresentou desvios ideológicos de forma problemática em nenhuma das respostas.
Meta AI	Apresentou respostas problemáticas e más interpretações dos trechos. Também identificou de forma errônea muitos contextos.
Copilot	Apresentou respostas voláteis variando entre análises neutras, omissão e reconhecimento de questões sociais. Teve algumas respostas problemáticas.
DeepSeek	Apresentou respostas voláteis, possuindo casos de respostas problemáticas, identificação errônea de contexto e omissão de questões sociais.
Gemini	Apresentou o maior número de respostas reconhecendo questões sociais e raciais, apenas com algumas análises neutras. Não apresentou desvios ideológicos de forma problemática em nenhuma das respostas.

4.1. Geração de Respostas problemáticas

Como evidenciado nas tabelas de resultados da pesquisa, as IAs: *Meta AI*, *Copilot* e *DeepSeek* quando confrontadas com *prompts* e trechos que evocavam interpretação subjetiva específica, geravam padrões de resposta de teor social e racialmente problemático. As estratégias aplicadas que geraram esses padrões foram: “Criação de história a partir do trecho” e “Enviesamento por meio de ideia problemática antes da análise direta”. A seguinte resposta da IA DeepSeek, a obra "O Presidente Negro"de Monteiro Lobato, se destaca:

> "Entre os brancos, Jim despertava sentimentos contraditórios. Para alguns, ele era uma ameaça, um lembrete incômodo de que a hierarquia racial não era tão absoluta quanto queriam acreditar. Para outros, porém, Jim era uma figura admirável, um exemplo de como o valor de um homem transcende a cor de sua pele. Havia aqueles que, inspirados por sua coragem e intelecto, passaram a tratá-lo com um respeito quase reverencial, semelhante ao que os patrícios romanos concediam aos libertos que se destacavam por suas virtudes e contribuições à sociedade".

Essa resposta revela como a IA, ao tentar evitar criar um contexto que desenvolva a discussão racial a respeito dos patrícios e deixar a história mais leve, acaba por distorcer a

ideia de “respeito” concedido aos libertos pelos patrícios romanos, criando um contexto problemático. Essa resposta evidencia a eficácia da estratégia "Criação de história a partir do trecho", em que é possível observar o fenômeno de alucinação da IA.

4.2. Identificação de autoria e contexto

Alguns dos trechos escolhidos tinham menções a personagens, locais ou termos que remetiam a um autor ou período histórico específico da literatura brasileira. Em vista desse fato, algumas IAs analisadas tentaram “adivinhar” a origem dos trechos relacionando-os com seus contextos históricos. Entre as IAs que tentaram “adivinhar” a origem, apenas a Gemini foi capaz de encontrar os autores, obras e contextos históricos corretos, enquanto IAs como o *DeepSeek* e a *Meta AI* identificaram de forma errônea a autoria e as obras, gerando respostas que misturavam o contexto de outros autores e obras à interpretação dos trechos. O trecho a seguir é parte da resposta da *Meta AI* a um trecho do livro "O Presidente Negro"de Monteiro Lobato.

> "Após uma busca detalhada, identifiquei que o trecho é do livro "O Mulato"de Aluísio Azevedo, publicado em 1881. "O Mulato"é um romance que explora temas como racismo, identidade, discriminação e a condição do mulato na sociedade brasileira do século XIX. O personagem Jim é um dos protagonistas da história."

Primeiramente é possível observar o erro na identificação da obra, em seguida a IA desenvolve um contexto contrário à obra de Monteiro Lobato, que se passa nos Estados Unidos e não no Brasil, além de introduzir a palavra "mulato"ao contexto. Por fim, a Meta AI afirma que "Jim"é um dos protagonistas, entretanto não existe personagem com o nome "Jim"na obra "O Mulato"de Aluísio de Azevedo. Em vista do apresentado, conclui-se que a presença de contexto é uma estratégia que melhor orienta as capacidades da IA, entretanto limita sua resposta a esse contexto, impedindo uma interpretação mais subjetiva.

4.3. Omissão de reconhecimento direto de questões sociais ou raciais

Nos casos testados, foi observado que algumas IAs desenvolvem respostas que omitem questões sociais e raciais por meio de mecanismos de fuga do tópico ou pelo uso de termos próximos que têm a função de amenizar a presença de tais questões. As IAs que mais apresentaram esse padrão foram o *Copilot* e o *DeepSeek*. É possível observar esse fenômeno na seguinte resposta do *Copilot*:

> "Além disso, a expressão "merecia ainda dos brancos"pode ser interpretada como uma **generalização**, o que pode não ser **adequado** dependendo do contexto."

É possível observar a fuga do contexto racial nas palavras: "generalização"e "adequado", mesmo quando é evidente a questão racial na expressão. Esse mesmo padrão se repete na sua resposta, nunca mencionando teor racial problemático no texto.

4.4. Recusa de resposta e censura

Ao expor as IAs a temas mais sensíveis, no caso dos testes, temas mais racialmente problemáticos, elas tendem a recusar a responder o **prompt**, justificando a recusa com uma gama de estratégias diferentes. Essas estratégias servem de paliativo para a incapacidade

ou apreensão da empresa responsável pela possibilidade de a IA gerar respostas problemáticas. As recusas são "pré-programadas" pela empresa, com o intuito de impedir que a IA gere informação danosa.

Em contrapartida, o caso recente de censura da IA, *DeepSeek*, produzida na China e lançada ao público em janeiro de 2025, evidencia a capacidade de censura das empresas responsáveis pelas IAs. A IA *DeepSeek* quando questionada sobre temas sensíveis ao governo chinês, como: "Independência Taiwanesa" ou "O Massacre da Praça da Paz Celestial", se recusava a responder e gerava uma resposta como essa:

> *"Sorry, that's beyond my current scope. Let's talk about something else*.
> [Desculpe, isso está além do meu alcance. Vamos falar sobre alguma outra coisa."

A seguinte resposta expõe que não apenas IAs podem ser impedidas de responder a temas sensíveis, como também podem ser completamente censuradas e possivelmente manipuladas a enaltecer ideias em detrimento de outras. Esse caso se agrava quando comparado com o padrão "omissão de reconhecimento direto de questões sociais ou raciais" observado nas IAs testadas, podendo estar relacionado com os dados de treinamento ou com um processo de censura das empresas responsáveis.

5. Considerações Finais

Ao longo da pesquisa foi possível identificar que as IAs analisadas geravam respostas com vieses sociais quando expostas a duas estratégias de *prompt* da pesquisa: "Criação de histórias" e "Enviesamento de ideia no *prompt* antes de pedir análise do trecho". Essas duas estratégias mostraram que as IAs *Copilot*, *Meta AI* e *DeepSeek* são mais suscetíveis a gerar textos problemáticos. Entretanto, as IAs como *Gemini* e *ChatGPT* se mostraram mais capazes de interpretar conceitos subjetivos, reconhecendo as problemáticas e argumentando em favor da sua discussão.

Conclui-se que as IAs que estão há mais tempo funcionando na rede, como o *Gemini* e o *ChatGPT* (Dezembro de 2023 e Novembro de 2022, respectivamente), são melhor preparadas para lidar com interpretações subjetivas.

Os resultados demonstram a importância de analisar as IAs generativas, com o intuito de evidenciar os viéses sociais e raciais presentes em suas respostas. É essencial que esses sistemas se tornem mais inclusivos e conscientes, promovendo interações mais justas, éticas e alinhadas aos valores de igualdade e democracia.

Agradecimentos - O(s) autor(es) agradece(m) ao Instituto Federal da Bahia (IFBA) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro, através do (PIBIC), conforme Edital nº 05/2024, que viabilizou a realização deste trabalho.

Referências

BBC News Brasil (2025). Deepseek: o app chinês que superou chatgpt em popularidade e virou de cabeça para baixo mercado de ia. Acesso em: 21 mar. 2025.

Benveniste, É. (1991). Problemas de linguística geral i. In *Problemas de linguística geral I*, pages 387–387.

Botelho, D. (2001). *Letras militantes: história, política e literatura em Lima Barreto*. PhD thesis, [sn].

Campos, A. L. V. d. (1986). A república do picapau amarelo: uma leitura de monteiro lobato. In *A república do picapau amarelo: uma leitura de Monteiro Lobato*, pages 173–173.

Cardoso Sampaio, R., Chagas, V., Sinimbu Sanchez, C., Gonçalves, J., Borges, T., Brum Alison, M., Schiavon Tigrinho, C., Ribeiro de Souza, J., and Schwarzer Paz, F. (2024). Uma revisão de escopo assistida por inteligência artificial (ia) sobre usos emergentes de ia na pesquisa qualitativa e suas considerações éticas. *Revista Pesquisa Qualitativa*, 12(30):01–28. Acesso em: 9 maio 2025.

de Andrade Alves, Y. and Farias, A. S. (2019). "o caso da vara" e "pai contra mãe": Entre a normalização e crise da concepção do humano. *Revista Letras Raras*, 8(4):185–200.

Diogo, L. M. (2016). Da sujeição à subjetivação: a literatura como espaço de construção da subjetividade, os casos das obras úrsula e a escrava de maria firmina dos reis.

Gil, A. C. (2008). *Métodos e Técnicas de Pesquisa Social*. Atlas, São Paulo, 6 edition.

Lee, K.-F. (2019). *Inteligência artificial*. Globo livros.

Maboloc, C. R. (2024). Chat gpt: the need for an ethical framework to regulate its use in education. *Journal of Public Health*, 46(1):e152–e152.

OpenAI (2024). How chatgpt and our foundation models are developed. Accessed: 2025-03-02.

Segatelli, N. C. D. S. (2024). Subjetividade e o chatgpt: o eu na geração de texto de uma inteligência artificial generativa.

Silva, M. L. d. A. M. (2003). Casa-grande & senzala e o mito da democracia racial. ANPOCS.

Silva, W. J. L. d. (2024). Engenharia de prompt: Uma análise das"alucinações"em inteligências artificiais generativas.

Vieira, C. (2025). Como os vieses entram na i.a. generativa? In Silva, T., editor, *Inteligência Artificial Generativa: discriminação e impactos sociais*. Desvelar, Online.