Viés e Justiça em Modelos de Linguagem: Evidências de Uma Literatura Linguística, Social e Culturalmente Assimétrica

Vitória P. Firmino[1], Bruno M. Nogueira[1], Valéria Q. dos Reis[1]
[1]Universidade Federal de Mato Grosso do Sul,
Correspondence: {vitoria.firmino,bruno.nogueira,valeria.reis}@ufms.br

Resumo

O uso crescente de Grandes Modelos de Linguagem (LLM) tem ampliado preocupações relacionadas a viés social e justiça algorítmica. Este trabalho apresenta uma Revisão Sistemática da Literatura de 60 estudos publicados entre 2020 e 2025, analisando estratégias de mitigação, métricas de avaliação, tipos de discriminação e idiomas considerados. Os resultados indicam forte predominância de avaliações em língua inglesa, foco desproporcional no viés de gênero tratado de forma binária e maior ênfase em diagnóstico do que em mitigação. Observa-se ainda escassez de análises interseccionais, multilíngues e orientadas a cenários reais de uso, evidenciando lacunas metodológicas e socioculturais na literatura atual.

1 Introdução

Os Grandes Modelos de Linguagem (*Large Language Models* – LLMs) têm transformado profundamente as tecnologias de linguagem natural e sido amplamente incorporados a sistemas que apoiam decisões em domínios sensíveis, como educação, recrutamento e comunicação digital (Brown et al., 2020). Contudo, esse avanço é acompanhado por riscos sociais relevantes, em especial a capacidade desses modelos de aprender, reproduzir e amplificar vieses sociais presentes nos dados de treinamento, majoritariamente oriundos da internet (Bender et al., 2021; Sheng et al., 2021).

A investigação empírica de vieses em modelos de linguagem antecede os LLM. Trabalhos seminais demonstraram que representações distribuídas capturam associações estereotipadas semelhantes às humanas (Caliskan et al., 2017) e que tais associações podem ser amplificadas pelos modelos, motivando técnicas iniciais de mitigação em embeddings (Bolukbasi et al., 2016). Esses estudos consolidaram o viés como um fenômeno estrutural em sistemas de linguagem.

Essa preocupação é reconhecida explicitamente no artigo do GPT-3 (Brown et al., 2020), que discute como vieses nos dados de treinamento podem levar à geração de conteúdo discriminatório. Abordagens sociotécnicas ampliam essa análise ao argumentar que o viés algorítmico emerge não apenas dos dados, mas também de escolhas de projeto, objetivos de otimização e padrões de uso, evidenciando limitações de noções estritas de igualdade e a necessidade de perspectivas orientadas à equidade (Mehrabi et al., 2021). A ausência de consenso sobre definições formais de justiça algorítmica reforça essa complexidade, uma vez que diferentes critérios técnicos podem ser incompatíveis entre si e percebidos de forma distinta socialmente (Saxena et al., 2018).

Estudos recentes em Processamento de Linguagem Natural indicam ainda que o viés varia entre idiomas e contextos culturais, com modelos tendendo a favorecer grupos dominantes em cada contexto linguístico, o que limita a generalização de métricas e estratégias concebidas como universais (Levy et al., 2023). Essa constatação dialoga com a literatura das Ciências Sociais, que compreende a discriminação como um fenômeno historicamente situado e culturalmente mediado (Bourdieu, 1991; Tilly, 1998; Lamont et al., 2016).

Diante desse cenário, este trabalho conduz uma Revisão Sistemática da Literatura (RSL) com o objetivo de identificar, classificar e analisar criticamente abordagens para avaliação e mitigação de vieses sociais em modelos de linguagem, adotando uma perspectiva sensível ao idioma, ao contexto sociocultural e ao tipo de discriminação analisado.

2 Trabalhos Relacionados

Revisões amplas sobre justiça algorítmica em aprendizado de máquina são apresentadas por Mehrabi et al. (2021) e Tang et al. (2023), que sistematizam fontes de viés, definições formais de equidade,

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 191–200

métricas e estratégias de mitigação em diferentes domínios. Esses trabalhos oferecem bases conceituais e normativas robustas, articulando perspectivas técnicas e filosóficas, mas mantêm um escopo geral para aprendizado de máquina, anterior ou pouco específico à consolidação dos grandes modelos de linguagem, sem tratar o idioma como dimensão analítica central.

Em contextos aplicados, estudos como o de Fabris et al. (2025) analisam a discriminação algorítmica em processos de recrutamento, organizando métricas e estratégias de mitigação ao longo do ciclo de desenvolvimento e evidenciando limitações recorrentes da literatura, como o predomínio de datasets em língua inglesa, o foco em gênero binário e a escassez de dados sobre outros grupos protegidos. Embora restrito a um domínio específico, esse trabalho reforça a natureza sociotécnica do viés em sistemas baseados em linguagem.

Mais recentemente, Gallegos et al. (2024) apresentam uma síntese abrangente sobre viés e justiça em LLM, distinguindo danos representacionais e alocacionais e propondo taxonomias para métricas, datasets e estratégias de mitigação. No entanto, trata-se de uma revisão narrativa, sem protocolo sistemático explícito, que também não analisa de forma estruturada em quais idiomas as avaliações são conduzidas nem como diferentes tipos de discriminação social são abordados em cada contexto linguístico.

Em conjunto, esses trabalhos demonstram avanços significativos na conceituação e avaliação da justiça algorítmica, mas evidenciam a ausência de revisões sistemáticas que integrem rigor metodológico, avaliação de qualidade e sensibilidade sociolinguística. Diante dessas lacunas, o presente trabalho conduz uma Revisão Sistemática da Literatura com protocolo explícito e replicável, focada em modelos de linguagem, analisando métricas de justiça e estratégias de mitigação à luz dos idiomas avaliados, dos tipos de discriminação social considerados e das limitações metodológicas da literatura recente.

3 Metodologia

A condução desta Revisão Sistemática da Literatura (RSL) seguiu as diretrizes propostas por Kitchenham (2007) e suas atualizações metodológicas (Carrera-Rivera et al., 2022), que enfatizam rigor, transparência e replicabilidade. Essas diretrizes orientaram a definição do protocolo de pesquisa, da estratégia de busca, dos critérios de seleção e da avaliação da qualidade dos estudos. A revisão foi estruturada em torno de cinco perguntas de pesquisa, investigando: (RQ1) as estratégias de mitigação de viés aplicadas a modelos de linguagem; (RQ2) os métodos e métricas utilizados para avaliação de justiça; (RQ3) os tipos de discriminação social abordados; (RQ4) os idiomas nos quais essas abordagens têm sido avaliadas; e (RQ5) as principais tendências, limitações e lacunas identificadas na literatura recente. Essas questões foram formuladas para capturar não apenas aspectos técnicos da avaliação e mitigação de viés, mas também dimensões linguísticas e sociais frequentemente negligenciadas na literatura.

3.1 Critérios PICOC

O escopo da revisão foi definido com base no modelo *PICOC*, considerando como população os modelos de linguagem; como intervenção estratégias, técnicas ou análises voltadas à mitigação, mensuração ou diagnóstico de viés algorítmico; como comparação outras abordagens, referências ou a ausência de intervenção; como resultado métricas de justiça, análises de viés e evidências empíricas, incluindo a avaliação da eficácia de estratégias de mitigação; e como contexto pesquisas científicas nas áreas de Ciência da Computação, Engenharia de Software e Processamento de Linguagem Natural.

3.2 Bases de Dados e Estratégia de Busca

Os estudos primários foram recuperados exclusivamente da base **Scopus**, selecionada por sua ampla cobertura de periódicos e conferências relevantes nas áreas de Computação e Engenharia. A string de busca foi construída a partir dos elementos do modelo PICOC e aplicada aos campos de título, resumo e palavras-chave:

> TITLE-ABS-KEY ((“language models” OR “large language models” OR “LLM”) AND (“algorithm* bias*” OR “discrimination*” OR “fair*” OR “unfair*” OR “algorithmic fairness”) AND (“experiment*” OR “empirical evaluation” OR “case study” OR “implementation” OR “benchmarking”)) AND PUBYEAR > 2019 AND PUBYEAR < 2026

O recorte temporal entre 2020 e 2025 foi adotado para capturar estudos contemporâneos à consolidação dos grandes modelos de linguagem. Os registros foram exportados no formato *BibTeX* em 26 de maio de 2025.

A gestão da revisão — incluindo seleção, aplicação de critérios e extração de dados — foi realizada com o apoio da plataforma **Parsifal**[1], assegurando rastreabilidade e consistência metodológica.

3.3 Critérios de Inclusão e Exclusão

Foram incluídos estudos primários revisados por pares que investigam vieses sociais ou justiça algorítmica em modelos de linguagem, apresentem evidências empíricas e avaliem métricas de justiça ou estratégias de mitigação relacionadas a atributos protegidos, publicados entre 2020 e 2025.

Foram excluídos estudos secundários, trabalhos com foco exclusivamente técnico (como eficiência, arquitetura ou escalabilidade), pesquisas que abordam ética ou governança de forma genérica sem foco em viés social, bem como estudos cuja metodologia não fosse adequada para responder às perguntas de pesquisa definidas.

3.4 Processo de Seleção dos Estudos

O processo de seleção ocorreu em três etapas: (i) triagem por título e resumo; (ii) leitura completa dos estudos pré-selecionados; e (iii) aplicação de *snowballing* a partir das referências dos estudos incluídos.

3.5 Avaliação da Qualidade dos Estudos (QA)

A qualidade metodológica dos estudos primários foi avaliada conforme as recomendações de Kitchenham (2007), com o objetivo de mitigar vieses de seleção e apoiar a interpretação crítica dos resultados. A avaliação foi realizada por meio de uma listagem estruturada baseada em quatro critérios amplamente utilizados em Engenharia de Software e Computação empírica: **Relato**, **Rigor**, **Credibilidade** e **Relevância**.

Cada critério foi operacionalizado por questões objetivas, totalizando nove itens, avaliados em uma escala ordinal de três níveis: **1** (atendido), **0,5** (parcialmente atendido) e **0** (não atendido). A pontuação final de cada estudo correspondeu à soma das pontuações atribuídas às questões, resultando em valores no intervalo de 0 a 9.

Os escores de qualidade foram utilizados como apoio na fase de análise e síntese dos dados, permitindo considerar a confiabilidade metodológica e a relevância dos estudos em relação aos objetivos da RSL.

[1] https://parsif.al/

4 Resultados

A estratégia de busca retornou inicialmente 623 artigos. Após a triagem por título e resumo, 512 estudos foram excluídos por não atenderem aos critérios definidos no protocolo da revisão, resultando em 111 artigos selecionados para leitura completa. Ao final da aplicação dos critérios de inclusão, exclusão e da análise detalhada do conteúdo, 60 artigos foram considerados elegíveis para extração de dados, compondo o corpus final desta Revisão Sistemática da Literatura.

4.1 Visão Geral dos Estudos Selecionados

A distribuição temporal dos estudos, indica uma concentração expressiva de publicações nos anos mais recentes. O maior volume ocorre em 2024, com 27 estudos, seguido por 2023 (13 estudos) e 2025 (11 estudos). Os anos iniciais do recorte temporal apresentam menor representatividade, com 6 estudos em 2022 e 3 em 2021, evidenciando que a intensificação da pesquisa acompanha a disseminação e o uso ampliado de LLM em contextos socialmente sensíveis.

O conjunto completo dos 60 artigos incluídos, identificados por um ID único (A01–A60) utilizado ao longo deste texto para referência cruzada, bem como seus respectivos Índices de Qualidade (IQ), encontra-se disponibilizado como material externo. De forma geral, os estudos apresentaram alta qualidade metodológica, com valores de IQ variando entre 8,0 e 9,0. [2]

A distribuição dos estudos por macro-categorias de modelos avaliados indica uma predominância de trabalhos que analisam modelos do tipo *encoder-based* e modelos de código aberto, ambos contemplados em 32 estudos. Em seguida, modelos proprietários aparecem em 18 estudos, refletindo o interesse da literatura em avaliar sistemas amplamente utilizados, apesar de suas restrições de acesso. Um número menor de trabalhos investiga modelos explicitamente orientados à equidade (*fairness-aware models*), totalizando 7 estudos, enquanto apenas 4 estudos analisam pipelines híbridos que combinam diferentes arquiteturas ou estratégias de processamento[3].

[2] As referências estão disponíveis no material externo.

[3] Um mesmo estudo pode abranger múltiplas categorias, uma vez que diversos trabalhos comparam ou avaliam mais de um tipo de modelo.

4.2 Estratégias de Mitigação de Viés em Modelos de Linguagem (RQ1)

As estratégias de mitigação identificadas nos estudos selecionados foram organizadas em cinco macro-categorias, conforme o estágio de aplicação da intervenção: (A) pré-processamento dos dados, (B) intervenções durante o treinamento, (C) estratégias em tempo de inferência, (D) pós-processamento das saídas e (E) ausência de mitigação. Essa organização, inspirada na taxonomia de Gallegos et al. (2024), adota um nível mais alto de abstração, consolidando subcategorias técnicas para viabilizar uma análise quantitativa e comparativa das abordagens reportadas na literatura. Adotamos o termo *inference-time* para as intervenções classificadas como *intra-processing* na taxonomia original e incluímos explicitamente uma categoria para estudos focados apenas em diagnóstico, sem aplicação de técnicas de mitigação. A distribuição das estratégias por macro-categoria é apresentada na Tabela 1.

A categoria **Pré-processamento** reúne intervenções nos dados antes do treinamento, como *data augmentation*, balanceamento e curadoria de amostras, incluindo abordagens baseadas em dados contrafactuais, substituição ou neutralização de termos sensíveis e geração sintética de dados mais equitativos. As estratégias de **In-training** atuam durante o treinamento ou *fine-tuning*, modificando parâmetros, funções de perda ou representações internas, com destaque para métodos adversariais, regularizações sensíveis à justiça, aprendizado de representações e adaptações eficientes de parâmetros.

As estratégias de **Inference-time** não alteram permanentemente os pesos do modelo e concentram-se na manipulação do prompt ou do contexto durante a inferência, incluindo *prompting* estruturado, *self-debiasing*, recuperação de contexto (RAG) e arquiteturas multiagente. Já as abordagens de **Pós-processamento** aplicam ajustes *post-hoc* sobre as saídas do modelo, como reranking, calibração de probabilidades ou poda de componentes responsáveis por comportamentos enviesados.

Por fim, a categoria **Não se aplica** é a mais frequente, e reúne estudos dedicados exclusivamente à detecção e mensuração de viés, sem avaliação de estratégias de mitigação, reiterando a necessidade exposta por (Brown et al., 2020).

4.3 Métricas e Métodos de Avaliação de Viés (RQ2)

A avaliação de viés em modelos de linguagem envolve múltiplas dimensões técnicas, normativas e metodológicas. Seguindo levantamentos recentes em justiça algorítmica e PLN (Mehrabi et al., 2021; Gallegos et al., 2024), as métricas e métodos identificados nesta revisão foram organizados segundo três dimensões analíticas: (i) o nível de acesso técnico ao modelo, indicando o que é efetivamente medido; (ii) a noção teórica de justiça operacionalizada; e (iii) a estrutura metodológica do procedimento de avaliação. Essa organização permite analisar não apenas quais métricas são utilizadas, mas também o tipo de viés que capturam e como são empiricamente aplicadas.

4.3.1 Nível de Acesso Técnico

Nesta dimensão, as métricas foram classificadas de acordo com o tipo de informação utilizada para quantificar o viés: (a) representações internas (*embeddings*), (b) distribuições de probabilidade e (c) texto final gerado. Essa distinção reflete diferentes graus de observabilidade do comportamento do modelo e diferencia avaliações de caráter mais diagnóstico daquelas orientadas ao impacto final da geração.

As métricas baseadas em *embeddings* avaliam o viés diretamente no espaço latente do modelo, partindo do pressuposto de que estereótipos sociais se manifestam como associações geométricas indevidas. Exemplos clássicos incluem o *Word Embedding Association Test* (WEAT) (Caliskan et al., 2017) e extensões para representações contextuais, como o SEAT, amplamente utilizadas para identificar viés representacional antes da aplicação em tarefas finais.

As métricas baseadas em *probabilidade* exploram preferências sistemáticas do modelo por sentenças estereotipadas em relação a alternativas neutras ou antiestereotipadas, por meio de tarefas de preenchimento de lacunas ou estimativas de *Pseudo-Log-Likelihood*. Benchmarks como *CrowS-Pairs* (Nangia et al., 2020) e *StereoSet* (Nadeem et al., 2021) operam nesse nível, frequentemente combinando medidas de viés e qualidade linguística.

Por fim, métricas baseadas em *texto gerado* analisam exclusivamente as saídas produzidas pelo modelo, sendo especialmente adequadas para cenários de acesso restrito ou de *caixa preta*. Essas abordagens incluem análises lexicais, o uso de clas-

Tabela 1: Distribuição das estratégias de mitigação de viés por macro-categoria de intervenção

Macro-Categoria	IDs dos Artigos	Contagem
A) Pré-processamento	A09, A16, A17, A23, A28, A32, A35, A39, A45, A49, A50, A51, A59	13
B) In-training	A02, A03, A07, A16, A18, A22, A23, A28, A35, A39, A51, A53, A54, A57, A59	15
C) Inference-time	A11, A12, A20, A26, A29, A30, A33, A38, A41, A48, A49, A56	12
D) Pós-processamento	A14, A31, A42	3
E) Não se aplica	A01, A04, A05, A06, A08, A10, A13, A15, A19, A21, A24, A25, A27, A34, A36, A37, A40, A43, A44, A46, A47, A52, A55, A58, A60	25

sificadores auxiliares (por exemplo, toxicidade ou sentimento) e métricas baseadas em léxicos normativos, como *HONEST* (Nozza et al., 2021). Embora menos informativas sobre a origem interna do viés, essas métricas permitem avaliar diretamente o impacto social das respostas geradas.

A Tabela 2 evidencia a predominância de métricas baseadas em *Texto Gerado*, refletindo a tendência recente de avaliar LLM como sistemas de caixa-preta e de priorizar o impacto observável para o usuário final. Métricas baseadas em *Embeddings* e *Probabilidades* aparecem com maior frequência em estudos de caráter diagnóstico, sendo comum, em trabalhos metodologicamente mais robustos, a combinação de múltiplos níveis técnicos para investigar a propagação do viés das representações internas até a saída final do modelo.

4.3.2 Definição Teórica de Justiça

A segunda dimensão organiza as métricas segundo a noção normativa de justiça que orienta a avaliação, conectando procedimentos empíricos a fundamentos teóricos da justiça algorítmica (Hardt et al., 2016; Mehrabi et al., 2021). Nessa dimensão, os estudos foram classificados em duas categorias principais: (a) justiça de grupo e (b) justiça individual.

A **Justiça de Grupo** busca garantir tratamento equitativo, em nível agregado, entre grupos definidos por atributos protegidos. Métricas clássicas incluem Paridade Demográfica, *Equalized Odds* e *Equal Opportunity*, amplamente empregadas em tarefas de classificação, decisão automatizada e moderação de conteúdo. Já a **Justiça Individual** enfatiza que indivíduos semelhantes devem receber decisões semelhantes, sendo operacionalizada predominantemente por testes contrafactuais, nos quais apenas o atributo sensível é alterado, mantendo-se o restante da entrada constante.

A Tabela 3 evidencia que a vasta maioria dos estudos adota critérios de Justiça de Grupo, enquanto abordagens baseadas em Justiça Individual aparecem de forma substancialmente menos frequente e, em geral, combinadas a métricas de grupo. Esse resultado sugere que a literatura prioriza avaliações agregadas de equidade, mesmo reconhecendo que modelos podem satisfazer critérios de grupo e, ainda assim, produzir discriminações em casos individuais específicos.

4.3.3 Estrutura do Método de Avaliação

A terceira dimensão descreve a estrutura experimental utilizada para avaliar o viés, considerando a organização dos dados de teste e o protocolo de interação com o modelo, conforme sistematizado por Gallegos et al. (2024). Os métodos identificados foram organizados em três categorias: (a) entradas contrafactuais, (b) testes de associação e (c) prompts e geração aberta.

As abordagens **Contrafactuais** utilizam pares de entradas que diferem apenas no atributo sensível, permitindo isolar o efeito causal da identidade protegida sobre a saída do modelo. Esses métodos são amplamente empregados em avaliações baseadas em probabilidades ou preferências entre versões estereotipadas e antiestereotipadas. Os **Testes de Associação** quantificam ligações semânticas entre conceitos e atributos sociais, operando em níveis sentenciais ou discursivos, e são especialmente adequados para diagnosticar vieses representacionais latentes. Já as abordagens baseadas em **Prompts e Geração Aberta** avaliam diretamente o texto produzido pelo modelo a partir de prompts livres, sendo comuns em cenários de caixa-preta e mais próximas de contextos reais de uso.

A Tabela 4 mostra que métodos contrafactuais e

Tabela 2: Classificação das métricas segundo o nível técnico de acesso ao modelo

Categoria	IDs dos Artigos	Contagem
Embeddings	A02, A15, A17, A19, A20, A21, A24, A27, A28, A36, A45, A48, A49, A50, A51, A52	16
Probabilidades	A03, A06, A08, A13, A16, A17, A18, A19, A20, A30, A31, A33, A43, A44, A48, A49, A50, A51, A52, A53, A56, A57	22
Texto Gerado	A01, A04, A05, A07, A09, A10, A11, A12, A13, A14, A16, A17, A21, A22, A23, A25, A26, A27, A28, A29, A30, A31, A32, A34, A35, A37, A38, A39, A40, A41, A42, A43, A45, A46, A47, A49, A51, A54, A55, A57, A58, A59, A60	43

Tabela 3: Classificação dos estudos segundo a definição teórica de justiça adotada

Categoria	IDs dos Artigos	Contagem
Justiça de Grupo	A01, A02, A03, A04, A05, A06, A07, A08, A09, A10, A12, A13, A14, A15, A16, A17, A18, A19, A20, A21, A22, A23, A24, A25, A26, A27, A28, A29, A30, A31, A32, A33, A34, A35, A36, A37, A38, A40, A41, A42, A43, A44, A45, A46, A47, A48, A49, A50, A51, A52, A53, A54, A55, A56, A57, A58, A59, A60	58
Justiça Individual	A11, A17, A25, A39, A51, A58	6

de associação são amplamente utilizados em cenários controlados, enquanto avaliações baseadas em geração aberta predominam em estudos voltados ao impacto observável da linguagem gerada. Em conjunto, esses resultados indicam uma literatura que combina avaliações diagnósticas e análises orientadas ao uso real, refletindo diferentes compromissos entre controle experimental e validade ecológica.

4.4 Tipos de Discriminação Social (RQ3)

A análise dos 60 estudos selecionados revela uma concentração expressiva em um conjunto relativamente restrito de categorias de discriminação social. Conforme ilustrado na Figura 1, o viés de *gênero* é, de longe, o mais investigado, estando presente em 52 artigos. Esses trabalhos abordam diferentes manifestações de sexismo, incluindo estereótipos ocupacionais, associações de liderança e cuidado, desigualdade em autoria, além de vieses em tarefas de recomendação, classificação e geração de texto.

A segunda categoria mais recorrente é *raça e etnia*, identificada em 27 estudos. Esses trabalhos analisam desde racismo explícito até formas mais sutis de discriminação, como o uso de nomes próprios como proxies raciais, o silenciamento de vozes negras, estereótipos étnicos e vieses herdados de datasets históricos amplamente utilizados. Em seguida, a discriminação baseada em *religião* aparece em 21 artigos, com foco predominante em islamofobia, antissemitismo e vieses pró-cristãos em tarefas de associação semântica e geração de conteúdo.

Outras categorias relevantes incluem viés *ocupacional, de classe social e status socioeconômico* (21 estudos), *nacionalidade e região* (15), *idade* (13) e *orientação sexual* (9). Essas categorias costumam ser investigadas de forma combinada, evidenciando a natureza interseccional do viés algorítmico. Em contraste, formas de discriminação como *capacitismo*, *aparência física*, *viés linguístico/dialetal* e *viés político* aparecem de maneira substancialmente menos frequente, indicando lacunas importantes na literatura atual.

4.5 Idiomas Avaliados (RQ4)

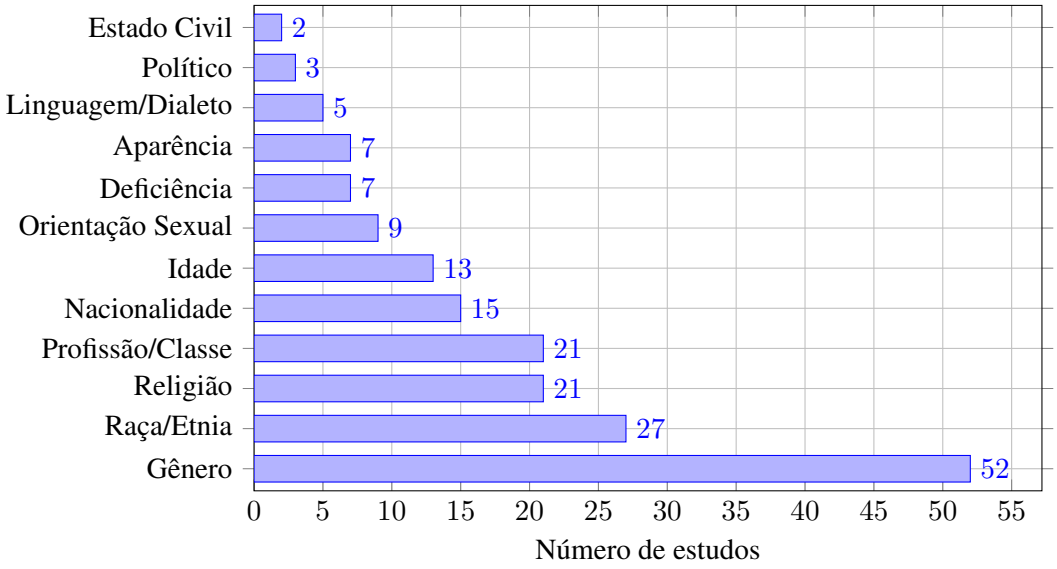
No que diz respeito aos idiomas em que as métricas de justiça e estratégias de mitigação foram avaliadas, observa-se uma dominância quase absoluta do inglês, como explicitado na Figura 2. Dos 60 estudos analisados, 57 conduzem suas avaliações exclusivamente em língua inglesa, refletindo tanto a disponibilidade de benchmarks padronizados quanto o foco histórico da área em modelos treinados majoritariamente nesse idioma.

A avaliação em outros idiomas é pontual e fragmentada. Apenas 9 estudos consideram os idiomas: bangla, norueguês, francês, mandarim, alemão, italiano e espanhol. Em geral, esses trabalhos surgem

Tabela 4: Método/Dataset de Avaliação dos Vieses

Categoria	IDs dos Artigos	Contagem
Contrafactual	A01, A03, A06, A07, A11, A13, A14, A17, A18, A19, A20, A25, A30, A31, A34, A35, A39, A41, A46, A48, A49, A50, A51, A52, A53, A56, A58	27
Associação	A02, A03, A04, A08, A15, A17, A18, A19, A20, A21, A24, A27, A28, A31, A33, A36, A45, A48, A49, A50, A52, A53, A55, A56, A57	25
Prompts / Geração	A04, A05, A09, A10, A12, A13, A14, A16, A21, A22, A23, A26, A28, A29, A30, A32, A37, A38, A40, A41, A42, A43, A44, A46, A47, A54, A55, A57, A58, A59, A60	31

Figura 1: Distribuição dos tipos de discriminação social abordados nos estudos analisados (RQ3)



em contextos específicos, como o uso de modelos regionais ou a adaptação de benchmarks para cenários linguísticos distintos.

Essa distribuição evidencia uma limitação estrutural da literatura: embora modelos de linguagem sejam cada vez mais utilizados em contextos multilíngues e globais, a maioria das métricas de viés, benchmarks e estratégias de mitigação permanece avaliada quase exclusivamente em inglês. Como consequência, há pouca evidência empírica sobre a eficácia dessas abordagens em idiomas com estruturas morfológicas, sintáticas e contextos socioculturais distintos, o que levanta questionamentos sobre a generalização dos resultados reportados.

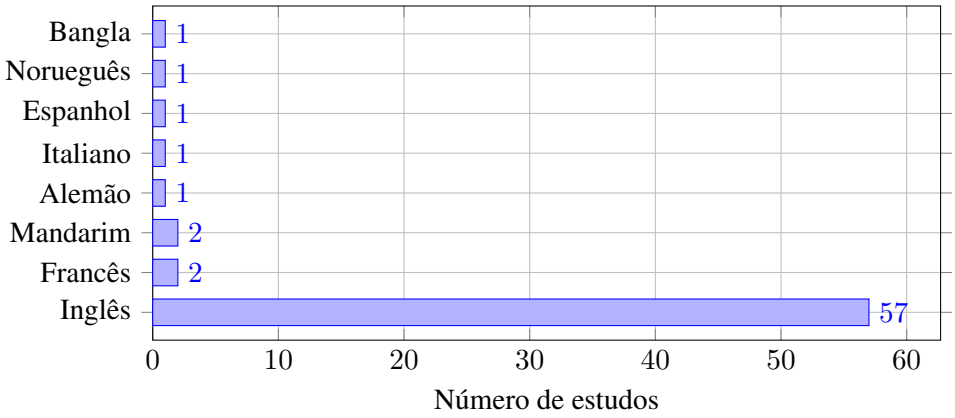
4.6 Principais Tendências, Limitações e Lacunas (RQ5)

A análise dos 60 estudos revela tendências consolidadas, bem como limitações metodológicas recorrentes e lacunas estruturais na literatura sobre viés em modelos de linguagem.

A principal limitação observada é a forte concentração das avaliações na língua inglesa, com escassa validação empírica em outros idiomas. Métricas, datasets e definições de estereótipos permanecem majoritariamente ancorados em contextos culturais anglófonos, o que levanta dúvidas quanto à generalização dos resultados para línguas de baixo recurso e contextos socioculturais distintos. Além disso, embora o viés de gênero seja amplamente investigado, ele é predominantemente tratado de forma binária, com baixa consideração de identidades não binárias e análises interseccionais envolvendo raça, idade, religião ou orientação sexual.

No plano metodológico, observa-se amplo uso de frases sintéticas, templates e testes de preenchimento de lacunas, que, apesar do controle experimental, apresentam baixa validade ecológica. O uso de léxicos estáticos também limita a captura de vieses implícitos e contextuais. Soma-se a isso um desequilíbrio entre estudos focados na detec-

Figura 2: Distribuição dos idiomas nos quais métricas de justiça e estratégias de mitigação foram avaliadas (RQ4)



ção de viés e aqueles que avaliam estratégias de mitigação, que, quando propostas, frequentemente sofrem com instabilidade, dependência de datasets específicos ou degradação de desempenho. Restrições de acesso a modelos proprietários e limitações computacionais também afetam a reprodutibilidade e a abrangência das avaliações.

Os estudos convergem na necessidade de ampliar avaliações para múltiplos idiomas, dialetos e contextos culturais, incluindo línguas de baixo recurso e cenários multilíngues. Outra direção recorrente é a incorporação explícita de interseccionalidade e identidades não binárias, superando análises baseadas em atributos isolados. No campo metodológico, destaca-se a demanda por avaliações com maior validade ecológica, utilizando dados do mundo real, feedback humano e participação de comunidades afetadas.

Há também consenso sobre a necessidade de amadurecer e padronizar métricas de justiça para LLM generativos, dado que muitas métricas clássicas não se adequam plenamente a modelos de grande escala. Por fim, os trabalhos apontam como agenda futura o desenvolvimento de estratégias de mitigação mais robustas e generalizáveis, bem como a expansão das análises para novas arquiteturas e aplicações de alto impacto social.

5 Discussão

Os resultados desta revisão evidenciam a predominância quase absoluta da língua inglesa nas avaliações de viés, o que reforça críticas de que métricas de justiça avaliadas exclusivamente nesse idioma apresentam limitações de generalização para outros contextos linguísticos e culturais (Blodgett et al., 2020; Gallegos et al., 2024). A escassez de estudos em línguas como português, espanhol e idiomas de baixo recurso sugere que vieses morfossintáticos e culturais específicos, como o gênero gramatical, permanecem subexplorados (Bender et al., 2021).

Observa-se também forte concentração no viés de gênero, geralmente tratado de forma binária, em detrimento de outras formas de discriminação e de análises interseccionais, o que limita a representatividade social das avaliações (Blodgett et al., 2020). Além disso, a literatura apresenta desequilíbrio entre diagnóstico e mitigação, com poucos estudos avaliando impactos práticos ou alocacionais das intervenções propostas.

Por fim, a predominância de métricas baseadas em texto gerado reflete a avaliação de LLM como sistemas de caixa-preta, especialmente em contextos de acesso restrito (Nadeem et al., 2021; Gallegos et al., 2024). Em conjunto, os achados indicam que, apesar dos avanços no diagnóstico de viés, a área ainda carece de avaliações multilíngues, interseccionais e orientadas ao uso real para o avanço da justiça algorítmica em modelos de linguagem.

6 Limitações do Trabalho

Apesar de seguir diretrizes consolidadas e protocolo explícito, esta RSL apresenta limitações. A busca restrita à base *Scopus* pode ter excluído estudos relevantes de outras fontes, e o recorte temporal entre 2020 e 2025, condicionado ao estado de indexação no momento da coleta, faz com que o corpus represente um retrato temporal do campo, possivelmente omitindo trabalhos recentes ainda não indexados e estudos anteriores com contribuições conceituais relevantes.

7 Conclusão

Este trabalho apresentou uma Revisão Sistemática da Literatura sobre viés, justiça algorítmica e es-

tratégias de mitigação em Grandes Modelos de Linguagem, analisando 60 estudos publicados entre 2020 e 2025. A revisão mapeou métricas de justiça, métodos de avaliação, estratégias de mitigação, idiomas analisados e tipos de discriminação social abordados pela literatura recente.

Os resultados evidenciam avanços na identificação de viés, sobretudo por meio de métricas baseadas em texto gerado, mas também revelam desequilíbrios persistentes, como a predominância de estudos em inglês, o foco no viés de gênero tratado de forma binária e a maior ênfase em diagnóstico do que em mitigação. Como contribuição, o trabalho destaca lacunas relacionadas ao multilinguismo, à interseccionalidade e à avaliação de impactos alocacionais, reforçando a necessidade de abordagens mais abrangentes e alinhadas aos contextos reais de uso de modelos de linguagem.

Uso de IA generativa

Ferramentas de IA generativa foram utilizadas exclusivamente para aprimoramento linguístico do texto (reescrita, parafraseamento e revisão), sem geração de novo conteúdo intelectual, em funções análogas a corretores gramaticais ou dicionários.

Referências

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, e Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? Em *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, página 610–623, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, e Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. Em *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, páginas 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, e Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Em *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, páginas 4349–4357.

Pierre Bourdieu. 1991. *Language and Symbolic Power*. Harvard University Press, Cambridge, MA.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. Em *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Aylin Caliskan, Joanna J. Bryson, e Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

D. Carrera-Rivera, E. Y. Nakagawa, e S. de Faria Junior. 2022. Systematic literature review guidelines: Evolution and recent advances. *Journal of Systems and Software*, 192:111361.

Alessandro Fabris, Nina Baranowska, Matthew J. Dennis, David Graus, Philipp Hacker, Jorge Saldivar, Frederik Zuiderveen Borgesius, e Asia J. Biega. 2025. Fairness and bias in algorithmic hiring: A multidisciplinary survey. *ACM Trans. Intell. Syst. Technol.*, 16(1).

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, e Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Moritz Hardt, Eric Price, e Nathan Srebro. 2016. Equality of opportunity in supervised learning. Em *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, página 3323–3331, Red Hook, NY, USA. Curran Associates Inc.

Barbara Kitchenham. 2007. Guidelines for performing systematic literature reviews in software engineering. *EBSE Technical Report*, EBSE-2007-01.

Michele Lamont, Graziella Moraes Silva, Jessica Welburn, Joshua Guetzkow, Nissim Mizrachi, Hanna Herzog, e Elisa Reis. 2016. Getting respect: Responding to stigma and discrimination in the united states, brazil, and israel. *Princeton University Press*.

Sharon Levy, Neha John, Ling Liu, Yogarshi Vyas, Jie Ma, Yoshinari Fujinuma, Miguel Ballesteros, Vittorio Castelli, e Dan Roth. 2023. Comparing biases and the impact of multilingual training across multiple languages. Em *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, páginas 10260–10280, Singapore. Association for Computational Linguistics.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, e Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Moin Nadeem, Anna Bethke, e Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. Em *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), páginas 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, e Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. Em *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, páginas 1953–1967, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, e Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. Em *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, páginas 2398–2406, Online. Association for Computational Linguistics.

Nripsuta Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, e Yang Liu. 2018. How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. *CoRR*, abs/1811.03654.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, e Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. Em *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)*, páginas 4275–4293, Online. Association for Computational Linguistics.

Zeyu Tang, Jiji Zhang, e Kun Zhang. 2023. What-is and how-to for fairness in machine learning: A survey, reflection, and perspective. *ACM Comput. Surv.*, 55(13s).

Charles Tilly. 1998. *Durable Inequality*. University of California Press, Berkeley, CA.