# **Benchmarks de Viés em Modelos de Inteligência Artificial: Foco na Fase de Treinamento**

## **I. Introdução ao Viés em IA e Benchmarking**

### **A. Definindo Viés em Modelos de IA: Do Erro Estatístico ao Dano Social**

No contexto da inteligência artificial (IA) e do aprendizado de máquina (ML), o termo "viés" refere-se fundamentalmente a erros sistemáticos ou desvios nas previsões de um modelo em relação aos resultados reais.1 Tecnicamente, pode ser entendido como uma forma de imprecisão que ocorre quando um modelo faz suposições ou generalizações com base em dados limitados ou incompletos.1 Na pesquisa tradicional de ML, o viés está intrinsecamente ligado à variância, onde o viés representa a distância (erro) entre a predição e o valor alvo real, enquanto a variância mede a dependência das predições da aleatoriedade na amostragem dos dados de treinamento.2 Ruído, viés e variância são os três componentes que contribuem para a degradação do desempenho do modelo e afetam negativamente sua capacidade de generalização para novos dados.2

No entanto, a compreensão do viés em IA transcendeu essa definição puramente estatística para abranger preocupações éticas e sociais mais amplas. O viés em modelos de IA pode levar a conclusões imprecisas e, mais criticamente, a decisões injustas ou discriminatórias.1 Ele não é uma propriedade inerente à IA em si, mas sim um reflexo das escolhas humanas, das limitações dos dados e das estruturas sociais embutidas nos dados de treinamento e no design do modelo.5 Os algoritmos, em vez de serem neutros, tendem a amplificar os vieses presentes nos dados sobre os quais são treinados.5 As fontes de viés são variadas e podem surgir em qualquer etapa do pipeline de ML, incluindo coleta de dados, curadoria, rotulagem, treinamento e implantação.2

As manifestações de viés são diversas e podem ter consequências significativas em vários setores. Por exemplo, modelos de recrutamento treinados com dados históricos de funcionários podem discriminar candidatos de origens sub-representadas.1 Na justiça criminal, modelos de previsão de reincidência podem tratar certos grupos demográficos de forma injusta.1 Na saúde, modelos de diagnóstico podem ser menos precisos para certas condições ou populações.1 Tecnologias de reconhecimento facial demonstraram taxas de erro significativamente mais altas para indivíduos de pele mais escura, especialmente mulheres.7 Esses exemplos ilustram como o viés algorítmico pode perpetuar e até exacerbar as desigualdades existentes.

É importante distinguir entre viés explícito e implícito. O viés explícito refere-se a preconceitos conscientes e intencionais, enquanto o viés implícito opera inconscientemente, moldado por condicionamento social e exposição cultural.7 Ambos os tipos podem ser codificados em sistemas de IA, seja através dos dados ou das suposições dos desenvolvedores.7 Os vieses implícitos são particularmente desafiadores porque podem influenciar o comportamento do modelo mesmo quando os criadores rejeitam conscientemente crenças discriminatórias.7

A evolução na compreensão do viés em IA é notável. Inicialmente enquadrado principalmente dentro do dilema técnico de viés-variância 2, o foco se expandiu criticamente para englobar os *danos sociais* que podem resultar de sistemas enviesados. Esses danos foram categorizados como danos de alocação (oportunidades ou recursos negados a certos grupos), danos de qualidade de serviço (o sistema funciona pior para alguns grupos), danos de estereotipagem (o sistema reforça estereótipos) e danos de apagamento (o sistema ignora a existência de certos grupos).14 Essa mudança reflete o reconhecimento de que o "viés" em IA não é apenas um conceito estatístico, mas um fenômeno sócio-técnico complexo com consequências éticas profundas e impactos no mundo real 1, exigindo abordagens de avaliação que vão além da simples medição de erros técnicos.

### **B. A Necessidade e Função dos Benchmarks de Viés**

Diante da prevalência e do impacto potencial do viés em IA, surgiu a necessidade de ferramentas e metodologias para avaliá-lo sistematicamente. Os benchmarks de viés cumprem essa função, fornecendo conjuntos de dados e protocolos padronizados para quantificar e comparar vieses em diferentes modelos de IA.10 Eles servem como instrumentos essenciais para auditar modelos 7, identificar tipos específicos de viés (como viés de gênero ou racial) 8, medir disparidades de desempenho entre diferentes subgrupos demográficos 10 e monitorar o progresso no desenvolvimento de sistemas de IA mais justos e equitativos.19 A existência e o uso generalizado desses benchmarks sinalizam um reconhecimento importante no campo: a justiça não emerge espontaneamente em sistemas de IA; ela requer medição deliberada, avaliação crítica e intervenção direcionada.16

No entanto, é crucial reconhecer que os próprios benchmarks não são infalíveis e possuem limitações. Questões sobre a qualidade e representatividade dos dados do benchmark foram levantadas, como no caso do CrowS-Pairs.27 Existe o risco de contaminação, onde dados de benchmark podem ter sido incluídos nos dados de treinamento de grandes modelos de linguagem (LLMs), inflando artificialmente as pontuações de desempenho.31 Muitos benchmarks existentes têm um foco predominantemente Anglo/US-cêntrico, limitando sua aplicabilidade e relevância em outros contextos culturais.28 Além disso, as métricas usadas dentro dos benchmarks podem, elas mesmas, incorporar vieses ou falhar em capturar a complexidade total da justiça.33 Alguns benchmarks também podem se tornar "saturados" à medida que os modelos melhoram, deixando de ser medidas desafiadoras de desempenho.32 Essas críticas destacam que o benchmarking de viés é um campo em evolução e não um problema resolvido. A confiança cega nas pontuações de benchmark é desaconselhada; em vez disso, os profissionais devem avaliar criticamente a construção do benchmark, suas falhas conhecidas e o tipo específico de viés que ele mede para interpretar os resultados de forma significativa.27

### **C. Significância de Abordar o Viés Durante o Treinamento do Modelo**

O viés em sistemas de IA pode ser detectado e abordado em diferentes estágios do ciclo de vida do desenvolvimento. A detecção pode ser reativa, observando-se um desempenho ruim ou resultados injustos após a implantação, ou proativa, tentando prevenir o viés antes que ele se manifeste.5 Abordar o viés *durante a fase de treinamento* do modelo, conhecido como mitigação *in-processing*, representa uma abordagem proativa fundamental.40

A importância de focar na fase de treinamento deriva do fato de que o viés frequentemente se origina nas escolhas feitas antes ou durante este estágio, seja nos dados de treinamento ou nos próprios algoritmos.2 Vieses incorporados durante o treinamento, devido a dados desequilibrados, históricos ou algoritmos inadequados, tendem a se propagar e podem ser amplificados nas etapas subsequentes e nas aplicações downstream.5 Portanto, intervir durante o treinamento permite abordar as causas raízes do viés, em vez de apenas tratar seus sintomas após o fato.

As técnicas de mitigação *in-processing* modificam o próprio algoritmo de aprendizado ou sua função objetivo para incorporar considerações de justiça diretamente no processo de otimização do modelo.16 Isso contrasta com as abordagens de pré-processamento, que modificam os dados antes do treinamento 40, e as abordagens de pós-processamento, que ajustam as saídas do modelo após o treinamento.40 Ao integrar a justiça ao processo de aprendizado, os métodos *in-processing* visam construir modelos que sejam inerentemente mais justos, influenciando diretamente as representações internas do modelo e sua lógica de tomada de decisão para serem menos enviesadas desde o início. Isso permite um trade-off explícito entre a precisão do modelo e as metas de justiça durante a otimização.2

## **II. Uma Taxonomia de Vieses Relevantes para o Treinamento de IA**

Compreender os diferentes tipos de viés que podem surgir durante o treinamento de modelos de IA é crucial para selecionar benchmarks apropriados e estratégias de mitigação eficazes. Esses vieses podem ser amplamente categorizados com base em sua origem: dados, algoritmo/modelo ou fatores humanos.

### **A. Vieses Centrados em Dados (Originários dos Dados de Treinamento)**

A qualidade e a composição dos dados de treinamento são fontes primárias de viés em modelos de IA. Vários tipos de viés relacionados a dados foram identificados:

* **Viés de Amostragem/Seleção (Sampling/Selection Bias):** Ocorre quando os dados usados para treinar o modelo não são representativos da população ou distribuição do mundo real onde o modelo será aplicado.4 Isso pode acontecer devido a métodos de coleta não aleatórios ou que favorecem sistematicamente certos grupos ou características.1 Inclui o *viés de cobertura* (coverage bias), onde certos subgrupos da população não estão presentes ou estão sub-representados nos dados, e o *viés de amostragem* (sampling bias) propriamente dito, resultante da falta de randomização adequada durante a coleta.13 Modelos treinados com esse viés podem ter um desempenho ruim quando confrontados com cenários não vistos nos dados de treinamento, levando a uma baixa generalização.2
* **Viés de Medição (Measurement Bias):** Refere-se a erros sistemáticos introduzidos durante o processo de coleta, registro ou interpretação dos dados.1 Isso pode resultar de equipamentos de medição defeituosos, protocolos inconsistentes ou erros na rotulagem dos dados, levando a valores que não refletem com precisão os atributos verdadeiros que estão sendo medidos.2 No contexto do dataset COMPAS, por exemplo, críticas apontam para possíveis vieses de medição, já que os registros de prisão não são um proxy perfeito para a criminalidade real, e a própria definição de reincidência pode ser enviesada.50
* **Viés de Representação (Representation Bias):** Surge da sub-representação ou super-representação de certos grupos, características ou cenários nos dados de treinamento.5 Exemplos clássicos incluem conjuntos de dados de reconhecimento facial treinados predominantemente em indivíduos de pele clara, resultando em desempenho inferior para pessoas de pele mais escura 7, ou dados para veículos autônomos que contêm muito mais cenários de dias ensolarados do que chuvosos ou noturnos.5 Isso pode levar o modelo a aprender padrões espúrios associados aos grupos super-representados.2
* **Viés Histórico (Historical Bias):** Ocorre quando os dados de treinamento refletem preconceitos, desigualdades ou práticas discriminatórias passadas da sociedade.2 Um modelo treinado com esses dados pode aprender e perpetuar essas injustiças históricas, mesmo que elas não sejam mais consideradas aceitáveis.7 Um exemplo é treinar um algoritmo de contratação com dados históricos de uma empresa que predominantemente contratava homens para certas funções.7
* **Viés de Exclusão (Exclusion Bias):** Resulta da remoção inadequada de amostras ou características importantes do conjunto de dados, que poderiam ter um impacto significativo no desempenho ou na justiça do modelo.4 Isso pode ocorrer por exclusão sistemática ou pela exclusão inadvertida de dados considerados sem importância, mas que na verdade agregam valor.5
* **Viés de Relato (Reporting Bias):** Surge quando a frequência com que eventos ou fenômenos são registrados nos dados não reflete sua frequência real no mundo.6 Um exemplo citado é um conjunto de dados de detecção de fraude onde quase todas as investigações em uma região remota foram rotuladas como fraude devido a práticas específicas dos investigadores, não à prevalência real de fraude.6
* **Viés de Proxy (Proxy Bias):** Ocorre quando o modelo utiliza características aparentemente neutras (não sensíveis) que são, na verdade, altamente correlacionadas com atributos sensíveis protegidos (como raça, gênero, etc.).1 Por exemplo, usar o CEP como uma característica em um modelo de concessão de crédito pode indiretamente discriminar com base na raça ou status socioeconômico, pois esses fatores estão frequentemente correlacionados com a localização geográfica.1

É fundamental notar que muitos desses vieses de dados estão interconectados. O viés histórico 2 frequentemente leva ao viés de representação 2, pois grupos historicamente marginalizados tendem a ser sub-representados nos dados. Além disso, esses vieses podem ser codificados através de variáveis proxy.1 Isso implica que a simples remoção de atributos protegidos (como raça ou gênero) dos dados de treinamento muitas vezes não é suficiente para eliminar o viés, pois o modelo ainda pode aprender associações indesejadas através dessas características correlacionadas.6 Isso demonstra a complexidade da mitigação do viés de dados, exigindo abordagens mais sofisticadas do que a mera exclusão de atributos.

### **B. Vieses Algorítmicos e Induzidos pelo Modelo**

Embora os dados sejam uma fonte primária, o próprio algoritmo de aprendizado de máquina e o processo de modelagem também podem introduzir ou amplificar vieses:

* **Viés Algorítmico (Algorithmic Bias):** Refere-se ao viés que surge das próprias escolhas feitas no design do modelo, sua arquitetura, função objetivo ou processo de otimização, independentemente de problemas nos dados.4 Por exemplo, um modelo pode fazer suposições simplificadoras (como linearidade) que não se sustentam no mundo real, levando a erros sistemáticos.48 A própria escolha de usar um algoritmo para uma determinada tarefa pode incorporar um viés.4
* **Dilema Viés-Variância (Bias-Variance Tradeoff):** Tradicionalmente, existe um trade-off na otimização de modelos: reduzir a variância (sensibilidade do modelo a flutuações nos dados de treinamento) pode aumentar o viés (erro sistemático do modelo) e vice-versa.2 A busca por minimizar o erro total geralmente envolve encontrar um equilíbrio entre esses dois. Embora modelos muito grandes (como redes neurais profundas) às vezes pareçam desafiar esse trade-off tradicional, a relação entre complexidade do modelo, viés e variância ainda é uma consideração importante no treinamento.2
* **Viés de Amplificação (Amplification Bias):** Os modelos de ML podem não apenas refletir os vieses presentes nos dados de treinamento, mas também amplificá-los.5 Pequenas representações desiguais ou correlações espúrias nos dados podem ser exageradas pelo modelo durante o aprendizado, levando a disparidades ainda maiores nas previsões.
* **Viés de Automação (Automation Bias):** Refere-se à tendência humana de confiar excessivamente ou acreditar acriticamente nas saídas de sistemas automatizados.13 Embora não seja um viés do modelo em si, ele interage com o viés algorítmico, pois pode levar à aceitação e implantação de modelos enviesados sem o escrutínio adequado, simplesmente porque se acredita que um sistema automatizado é inerentemente melhor ou mais objetivo do que a decisão humana.13

Esses pontos destacam que o algoritmo e o modelo não são entidades neutras.5 As escolhas inerentes ao design do modelo, aos objetivos de otimização (como o equilíbrio viés-variância 2) e à própria decisão de automatizar uma tarefa 13 podem introduzir ou exacerbar o viés, mesmo na presença de dados hipoteticamente "perfeitos". Isso sublinha a importância de considerar o processo de modelagem como um local ativo de introdução de viés, não apenas como um refletor passivo dos vieses dos dados.

### **C. Vieses Induzidos por Humanos no Pipeline de Treinamento**

As ações e predisposições dos humanos envolvidos no desenvolvimento de sistemas de IA também são uma fonte significativa de viés:

* **Viés de Preconceito (Prejudice Bias):** Noções preconcebidas, estereótipos ou crenças tendenciosas mantidas pelos desenvolvedores, cientistas de dados ou anotadores de dados podem influenciar decisões em várias etapas, como coleta de dados, engenharia de características, rotulagem ou avaliação do modelo.4 Essas crenças podem ser conscientes ou inconscientes.
* **Viés de Confirmação (Confirmation Bias):** A tendência de buscar, interpretar ou favorecer informações que confirmem crenças ou hipóteses preexistentes.7 No desenvolvimento de ML, isso pode se manifestar quando os construtores de modelos processam dados inconscientemente ou selecionam características de maneiras que apoiam suas suposições iniciais, como descartar dados que contradizem uma crença sobre a agressividade de uma determinada raça de cachorro.13
* **Viés do Experimentador (Experimenter's Bias):** Ocorre quando um desenvolvedor continua a treinar, ajustar ou modificar um modelo repetidamente até que ele produza um resultado que se alinhe com sua hipótese ou expectativa original, em vez de aceitar os resultados objetivamente.13
* **Viés Implícito (Implicit Bias):** Vieses inconscientes, muitas vezes moldados por normas sociais e culturais, que afetam o julgamento e as decisões sem que o indivíduo perceba.7 Esses vieses podem influenciar sutilmente desde a formulação do problema até a interpretação dos resultados.7
* **Viés de Atribuição de Grupo (Group Attribution Bias):** A tendência de generalizar características de um grupo para um indivíduo (estereotipagem) ou de um indivíduo para um grupo.6 Isso inclui o *viés endogrupal* (in-group bias), que é a preferência pelo próprio grupo, e o *viés de homogeneidade exogrupal* (out-group homogeneity bias), que é a percepção de que os membros de outros grupos são mais semelhantes entre si do que realmente são.7 Esse último pode levar, por exemplo, a sistemas de reconhecimento facial que têm dificuldade em distinguir indivíduos de grupos minoritários devido à falta de diversidade nos dados de treinamento.7

A presença desses vieses humanos ressalta que o pipeline de ML não é um processo puramente técnico e objetivo.5 Decisões humanas, conscientes ou inconscientes, permeiam todas as etapas, desde a definição do problema e coleta de dados 7 até a construção do modelo 13 e avaliação.4 Isso enfatiza a necessidade crítica de equipes de desenvolvimento diversas, práticas de documentação rigorosas e uma cultura de auto-reflexão crítica para identificar e mitigar o impacto dos vieses humanos na criação de sistemas de IA.21

### **D. Estruturas Conceituais (por exemplo, Viés Tipo I vs. Tipo II)**

A ambiguidade em torno do termo "viés" levou a esforços de pesquisa para desenvolver estruturas conceituais mais precisas para categorizar diferentes tipos de problemas de justiça em IA. Uma distinção útil proposta na literatura diferencia entre Viés Tipo I e Viés Tipo II 16:

* **Viés Tipo I (Desempenho Desigual):** Este tipo de viés refere-se a disparidades no desempenho do modelo entre diferentes grupos demográficos.16 O modelo funciona sistematicamente melhor para um grupo do que para outro em métricas de desempenho padrão. Um exemplo proeminente é o desempenho inferior de sistemas de reconhecimento facial para mulheres de pele mais escura em comparação com homens de pele clara.10 A avaliação do Viés Tipo I geralmente requer acesso aos rótulos verdadeiros (ground truth) da tarefa principal para calcular métricas como taxa de erro, acurácia, taxa de verdadeiros positivos (TPR), taxa de falsos positivos (FPR) ou área sob a curva ROC (AUC) e comparar essas métricas entre os grupos.16
* **Viés Tipo II (Dependência Indesejada):** Este viés ocorre quando as previsões do modelo (Y^) mostram uma dependência estatística indesejada de um atributo protegido (A), mesmo que o desempenho geral do modelo (por exemplo, acurácia) seja alto.16 O problema aqui não é necessariamente que o modelo erra mais para um grupo, mas que suas previsões são influenciadas pelo atributo protegido de uma forma considerada injusta. Um exemplo seria um sistema de aprovação de empréstimos que tende a aprovar empréstimos com mais frequência para um grupo demográfico do que para outro, mesmo controlando para fatores de risco relevantes.16 A avaliação do Viés Tipo II foca na relação entre a previsão (Y^) e o atributo protegido (A) e nem sempre requer os rótulos verdadeiros da tarefa principal para sua avaliação.16

Essa distinção é valiosa porque clarifica que "viés" não é um conceito monolítico. Falhas de desempenho (Tipo I) e influência indevida de atributos protegidos (Tipo II) representam modos de falha distintos. Um modelo pode ter alta acurácia geral, mas ainda assim exibir Viés Tipo I (desempenho desigual entre grupos) ou Viés Tipo II (previsões correlacionadas com atributos protegidos). Abordar esses diferentes tipos de viés pode exigir estratégias de mitigação e métricas de avaliação distintas.16 Simplesmente otimizar a acurácia geral do modelo não garante a justiça em nenhum desses sentidos.

## **III. Principais Benchmarks de Viés: Conjuntos de Dados e Metodologias**

Para operacionalizar a avaliação dos vieses descritos anteriormente, a comunidade de pesquisa desenvolveu vários benchmarks. Esses benchmarks fornecem conjuntos de dados e metodologias padronizadas para testar sistemas de IA em relação a vieses específicos.8 Eles cobrem uma variedade de domínios de aplicação, incluindo análise facial, justiça criminal, classificação de texto e modelos de linguagem, e geralmente se concentram em tipos específicos de viés social, como gênero, raça, tipo de pele, religião ou idade.8

### **A. Visão Geral dos Principais Benchmarks**

Alguns dos benchmarks de viés mais proeminentes incluem:

* **Gender Shades:** Focado em viés interseccional (gênero e tipo de pele) em sistemas comerciais de classificação de gênero facial.10
* **COMPAS:** Analisa o viés racial na ferramenta de avaliação de risco de reincidência COMPAS.8 Embora seu uso como benchmark seja desencorajado devido a problemas inerentes.50
* **Civil Comments (Jigsaw Unintended Bias):** Avalia o viés não intencional em modelos de classificação de toxicidade de texto, especificamente a associação incorreta de termos de identidade com toxicidade.22
* **BOLD (Bias in Open-Ended Language Generation):** Mede vieses sociais (profissão, gênero, raça, religião, ideologia política) em texto gerado por modelos de linguagem de forma aberta.36
* **StereoSet:** Quantifica vieses estereotipados (gênero, profissão, raça, religião) em modelos de linguagem pré-treinados usando tarefas intra e intersentenciais.23
* **CrowS-Pairs:** Mede vieses estereotipados em modelos de linguagem mascarados (MLMs) em nove categorias de viés social nos EUA, usando pares de sentenças minimamente diferentes.19

### **B. Perfis Detalhados de Benchmarks**

A seguir, detalhamos a metodologia, métricas e achados de cada um desses benchmarks-chave.

**1. Gender Shades**

* **Objetivo:** Avaliar a precisão e o viés em produtos comerciais de classificação de gênero baseados em IA, com foco particular na interseção entre gênero e tipo de pele.10
* **Dataset (PPB):** O benchmark utiliza o Pilot Parliaments Benchmark (PPB), um conjunto de dados composto por 1270 imagens de parlamentares de três países africanos e três europeus.11 Este dataset foi construído para ser equilibrado em termos de gênero percebido (feminino/masculino) e tipo de pele.10 O tipo de pele foi rotulado usando a escala Fitzpatrick, um sistema aprovado por dermatologistas, e posteriormente agrupado em categorias "mais clara" e "mais escura".24 A rotulagem de gênero foi binária (feminino/masculino), refletindo as saídas dos sistemas avaliados, mas reconhecendo que isso é uma simplificação da complexidade do gênero.24
* **Metodologia:** A metodologia consistiu em testar as APIs de classificação de gênero de três grandes empresas (IBM, Microsoft, Face++) usando o dataset PPB.11 A análise principal envolveu o cálculo e a comparação das taxas de erro de classificação entre diferentes subgrupos, especialmente os quatro grupos interseccionais: mulheres de pele escura, homens de pele escura, mulheres de pele clara e homens de pele clara.10
* **Tipo de Viés Medido:** O benchmark mede principalmente o viés interseccional de gênero e tipo de pele, especificamente o Viés Tipo I, que se manifesta como disparidades significativas nas taxas de erro entre os subgrupos.10
* **Métricas Principais:** A métrica central é a Taxa de Erro de Classificação de Gênero.24 A análise foca nas diferenças nessas taxas entre os grupos, como a diferença absoluta entre a taxa de erro para homens de pele clara e mulheres de pele escura.24
* **Principais Achados:** O estudo revelou disparidades substanciais no desempenho dos classificadores. Mulheres de pele escura consistentemente apresentaram as maiores taxas de erro (chegando a 34.7% em um dos sistemas), enquanto homens de pele clara tiveram as menores taxas de erro (tão baixas quanto 0.0% e 0.3% em dois sistemas).10 Esses achados apontaram para a sub-representação de certos grupos nos dados de treinamento como uma causa provável do viés.10
* **Limitações e Críticas:** Os próprios autores reconhecem limitações, como a classificação binária de gênero que não abrange identidades trans ou não-binárias.24 A escala Fitzpatrick é um sistema de classificação e um proxy para a cor da pele.24 O estudo se concentrou em APIs específicas disponíveis em um determinado momento.24 O trabalho enfatizou a necessidade urgente de maior transparência no desempenho de sistemas de IA, testes inclusivos e relatórios desagregados por subgrupos interseccionais.10

**2. COMPAS**

* **Objetivo:** O dataset COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) foi originalmente analisado pela ProPublica para investigar o viés racial na ferramenta de avaliação de risco de reincidência criminal usada no sistema de justiça dos EUA.8 Subsequentemente, tornou-se um dataset frequentemente utilizado (embora agora desencorajado) para testar e comparar algoritmos de mitigação de viés.48
* **Dataset:** Contém dados de mais de 10.000 réus criminais do Condado de Broward, Flórida, coletados entre 2013-2014.8 Inclui informações demográficas (idade, raça, sexo), histórico criminal (contagem de antecedentes, grau da acusação), a pontuação de risco decil atribuída pelo COMPAS e o resultado real de reincidência (geral e violenta) dentro de um período de dois anos.8
* **Metodologia:** A análise original da ProPublica comparou as previsões de risco do COMPAS com os resultados reais de reincidência, focando nas taxas de erro (especialmente falsos positivos e falsos negativos) entre réus negros e brancos.8 Pesquisadores frequentemente usam este dataset para treinar seus próprios modelos preditivos de reincidência (às vezes excluindo a raça como característica explícita) e, em seguida, avaliam a justiça desses modelos usando várias métricas de equidade.8
* **Tipo de Viés Medido:** Principalmente viés racial, que pode se manifestar como Viés Tipo I (disparidades nas taxas de erro, como taxas de falsos positivos diferentes para grupos raciais) e potencialmente Viés Tipo II (dependência da pontuação de risco da raça, mesmo após controlar outros fatores relevantes).8
* **Métricas Principais:** Taxa de Falsos Positivos (FPR), Taxa de Falsos Negativos (FNR), Acurácia, Precisão (Taxa de Verdadeiros Positivos - TPR ou Sensibilidade).8 Ao avaliar modelos treinados nesses dados, métricas como paridade demográfica, oportunidade igualitária e probabilidades equalizadas são comumente calculadas e comparadas entre grupos raciais.
* **Principais Achados (ProPublica):** A análise da ProPublica encontrou que réus negros que não reincidiram tinham quase o dobro da probabilidade de serem classificados erroneamente como de alto risco (FPR mais alto) em comparação com réus brancos.8 Por outro lado, réus brancos que reincidiram tinham quase o dobro da probabilidade de serem rotulados erroneamente como de baixo risco (FNR mais alto) em comparação com réus negros que reincidiram.8 Essas disparidades persistiram mesmo ao controlar por histórico criminal, idade e gênero.8
* **Limitações e Críticas:** O algoritmo COMPAS é proprietário e funciona como uma "caixa preta".48 A metodologia da ProPublica foi criticada por suas suposições e métricas.48 Mais fundamentalmente, o próprio dataset COMPAS é considerado problemático para benchmarking de justiça. Ele sofre de viés de medição (prisão não é igual a crime; a definição de reincidência pode ser enviesada) e a tarefa de predição derivada dele pode ter pouca conexão com tarefas relevantes no mundo real para instrumentos de avaliação de risco.50 Por essas razões, seu uso para benchmark de *novos* algoritmos de justiça é agora geralmente desencorajado pela comunidade de pesquisa.50

**3. Civil Comments (Jigsaw Unintended Bias)**

* **Objetivo:** Medir o viés não intencional em modelos de classificação de toxicidade, especificamente a tendência dos modelos de associarem erroneamente comentários não tóxicos que mencionam certas identidades (por exemplo, raça, gênero, orientação sexual) como sendo tóxicos.22
* **Dataset:** Consiste em aproximadamente 2 milhões de comentários públicos da plataforma Civil Comments (ativa de 2015 a 2017).22 Os comentários foram anotados por múltiplos avaliadores humanos (até 10 ou mais por comentário) para determinar um nível de toxicidade (rótulo alvo, representado como uma fração; >= 0.5 é considerado tóxico para avaliação).22 Anotações adicionais incluem subtipos de toxicidade (toxicidade severa, obsceno, ameaça, insulto, ataque à identidade, sexualmente explícito) e, para um subconjunto de comentários, menções a vários grupos de identidade (gênero, orientação sexual, religião, raça/etnia, deficiência, etc.).22
* **Metodologia:** O benchmark envolve treinar modelos de classificação de toxicidade no conjunto de dados e avaliá-los não apenas pela performance geral (AUC), mas também usando métricas de viés especializadas.22 Essas métricas comparam o desempenho do modelo em subconjuntos de dados que mencionam identidades específicas versus o desempenho em dados de fundo (comentários que não mencionam essa identidade).57
* **Tipo de Viés Medido:** Viés de identidade ou viés de correlação, onde o modelo aprende associações espúrias entre termos de identidade e toxicidade.22 Isso pode ser visto como uma manifestação de Viés Tipo II (dependência indesejada da previsão em menções de identidade) que leva a danos do Tipo I (por exemplo, taxas de falsos positivos mais altas para comentários não tóxicos mencionando certos grupos).57
* **Métricas Principais:** AUC ROC geral. AUCs de Viés: AUC de Subgrupo (AUC nos comentários mencionando a identidade), AUC BPSN (Background Positive, Subgroup Negative - mede a confusão entre comentários tóxicos sem a identidade e comentários não tóxicos com a identidade), AUC BNSP (Background Negative, Subgroup Positive - mede a confusão entre comentários não tóxicos sem a identidade e comentários tóxicos com a identidade).22 Uma Média Generalizada das AUCs de Viés (frequentemente com p=−5) é usada para combinar as métricas por identidade em uma pontuação geral de viés, dando mais peso aos subgrupos com pior desempenho.57
* **Principais Achados:** Modelos de toxicidade frequentemente exibem viés, como taxas de falsos positivos mais altas para comentários não tóxicos que mencionam identidades como 'gay' ou 'negro'.57 Diferentes arquiteturas de modelo (BERT, RNN) mostram diferentes níveis e tipos de viés.66 A própria subjetividade e os vieses potenciais dos anotadores humanos também são uma consideração.61
* **Limitações e Críticas:** A definição e anotação de toxicidade são inerentemente subjetivas e dependem da agregação de julgamentos de múltiplos avaliadores.57 A qualidade dos dados reflete a natureza bruta dos comentários online.69 O benchmark foca em *menções* de identidades, o que não é o mesmo que avaliar o viés em relação a comentários escritos *por* membros desses grupos.

**4. BOLD (Bias in Open-Ended Language Generation)**

* **Objetivo:** Avaliar e comparar sistematicamente vieses sociais em modelos de linguagem que geram texto de forma aberta (open-ended), abrangendo múltiplos domínios sociais.18
* **Dataset:** Consiste em 23.679 prompts em inglês, extraídos de frases da Wikipedia para simular gatilhos de linguagem natural.36 Os prompts cobrem cinco domínios: profissão, gênero, raça, religião e ideologia política, abrangendo 43 subgrupos diferentes.36
* **Metodologia:** Usar os prompts BOLD para estimular modelos de linguagem (LLMs) a gerar continuações de texto.36 Analisar os textos gerados usando um conjunto de métricas automatizadas que medem diferentes dimensões de viés social: sentimento, toxicidade, consideração (regard), normas psicolinguísticas e polaridade de gênero.36 A análise principal envolve comparar as distribuições dessas métricas entre diferentes subgrupos demográficos dentro de cada domínio (por exemplo, comparar o sentimento médio do texto gerado para diferentes grupos religiosos).36 As métricas automatizadas foram validadas comparando-as com julgamentos humanos coletados via Amazon Mechanical Turk (AMT).36
* **Tipo de Viés Medido:** Mede a manifestação de vários vieses sociais (estereotipagem, sentimento negativo, toxicidade, disparidades de consideração) no texto gerado associado a diferentes grupos demográficos.36 Isso se alinha com a avaliação de manifestações do Viés Tipo II.
* **Métricas Principais:** Sentimento (usando VADER), Toxicidade (usando um classificador BERT), Consideração (Regard, usando um classificador BERT treinado), Normas Psicolinguísticas (Valence, Arousal, Dominance - VAD; e Joy, Anger, Sadness, Fear, Disgust - BE5, usando redes neurais treinadas em léxicos), Polaridade de Gênero (contagem de unigramas específicos de gênero e projeção de embeddings de palavras em um eixo de gênero).36
* **Principais Achados:** Modelos populares avaliados (GPT-2, BERT, CTRL) geralmente exibiram vieses sociais mais pronunciados do que o texto base da Wikipedia de onde os prompts foram extraídos, especialmente em relação a grupos historicamente desfavorecidos.18 Exemplos incluem sentimento mais negativo em relação ao Ateísmo e Islamismo (GPT-2) e associações de gênero estereotipadas com profissões.71
* **Limitações e Críticas:** Uma limitação significativa é a dependência de classificadores automatizados (para toxicidade, sentimento, consideração) como modelos de métrica.33 Esses próprios classificadores podem ser enviesados, introduzindo assim viés na própria avaliação do viés.33 A escolha das métricas específicas influencia fortemente o tipo e a magnitude do viés detectado. A extração de prompts da Wikipedia pode incorporar os vieses inerentes à própria Wikipedia. O benchmark é focado no idioma inglês.

**5. StereoSet**

* **Objetivo:** Medir vieses estereotipados em modelos de linguagem pré-treinados (PLMs) em quatro domínios sociais: gênero, profissão, raça e religião.20
* **Dataset:** Um conjunto de dados em inglês de larga escala (aproximadamente 17.000 frases) 72, com exemplos criados por crowdsourcing.23 Cada exemplo geralmente consiste em um contexto e três opções de associação ou continuação: uma estereotipada, uma anti-estereotipada e uma não relacionada ou sem sentido.72
* **Metodologia:** O benchmark inclui duas tarefas principais 55:
  + *Tarefa Intrasentencial:* Avalia o viés dentro de uma única frase, geralmente envolvendo a previsão de uma palavra mascarada em um contexto que pode ativar estereótipos.55
  + *Tarefa Intersentencial:* Avalia o viés entre frases, apresentando ao modelo uma frase de contexto e pedindo que escolha a continuação mais provável ou coerente entre as três opções (estereotipada, anti-estereotipada, não relacionada).55 A avaliação se baseia na preferência do modelo (medida por probabilidade ou verossimilhança) por cada tipo de opção.30
* **Tipo de Viés Medido:** Viés de associação estereotipada (Viés Tipo II).23
* **Métricas Principais:** Language Modeling Score (LMS), Stereotype Score (SS) e Idealized Context Association Test (iCAT) Score.23
  + *LMS:* Mede a capacidade do modelo de distinguir associações significativas (estereótipo ou anti-estereótipo) das não relacionadas. É a porcentagem de casos em que o modelo prefere uma opção significativa.72 Uma pontuação alta indica boa compreensão contextual básica.
  + *SS:* Mede a preferência do modelo pelo estereótipo em relação ao anti-estereótipo, entre os casos em que escolheu uma opção significativa. É a porcentagem de instâncias significativas onde o estereótipo foi preferido.72 Uma pontuação idealmente imparcial é 50%. Pontuações distantes de 50% indicam viés.
  + *iCAT:* Combina LMS e SS para fornecer uma única pontuação que recompensa a compreensão (alto LMS) e a imparcialidade (SS próximo de 50%). A fórmula é: iCAT=lms×50min(ss,100−ss)​.72 Uma pontuação de 100 é ideal (alto LMS, SS=50), enquanto 0 sugere viés máximo ou baixa compreensão.
* **Principais Achados:** Modelos de linguagem populares como BERT, GPT-2, RoBERTa e XLNet demonstraram fortes vieses estereotipados, com pontuações SS significativamente diferentes de 50%.23 O desempenho varia entre modelos e domínios de viés.23
* **Limitações e Críticas:** A qualidade dos dados do StereoSet foi questionada, com estudos indicando que uma porcentagem significativa das amostras pode ter problemas de validade ou não representar estereótipos genuínos.28 A metodologia de cálculo de pontuação e as suposições subjacentes também foram criticadas.28 Foram observadas inconsistências nos resultados quando comparados com outros benchmarks como o CrowS-Pairs.38 O foco em estruturas de frases específicas pode limitar a generalização.

**6. CrowS-Pairs**

* **Objetivo:** Medir vieses sociais, especificamente estereótipos, em Modelos de Linguagem Mascarados (MLMs) contra grupos demográficos protegidos nos EUA.19 O benchmark cobre nove categorias de viés: raça/cor, gênero/identidade de gênero, orientação sexual, religião, idade, nacionalidade, deficiência, aparência física e status socioeconômico.19
* **Dataset:** Contém 1.508 pares de frases em inglês, criados por crowdsourcing.19 Cada par consiste em duas frases minimamente diferentes: uma (sent\_more) é mais estereotipada em relação a um grupo desfavorecido, e a outra (sent\_less) é menos estereotipada, frequentemente envolvendo um grupo favorecido contrastante.19 O dataset inclui rótulos indicando a direção do estereótipo (se sent\_more expressa um estereótipo ou sent\_less viola um) e o tipo de viés.27 Existe também uma extensão para o francês.80
* **Metodologia:** A metodologia envolve apresentar os pares de frases a um MLM e calcular uma pontuação para cada frase.25 A pontuação é baseada na pseudo-log-verossimilhança (PLL) dos tokens não modificados da frase, condicionada aos tokens modificados (que definem o grupo).25 Isso é feito mascarando iterativamente cada token não modificado e calculando sua probabilidade logarítmica dada pelo MLM.25 As pontuações das duas frases do par são então comparadas para determinar qual delas o modelo "prefere" (atribui maior verossimilhança).25
* **Tipo de Viés Medido:** Viés de associação estereotipada (Viés Tipo II) em múltiplas dimensões sociais.19
* **Métricas Principais:** A métrica primária é a porcentagem de pares em que o modelo atribui uma pontuação de verossimilhança maior à frase mais estereotipada (sent\_more).25 Uma pontuação idealmente imparcial seria 50%. Pontuações significativamente acima de 50% indicam um viés do modelo em favor dos estereótipos.25 Análises de confiança baseadas na razão das pontuações também podem ser realizadas.25
* **Principais Achados:** Avaliações iniciais em MLMs amplamente utilizados (BERT, RoBERTa, ALBERT) mostraram uma preferência substancial por sentenças estereotipadas em todas as categorias de viés.19 O nível de viés varia entre diferentes modelos e tipos de viés.86
* **Limitações e Críticas:** O benchmark CrowS-Pairs enfrentou críticas significativas quanto à confiabilidade e validade de seus dados.27 Blodgett et al. (2021) e outros pesquisadores identificaram problemas substanciais, sugerindo que apenas uma pequena fração das amostras pode ser válida ou representar adequadamente estereótipos prejudiciais.28 A metodologia de pontuação baseada em PLL também foi questionada.28 O foco do dataset é centrado nos EUA.25 Além disso, foram relatadas inconsistências nos resultados em comparação com o StereoSet.38 Devido a essas questões, a adequação do CrowS-Pairs como um indicador confiável de viés social em LMs é agora questionada por parte da comunidade.27

As críticas substanciais dirigidas a benchmarks amplamente utilizados como StereoSet e, especialmente, CrowS-Pairs 27 apontam para um desafio crítico no campo da avaliação da justiça em IA: a dificuldade em construir benchmarks que sejam simultaneamente abrangentes em termos de tipos de viés e domínios, e psicometricamente sólidos e válidos. Os problemas identificados não são apenas falhas menores, mas potencialmente questões fundamentais com a coleta de dados, construção de exemplos e metodologias de pontuação.28 Isso sugere uma necessidade premente de metodologias de desenvolvimento de benchmark mais rigorosas, maior transparência na documentação e, potencialmente, uma mudança para abordagens de avaliação mais dinâmicas, contextuais ou adversariais que possam superar as limitações dos atuais benchmarks estáticos.83 A confiabilidade dos benchmarks é fundamental, pois eles informam tanto a pesquisa quanto o desenvolvimento de técnicas de mitigação.

## **IV. Utilizando Benchmarks na Fase de Treinamento de IA**

Embora os benchmarks sejam frequentemente associados à avaliação final de modelos treinados, seu valor se estende significativamente à fase de treinamento. As informações obtidas a partir da avaliação de benchmarks podem informar diretamente as estratégias de pré-processamento de dados e guiar o desenvolvimento e a aplicação de técnicas de mitigação de viés *in-processing*.

### **A. Conectando Insights de Benchmarks ao Pré-processamento e Aumento de Dados**

Os benchmarks atuam como ferramentas diagnósticas que podem revelar deficiências nos dados de treinamento. Por exemplo, o benchmark Gender Shades expôs a sub-representação de mulheres de pele escura em conjuntos de dados de análise facial e o consequente desempenho inferior dos modelos para esse grupo.10 Esse tipo de descoberta motiva diretamente a coleta de dados mais diversos e representativos ou o uso de técnicas de aumento de dados direcionadas para grupos sub-representados.5

Além disso, as disparidades de grupo identificadas por benchmarks podem guiar a aplicação de técnicas de pré-processamento como reponderação (reweighting) ou reamostragem (resampling).46 A reponderação atribui pesos diferentes às amostras de treinamento, dando maior importância a exemplos de grupos minoritários ou sub-representados, conforme identificado pela análise de benchmark.8 A reamostragem envolve a subamostragem de grupos majoritários (downsampling) ou a superamostragem de grupos minoritários (upsampling), possivelmente através da duplicação de exemplos ou da geração de dados sintéticos.46

Outra abordagem de pré-processamento informada por benchmarks é o aprendizado de representações justas (fair representation learning).5 O objetivo é transformar os dados de entrada em um novo espaço latente onde a informação sobre atributos sensíveis é removida ou obscurecida, enquanto a informação útil para a tarefa principal é preservada.46 Benchmarks podem ser usados para avaliar o sucesso dessas transformações em reduzir o viés.

Finalmente, a estrutura de alguns benchmarks, como os pares contrafatuais usados no CrowS-Pairs 25, pode inspirar técnicas de aumento de dados contrafatuais. Isso envolve criar novas amostras de treinamento alterando atributos sensíveis (por exemplo, trocando pronomes de gênero) para ajudar o modelo a aprender a ser menos dependente desses atributos.90

Essas conexões demonstram um ciclo de feedback valioso: os benchmarks não servem apenas para avaliação pós-treinamento, mas fornecem diagnósticos cruciais que informam diretamente como os dados de treinamento devem ser curados, aumentados ou reponderados na fase de pré-processamento. Essa abordagem proativa visa mitigar os vieses identificados pelo benchmark antes mesmo que o treinamento do modelo comece.8

### **B. Técnicas de Mitigação de Viés In-Processing Guiadas por Benchmarks**

As técnicas de mitigação *in-processing* intervêm diretamente no processo de aprendizado do modelo, modificando o algoritmo de treinamento ou a função objetivo para reduzir o viés.16 Os benchmarks desempenham um papel crucial ao fornecer as métricas de justiça que essas técnicas visam otimizar ou as restrições que buscam satisfazer.

**1. Regularização e Restrições Cientes da Justiça (Fairness-Aware Regularization and Constraints)**

Uma abordagem comum é incorporar a justiça na função de perda que o modelo otimiza durante o treinamento. Isso pode ser feito adicionando termos de regularização que penalizam resultados enviesados ou dependências indesejadas de atributos sensíveis.40 Por exemplo, um termo de regularização pode penalizar a diferença nas taxas de erro entre diferentes grupos demográficos.12 A força dessa penalidade (controlada por um hiperparâmetro) permite ajustar o trade-off entre precisão e justiça.91

Alternativamente, restrições de justiça podem ser aplicadas diretamente durante o processo de otimização.14 Essas restrições exigem que o modelo satisfaça certas condições de paridade, como paridade demográfica (taxas de previsão positiva iguais entre grupos) ou probabilidades equalizadas (taxas de verdadeiros positivos e falsos positivos iguais entre grupos).14 Frameworks como o Fairlearn 14 e o AI Fairness 360 5 implementam algoritmos (por exemplo, ExponentiatedGradient, GridSearch 14) que podem treinar modelos sujeitos a essas restrições. A escolha da métrica ou restrição de justiça apropriada depende crucialmente da definição de justiça relevante para o contexto da aplicação e do tipo de viés identificado através da análise de benchmarks.14

**2. Debiasing Adversarial (Adversarial Debiasing)**

Esta técnica emprega uma abordagem inspirada em redes adversariais generativas (GANs) para promover a justiça.12 Ela envolve o treinamento simultâneo de dois modelos: um modelo principal (o preditor) que realiza a tarefa desejada (por exemplo, classificação) e um modelo adversário.44

O preditor é treinado para minimizar sua própria perda de predição (melhorar a precisão) e, ao mesmo tempo, *maximizar* a perda do adversário.44 O adversário, por sua vez, é treinado para prever o atributo sensível protegido (como gênero ou raça) a partir das previsões (Y^) ou das representações internas do preditor.44

Esse jogo adversarial força o preditor a aprender a fazer previsões ou a gerar representações que contenham o mínimo possível de informação sobre o atributo sensível, tornando difícil para o adversário adivinhá-lo.44 Isso resulta em um classificador mais justo, pois suas previsões não carregam informações discriminatórias que o adversário possa explorar.93 Dependendo de como o adversário é configurado (por exemplo, quais informações ele recebe como entrada), a técnica pode ser usada para impor diferentes noções de justiça, como paridade demográfica ou igualdade de probabilidades.44 Essa abordagem é implementada em toolkits como o AI Fairness 360.93

**3. Aprendizado de Representação Justa (In-Processing)**

Semelhante à abordagem de pré-processamento, o aprendizado de representação justa pode ocorrer durante o treinamento do modelo principal.5 O objetivo é aprender um espaço latente (representação interna dos dados) que seja útil para a tarefa principal, mas que seja estatisticamente independente ou contenha informações mínimas sobre os atributos sensíveis. Técnicas como regularização ou aprendizado adversarial são frequentemente usadas para impor essa restrição de justiça ao espaço de representação durante o treinamento.17

**4. Justiça em Fine-Tuning Eficiente em Parâmetros (PEFT)**

Com o advento de grandes modelos de linguagem (LLMs), as técnicas de PEFT, como LoRA (Low-Rank Adaptation), tornaram-se populares para adaptar esses modelos a tarefas específicas com custo computacional reduzido.43 A pesquisa está explorando como garantir ou melhorar a justiça durante o PEFT.43 Isso pode envolver a aplicação de restrições de justiça ou funções objetivo modificadas especificamente durante a fase de fine-tuning de baixo posto (por exemplo, FairLoRA) ou a busca heurística por conjuntos de parâmetros justos para atualizar.43 A relação entre PEFT e justiça ainda está sob investigação, com resultados indicando que o impacto na justiça pode depender do modelo pré-treinado, do conjunto de dados e da tarefa.43

O mecanismo fundamental subjacente a todas essas técnicas *in-processing* é a modificação da paisagem de otimização do modelo.14 Elas introduzem explicitamente considerações de justiça (conforme definidas e quantificadas por métricas ou restrições derivadas de benchmarks) na função objetivo ou no algoritmo de treinamento. Isso força o modelo a equilibrar a precisão com a justiça durante o aprendizado, resultando em modelos que são projetados para serem inerentemente mais justos, em vez de terem seus inputs ou outputs ajustados retrospectivamente.

### **C. Utilizando Métricas de Benchmark para Otimização e Validação do Treinamento**

As métricas quantitativas fornecidas pelos benchmarks ou inspiradas por eles são essenciais para operacionalizar as técnicas de mitigação *in-processing*. Elas servem a múltiplos propósitos durante o ciclo de treinamento:

* **Otimização Direta:** Algumas abordagens *in-processing* podem tentar otimizar diretamente uma métrica de justiça (ou minimizar uma métrica de disparidade) como parte da função objetivo, geralmente em combinação com um termo de perda de precisão.41
* **Aplicação de Restrições:** Métricas de justiça definem as restrições de paridade (por exemplo, diferença de TPR entre grupos para Oportunidade Igualitária) que o modelo deve satisfazer durante o treinamento, conforme imposto por algoritmos específicos.14
* **Validação Cruzada e Ajuste de Hiperparâmetros:** Assim como a precisão é avaliada em um conjunto de validação para ajustar hiperparâmetros, as métricas de justiça (calculadas usando metodologias de benchmark em dados de validação) podem ser usadas para ajustar hiperparâmetros relacionados à justiça.21 Isso inclui, por exemplo, a força de um termo de regularização de justiça ou o peso da perda adversarial.2 O objetivo é encontrar um equilíbrio aceitável entre o desempenho da tarefa e a justiça no conjunto de validação.
* **Monitoramento Durante o Treinamento:** Rastrear métricas de justiça juntamente com a precisão em dados semelhantes aos de benchmark ao longo das épocas de treinamento pode fornecer informações valiosas.21 Isso pode ajudar a identificar se o modelo está aprendendo ou amplificando vieses durante o treinamento e se as técnicas de mitigação estão tendo o efeito desejado.

Em essência, os benchmarks e as métricas associadas fornecem as *definições quantificáveis* de justiça que são necessárias para traduzir conceitos abstratos de equidade em objetivos matemáticos concretos que podem ser incorporados aos algoritmos de otimização.14 Sem essas métricas, as metas de justiça permaneceriam vagas e não poderiam ser diretamente integradas ao processo de treinamento *in-processing*. Elas formam a ponte entre os ideais de justiça e a implementação matemática na otimização de modelos de ML.

## **V. Análise Comparativa de Desempenho: Modelos vs. Benchmarks**

A avaliação de modelos de IA em benchmarks de viés permite uma análise comparativa de seu desempenho em relação à justiça. Vários modelos proeminentes, desde sistemas de visão computacional até grandes modelos de linguagem, foram submetidos a esses testes.

### **A. Pesquisa de Modelos de IA Avaliados por Viés**

Uma ampla gama de modelos de IA foi avaliada usando benchmarks de viés:

* **Análise/Reconhecimento Facial:** Sistemas comerciais da IBM, Microsoft e Face++ foram avaliados no benchmark Gender Shades, revelando vieses significativos.11 Outros estudos também confirmaram vieses em tecnologias de reconhecimento facial (FRT).12
* **Predição de Reincidência:** O algoritmo proprietário COMPAS foi extensivamente analisado quanto ao viés racial.8 Modelos preditivos treinados com dados COMPAS são frequentemente usados em pesquisas sobre justiça algorítmica.48
* **Classificação de Toxicidade:** Diversos modelos, incluindo Redes Neurais Convolucionais (CNNs), Redes Neurais Recorrentes (RNNs, como BiLSTM) e variantes do BERT (RoBERTa, DistilBERT, ALBERT), foram avaliados no dataset Civil Comments (Jigsaw), mostrando diferentes perfis de viés de identidade.57 As versões da API Perspective do Google também foram comparadas.57
* **Modelos de Linguagem (Geral):**
  + *Modelos de Linguagem Pré-treinados (PLMs):* Modelos como BERT, GPT-2, RoBERTa e XLNet foram avaliados nos benchmarks StereoSet 23 e CrowS-Pairs 19, geralmente exibindo vieses estereotipados significativos.
  + *Modelos de Geração Aberta:* GPT-2, BERT e CTRL foram avaliados no benchmark BOLD, mostrando vieses sociais em relação a profissões, gênero, raça, religião e ideologia política.18
  + *Grandes Modelos de Linguagem (LLMs):* Modelos iniciais como GPT-3, InstructGPT, OPT, GLM, Cohere e Anthropic-LM foram avaliados em tarefas como sumarização, considerando aspectos de justiça.96 Mais recentemente, LLMs de última geração, incluindo GPT-4 (e variantes como 4o), Llama (2 e 3), Gemini (1.0, 1.5 Pro), Claude (3 variantes), Mixtral, Qwen, GAL 120B, entre outros, estão sendo cada vez mais avaliados em uma variedade de benchmarks de justiça como BBQ, StereoSet, CrowS-Pairs e BOLD, além de benchmarks de segurança e toxicidade.23

### **B. Tabela de Desempenho Comparativo**

A tabela a seguir sintetiza alguns resultados de desempenho de modelos de IA em benchmarks de viés selecionados, com base nas informações disponíveis nas fontes de pesquisa. É importante notar que a comparação direta pode ser desafiadora devido a variações nas metodologias de avaliação, versões dos modelos e métricas relatadas entre os estudos. A tabela visa fornecer uma visão geral ilustrativa.

| **Modelo** | **Benchmark** | **Domínio/Tipo de Viés** | **Métrica Principal** | **Resultado Exemplo** | **Fonte(s)** |
| --- | --- | --- | --- | --- | --- |
| IBM (API 2017) | Gender Shades | Gênero x Tipo de Pele | Taxa de Erro (Max Gap) | 34.4% (Dif. entre Homens Claros e Mulheres Escuras) | 24 |
| Microsoft (API 2017) | Gender Shades | Gênero x Tipo de Pele | Taxa de Erro (Max Gap) | 20.8% (Dif. entre Homens Claros e Mulheres Escuras) | 24 |
| Face++ (API 2017) | Gender Shades | Gênero x Tipo de Pele | Taxa de Erro (Max Gap) | 33.8% (Dif. entre Homens Escuros e Mulheres Escuras) | 24 |
| Jigsaw TOXICITY@1 | Civil Comments | Identidade (Geral) | Média Generalizada AUC Viés | (Benchmark Baseline) | 57 |
| Jigsaw TOXICITY@6 | Civil Comments | Identidade (Geral) | Média Generalizada AUC Viés | (Benchmark Baseline, melhoria no BPSN vs @1) | 57 |
| GPT-2 (small) | StereoSet | Gênero, Raça, etc. | ICAT Score | 72.97 (Rank 1 em 2020) | 23 |
| XLNet (large) | StereoSet | Gênero, Raça, etc. | ICAT Score | 72.03 (Rank 2 em 2020) | 23 |
| BERT (base) | StereoSet | Gênero, Raça, etc. | ICAT Score | 71.21 (Rank 4 em 2020) | 23 |
| RoBERTa (base) | StereoSet | Gênero, Raça, etc. | ICAT Score | 67.50 (Rank 7 em 2020) | 23 |
| GAL 120B | StereoSet | Gênero, Raça, etc. | ICAT Score (LMS, SS) | 65.6 (75, 56.2) | 23 |
| GPT-3 (text-davinci-002) | StereoSet | Gênero, Raça, etc. | ICAT Score (LMS, SS) | 60.8 (77.6, 60.8) | 23 |
| OPT 175B | StereoSet | Gênero, Raça, etc. | ICAT Score (LMS, SS) | 60 (74.8, 59.9) | 23 |
| GAL 120B | CrowS-Pairs | Gênero, Raça, etc. | % Preferência Estereótipo | 60.5 (Overall, Rank 1 em 2022) | 56 |
| GPT-3 | CrowS-Pairs | Gênero, Raça, etc. | % Preferência Estereótipo | 67.2 (Overall) | 86 |
| OPT-175B | CrowS-Pairs | Gênero, Raça, etc. | % Preferência Estereótipo | 69.5 (Overall) | 86 |
| LLaMA 65B | CrowS-Pairs | Gênero, Raça, etc. | % Preferência Estereótipo | 66.6 (Overall) | 86 |
| GPT-2 | BOLD | Religião | Sentimento Negativo | Mais alto para Ateísmo, Islamismo; mais baixo para Hinduísmo, Budismo | 71 |
| BERT | BOLD | Profissão | Polaridade de Gênero | Mais masculino para Ciência/Tecnologia; mais feminino para Saúde/Medicina | 71 |
| GPT-4o | Vários (MMLU, MATH, etc.) | Conhecimento, Raciocínio | Acurácia | 88.7 (MMLU), 76.6 (MATH) - Comparável ou superior a Claude 3 Opus, Gemini 1.5 Pro | 98 |
| Claude 3 Opus | Vários (MMLU, MATH, etc.) | Conhecimento, Raciocínio | Acurácia | 86.8 (MMLU), 60.1 (MATH) - Forte, mas ligeiramente abaixo de GPT-4o em alguns | 98 |
| Gemini 1.5 Pro | Vários (MMLU, MATH, etc.) | Conhecimento, Raciocínio | Acurácia | 81.9 (MMLU), 58.5 (MATH) - Competitivo, forte em contexto longo | 98 |
| Llama 3 (70B) | Vários (MMLU, MATH, etc.) | Conhecimento, Raciocínio | Acurácia | 80.2 (MMLU), 52.8 (MATH) - Forte para modelo aberto | 98 |
| Llama 3 (405B) | Vários (MMLU, MATH, etc.) | Conhecimento, Raciocínio | Acurácia | ~87.3 (MMLU), 73.8 (MATH) - Paridade com modelos proprietários de topo | 98 |

*Nota: Os resultados são de diferentes fontes e momentos, podendo não ser perfeitamente comparáveis. MMLU e MATH não são benchmarks de viés, mas incluídos para contexto de desempenho geral dos LLMs frequentemente avaliados em justiça.*

Esta tabela centraliza informações fragmentadas de várias fontes, permitindo uma comparação mais direta do desempenho de modelos proeminentes em testes de viés padronizados. Ela aborda diretamente a necessidade do usuário de visualizar dados comparativos, facilitando a identificação de quais modelos podem exibir mais ou menos viés em métricas específicas e a análise de tendências gerais no campo.

### **C. Análise das Tendências de Desempenho**

A análise comparativa dos resultados dos modelos em benchmarks de viés revela várias tendências e complexidades:

* **Escala do Modelo vs. Justiça:** Não há uma relação direta e simples entre o tamanho do modelo (número de parâmetros) e seu nível de justiça. Modelos maiores não são inerentemente mais justos; o viés pode persistir ou até ser amplificado em modelos maiores.29 Embora alguns estudos sugiram que modelos maiores podem ter melhor desempenho em certos benchmarks devido à melhor generalização 71, outros fatores como os dados de treinamento e o fine-tuning parecem ser mais críticos.96 O desempenho da justiça é altamente dependente do modelo específico, da tarefa, do conjunto de dados e da métrica utilizada.43
* **Arquitetura do Modelo:** Diferentes arquiteturas podem exibir perfis de viés distintos. Por exemplo, em tarefas de classificação de toxicidade, modelos baseados em RNNs podem ter tempos de inferência mais rápidos, mas desempenho de viés diferente em comparação com modelos baseados em Transformer como o BERT.65
* **Dados de Treinamento:** Consistentemente, a fonte e a natureza dos dados de treinamento são identificadas como os principais impulsionadores do viés.5 Modelos treinados em dados da web em grande escala inevitavelmente herdam e refletem os vieses presentes nesses dados.89
* **Ajuste por Instrução (Instruction Tuning) e RLHF:** Técnicas de alinhamento como o ajuste por instrução e o Aprendizado por Reforço com Feedback Humano (RLHF) parecem desempenhar um papel crucial não apenas nas capacidades gerais do modelo 96, mas também em seus perfis de segurança e justiça.42 Modelos explicitamente ajustados para segurança podem recusar-se a responder a prompts enviesados com mais frequência.112 No entanto, o alinhamento para segurança geral não garante a ausência de viés em todos os cenários ou para todos os grupos, e pode até introduzir novos vieses ou limitações.83
* **Eficácia da Mitigação:** As técnicas de mitigação de viés mostram eficácia variável. Algumas podem funcionar bem para certos tipos de viés (por exemplo, viés de gênero) em determinados benchmarks, mas ser menos consistentes para outros (por exemplo, viés racial ou religioso).90 Além disso, a mitigação de viés pode ter custos, como uma redução na utilidade geral do modelo ou no desempenho em outras tarefas ou idiomas.29
* **Discrepâncias entre Benchmarks:** O desempenho de um modelo pode variar significativamente entre diferentes benchmarks de viés (por exemplo, StereoSet vs. CrowS-Pairs).38 Isso pode ser devido às diferentes metodologias, focos de viés ou problemas de validade dos próprios benchmarks. Além disso, as pontuações de benchmark relatadas podem ser infladas devido à contaminação dos dados de treinamento com dados de teste.31

A complexidade dessas tendências sublinha que não existe uma solução única ou uma relação simples entre as características de um modelo e seu desempenho em termos de justiça. A avaliação da justiça é altamente contextual, dependente do tipo específico de viés, do benchmark utilizado, dos dados de treinamento, das técnicas de alinhamento aplicadas e da métrica de avaliação escolhida. Alcançar a justiça frequentemente envolve trade-offs complexos com outras qualidades desejáveis do modelo, como precisão ou fluência.48

### **D. Desafios e Considerações na Avaliação Comparativa de Viés**

A comparação do desempenho de diferentes modelos em benchmarks de viés enfrenta vários desafios significativos:

* **Validade e Confiabilidade do Benchmark:** Como discutido anteriormente, a qualidade, validade e confiabilidade dos próprios benchmarks são uma preocupação central.27 Usar um benchmark falho pode levar a conclusões enganosas sobre o viés de um modelo.
* **Viés do Modelo de Métrica:** Benchmarks como o BOLD dependem de outros modelos (classificadores de toxicidade, sentimento, etc.) para calcular suas métricas.33 Se esses modelos de métrica forem eles próprios enviesados, eles introduzirão um viés na avaliação, confundindo os resultados.
* **Falta de Padronização:** Diferentes estudos podem usar versões diferentes de um benchmark, estratégias de prompting variadas (para LLMs), ou configurações de avaliação distintas, dificultando a comparação direta dos resultados relatados.32 A normalização de pontuações também pode afetar rankings.32
* **Limitações das Métricas Quantitativas:** A justiça é um conceito sócio-técnico complexo. Métricas quantitativas, embora úteis, não capturam todos os aspectos relevantes da justiça, como devido processo, contexto ou impacto diferencial no mundo real.14 A análise qualitativa e a consideração do contexto são indispensáveis.14
* **Dependência do Contexto:** Os requisitos de justiça variam enormemente dependendo do domínio da aplicação e dos danos potenciais envolvidos.14 Uma pontuação única em um benchmark genérico raramente é suficiente para determinar se um modelo é "justo" para um caso de uso específico.
* **Acesso a Modelos Abertos vs. Fechados:** Avaliar modelos fechados (como GPT-4, Claude via API) muitas vezes se limita à observação de suas saídas, impedindo o acesso a estados internos ou probabilidades de logits que são necessários para algumas metodologias de benchmark (como as usadas em StereoSet ou CrowS-Pairs para PLMs).72 Modelos abertos permitem uma análise mais profunda e a aplicação de uma gama mais ampla de métricas.72

Esses desafios convergem para uma conclusão importante: a avaliação comparativa baseada unicamente em classificações de benchmarks (leaderboards) pode ser superficial e potencialmente enganosa. Uma compreensão verdadeiramente nuançada do perfil de justiça de um modelo requer uma consideração crítica das limitações do benchmark, das escolhas de métricas, das especificidades do modelo (dados de treinamento, alinhamento) e, crucialmente, do contexto sócio-técnico em que a justiça está sendo avaliada. Uma visão holística é necessária para uma comparação significativa.14

## **VI. Conclusão e Recomendações**

A avaliação e mitigação de viés em modelos de Inteligência Artificial tornaram-se áreas de pesquisa e prática cruciais, impulsionadas pela crescente implantação desses sistemas em domínios de alto impacto. Os benchmarks de viés surgiram como ferramentas essenciais nesse esforço, mas sua aplicação eficaz, especialmente durante a fase de treinamento, requer uma compreensão nuançada de suas capacidades e limitações.

### **A. Síntese dos Principais Achados sobre Benchmarking de Viés no Treinamento de IA**

A análise apresentada neste relatório destaca vários pontos-chave:

1. **Natureza Multifacetada do Viés:** O viés em IA não é um fenômeno único, mas um termo abrangente para erros sistemáticos e resultados injustos que podem originar-se de dados, algoritmos ou fatores humanos ao longo do ciclo de vida do desenvolvimento. A compreensão evoluiu de uma perspectiva puramente estatística para uma sócio-técnica, reconhecendo os diversos danos sociais potenciais.
2. **Papel Crítico dos Benchmarks:** Benchmarks fornecem mecanismos padronizados indispensáveis para identificar, quantificar e comparar vieses. Eles permitem auditorias, medição de disparidades e monitoramento do progresso. No entanto, benchmarks existentes possuem limitações significativas relacionadas à validade dos dados, metodologias de pontuação, foco cultural e potencial para contaminação ou saturação.
3. **Importância da Intervenção no Treinamento:** Abordar o viés durante a fase de treinamento (*in-processing*) é uma estratégia proativa fundamental, pois permite mitigar as causas raízes do viés incorporadas nos dados ou algoritmos. Técnicas como regularização ciente da justiça e debiasing adversarial, guiadas por métricas de benchmark, podem construir modelos inerentemente mais justos.
4. **Complexidade do Desempenho de Justiça:** Não há uma relação simples entre características do modelo (tamanho, arquitetura) e seu desempenho em benchmarks de viés. A justiça é altamente dependente do contexto, influenciada por dados de treinamento, técnicas de alinhamento, tipo de viés, benchmark específico e métrica de avaliação. Frequentemente, existem trade-offs entre justiça e outras métricas de desempenho.
5. **Desafios Contínuos:** A criação de benchmarks de viés robustos, válidos e culturalmente sensíveis permanece um desafio. O viés potencial nos próprios modelos de métrica usados para avaliação e a falta de padronização complicam ainda mais as comparações. A avaliação da justiça requer ir além das métricas quantitativas para incluir análises qualitativas e considerações contextuais.

### **B. Recomendações Acionáveis para Profissionais de IA**

Com base nesses achados, as seguintes recomendações são propostas para profissionais que desenvolvem e implantam sistemas de IA:

1. **Diligência nos Dados:** Examine criticamente os dados de treinamento em busca de vieses potenciais (representação, histórico, medição, etc.). Utilize ferramentas de análise de dados e envolva perspectivas diversas nesse processo.5 Priorize a coleta ou criação de conjuntos de dados diversos e representativos sempre que possível.5
2. **Alfabetização em Benchmarks:** Desenvolva uma compreensão profunda dos benchmarks de viés utilizados. Conheça sua metodologia, as métricas empregadas, o tipo específico de viés que visam medir e suas limitações conhecidas.21 Evite confiar excessivamente em pontuações únicas ou classificações de leaderboard sem contexto.
3. **Integração Precoce e Contínua da Justiça:** Incorpore considerações de justiça desde as fases iniciais de concepção do projeto e continue a avaliá-las ao longo de todo o ciclo de vida do desenvolvimento, com ênfase especial nas fases de preparação de dados e treinamento do modelo.5
4. **Utilização de Mitigação In-Processing:** Quando apropriado e viável, empregue técnicas de mitigação *in-processing*, como regularização ciente da justiça ou debiasing adversarial, durante o treinamento do modelo. Use métricas de justiça relevantes, informadas por benchmarks, para guiar essas técnicas.14
5. **Testes Abrangentes e Interseccionais:** Avalie os modelos usando múltiplas métricas de justiça e diversos conjuntos de dados de benchmark. Desagregue os resultados por subgrupos demográficos relevantes, prestando atenção especial às interseções (por exemplo, raça e gênero combinados).10 Continue monitorando o desempenho e a justiça após a implantação para detectar vieses emergentes ou mudanças de distribuição.7
6. **Transparência e Documentação:** Documente de forma transparente as fontes de dados, as limitações conhecidas do modelo, os resultados da avaliação de justiça e as estratégias de mitigação empregadas. Considere o uso de ferramentas como "model cards" para comunicar essas informações.11
7. **Supervisão Humana:** Mantenha a supervisão humana no ciclo, especialmente em aplicações de alto risco onde decisões enviesadas podem causar danos significativos. A IA deve auxiliar, não substituir completamente, o julgamento humano em contextos críticos.7

### **C. Perspectivas: O Futuro do Benchmarking e Desenvolvimento de IA Justa**

O campo da justiça em IA está em rápida evolução. As direções futuras provavelmente incluirão:

* **Benchmarks Mais Robustos:** Desenvolvimento de benchmarks com maior rigor metodológico, validação psicométrica aprimorada, maior sensibilidade cultural e que abordem uma gama mais ampla de tipos de viés e grupos sociais.28
* **Métricas Aprimoradas:** Criação de métricas de avaliação de viés que sejam menos suscetíveis a vieses próprios e que capturem melhor as nuances da justiça em diferentes contextos.33
* **Mitigação Mais Eficaz:** Avanços em técnicas de mitigação (especialmente *in-processing*) que equilibrem melhor a justiça e a utilidade do modelo, minimizando trade-offs negativos.43
* **Foco na Interseccionalidade:** Maior atenção à avaliação e mitigação de vieses interseccionais e vieses que afetam grupos sub-representados na pesquisa atual.10
* **Avaliação Dinâmica e Realista:** Transição de benchmarks estáticos para cenários de avaliação mais dinâmicos, adversariais ou baseados no mundo real que reflitam melhor como os vieses se manifestam na prática.83
* **Abordagem Sócio-Técnica:** Aprofundamento da integração de considerações éticas, perspectivas das ciências sociais e envolvimento das partes interessadas no processo de desenvolvimento e avaliação de IA, reconhecendo a justiça como um desafio inerentemente sócio-técnico.14

Em suma, garantir a justiça em sistemas de IA é um esforço contínuo que exige vigilância constante, ferramentas de avaliação críticas e um compromisso com a mitigação proativa, especialmente durante a fase crucial de treinamento do modelo. Os benchmarks de viés, apesar de suas imperfeições, são componentes vitais desse esforço, fornecendo os meios para medir, compreender e, finalmente, reduzir os vieses indesejados em tecnologias de IA.

#### Referências citadas

1. Bias Estimation in Machine Learning: Definition, Causes, and Mitigation Strategies, acessado em abril 22, 2025, <https://kili-technology.com/data-labeling/machine-learning/bias-estimation-a-complete-guide-for-machine-learning-engineers>
2. Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects - Diagnostic and Interventional Radiology, acessado em abril 22, 2025, <https://www.dirjournal.org/articles/bias-in-artificial-intelligence-for-medical-imaging-fundamentals-detection-avoidance-mitigation-challenges-ethics-and-prospects/doi/dir.2024.242854>
3. kili-technology.com, acessado em abril 22, 2025, <https://kili-technology.com/data-labeling/machine-learning/bias-estimation-a-complete-guide-for-machine-learning-engineers#:~:text=Bias%20in%20Machine%20Learning%3A%20Definition,-Bias%20in%20machine&text=In%20other%20words%2C%20it%20is,on%20limited%20or%20incomplete%20data.>
4. Unveiling AI Bias: Understanding Its Ethical Implications - Robert F. Smith, acessado em abril 22, 2025, <https://robertsmith.com/blog/ai-bias/>
5. Bias in Machine Learning - Lightly.ai, acessado em abril 22, 2025, <https://www.lightly.ai/blog/bias-in-machine-learning>
6. AI Bias: Definition, Types, Examples, and Debiasing Strategies - ITRex, acessado em abril 22, 2025, <https://itrexgroup.com/blog/ai-bias-definition-types-examples-debiasing-strategies/>
7. Bias in AI - Chapman University, acessado em abril 22, 2025, <https://www.chapman.edu/ai/bias-in-ai.aspx>
8. COMPAS Recidivism Racial Bias - Kaggle, acessado em abril 22, 2025, <https://www.kaggle.com/datasets/danofer/compass/data>
9. COMPAS Recidivism Racial Bias - Kaggle, acessado em abril 22, 2025, <https://www.kaggle.com/datasets/danofer/compass>
10. Overview ‹ Gender Shades - MIT Media Lab, acessado em abril 22, 2025, <https://www.media.mit.edu/projects/gender-shades/overview/>
11. How 'Gender Shades' Sheds Light on Bias in Machine Learning, acessado em abril 22, 2025, <https://www.dataprivacyadvisory.com/how-gender-shades-sheds-light-on-bias-in-machine-learning/>
12. Addressing Gender Bias in Facial Recognition Technology: An Urgent Need for Fairness and Inclusion - Cogent Infotech, acessado em abril 22, 2025, <https://www.cogentinfo.com/resources/addressing-gender-bias-in-facial-recognition-technology-an-urgent-need-for-fairness-and-inclusion>
13. Fairness: Types of bias | Machine Learning - Google for Developers, acessado em abril 22, 2025, <https://developers.google.com/machine-learning/crash-course/fairness/types-of-bias>
14. Machine learning fairness - Azure Machine Learning | Microsoft Learn, acessado em abril 22, 2025, <https://learn.microsoft.com/en-us/azure/machine-learning/concept-fairness-ml?view=azureml-api-2>
15. Fairness in Machine Learning — Fairlearn 0.13.0.dev0 documentation, acessado em abril 22, 2025, <https://fairlearn.org/main/user_guide/fairness_in_machine_learning.html>
16. A Critical Review of Predominant Bias in Neural Networks - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2502.11031v1>
17. Benchmarking Bias Mitigation Algorithms in Representation Learning through Fairness Metrics, acessado em abril 22, 2025, <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/2723d092b63885e0d7c260cc007e8b9d-Paper-round1.pdf>
18. BOLD: Dataset and metrics for measuring biases in open-ended language generation, acessado em abril 22, 2025, <http://web.cs.ucla.edu/~kwchang/bibliography/Dhamala2021bold/>
19. Daily Papers - Hugging Face, acessado em abril 22, 2025, <https://huggingface.co/papers?q=Crows-Pairs>
20. Evaluate model and system for safety | Responsible Generative AI Toolkit - Gemini API, acessado em abril 22, 2025, <https://ai.google.dev/responsible/docs/evaluation>
21. Fairness in Machine Learning | dida blog, acessado em abril 22, 2025, <https://dida.do/blog/fairness-in-ml>
22. Civil Comments Dataset - Papers With Code, acessado em abril 22, 2025, <https://paperswithcode.com/dataset/civil-comments>
23. StereoSet Benchmark (Bias Detection) | Papers With Code, acessado em abril 22, 2025, <https://paperswithcode.com/sota/bias-detection-on-stereoset-1>
24. Gender Shades, acessado em abril 22, 2025, <http://gendershades.org/overview.html>
25. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models - ACL Anthology, acessado em abril 22, 2025, <https://aclanthology.org/2020.emnlp-main.154.pdf>
26. StereoSet: Measuring stereotypical bias in pretrained language models - Papers With Code, acessado em abril 22, 2025, <https://paperswithcode.com/paper/stereoset-measuring-stereotypical-bias-in/review/>
27. README.md - nyu-mll/crows-pairs - GitHub, acessado em abril 22, 2025, <https://github.com/nyu-mll/crows-pairs/blob/master/README.md>
28. arXiv:2302.12640v1 [cs.CL] 24 Feb 2023, acessado em abril 22, 2025, <https://arxiv.org/pdf/2302.12640>
29. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets, acessado em abril 22, 2025, <https://www.researchgate.net/publication/353487240_Stereotyping_Norwegian_Salmon_An_Inventory_of_Pitfalls_in_Fairness_Benchmark_Datasets>
30. In-Depth Look at Word Filling Societal Bias Measures - ACL Anthology, acessado em abril 22, 2025, <https://aclanthology.org/2023.eacl-main.265.pdf>
31. Benchmark Inflation: Revealing LLM Performance Gaps Using Retro-Holdouts, acessado em abril 22, 2025, <https://openreview.net/forum?id=WdA5H9ARaa>
32. Open-LLM performances are plateauing, let's make the leaderboard steep again - Hugging Face, acessado em abril 22, 2025, <https://huggingface.co/spaces/open-llm-leaderboard/blog>
33. Assessing Bias in Metric Models for LLM Open-Ended Generation Bias Benchmarks - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2410.11059v1>
34. Assessing Bias in Metric Models for LLM Open-Ended Generation Bias Benchmarks - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2410.11059>
35. debiaSAE: Benchmarking and Mitigating Vision-Language Model Bias - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2410.13146v2>
36. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation - arXiv, acessado em abril 22, 2025, <https://arxiv.org/pdf/2101.11718>
37. You Reap What You Sow: On the Challenges of Bias Evaluation Under Multilingual Settings - ACL Anthology, acessado em abril 22, 2025, <https://aclanthology.org/2022.bigscience-1.3.pdf>
38. Blind Men and the Elephant: Diverse Perspectives on Gender Stereotypes in Benchmark Datasets - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2501.01168v1>
39. Undesirable Biases in NLP: Addressing Challenges of Measurement - Journal of Artificial Intelligence Research, acessado em abril 22, 2025, <https://jair.org/index.php/jair/article/download/15195/26998/37495>
40. Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2207.07068>
41. Recommendations for Bias Mitigation Methods: Applicability and Legality - King's Research Portal, acessado em abril 22, 2025, <https://kclpure.kcl.ac.uk/portal/files/228505595/WallerAEQUITASWorkshopFinal.pdf>
42. A Survey on Fairness in Large Language Models - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2308.10149v2>
43. FairLoRA: Unpacking Bias Mitigation in Vision Models with Fairness-Driven Low-Rank Adaptation - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2410.17358v1>
44. Adversarial Debiasing — holisticai documentation - Read the Docs, acessado em abril 22, 2025, <https://holisticai.readthedocs.io/en/latest/getting_started/bias/mitigation/inprocessing/bc_adversarial_debiasing_adversarial_debiasing.html>
45. What role does regularization play in promoting fairness in machine learning models?, acessado em abril 22, 2025, [https://infermatic.ai/ask/?question=What%20role%20does%20regularization%20play%20in%20promoting%20fairness%20in%20machine%20learning%20models?](https://infermatic.ai/ask/?question=What+role+does+regularization+play+in+promoting+fairness+in+machine+learning+models?)
46. Bias Mitigation Strategies and Techniques for Classification Tasks - Holistic AI, acessado em abril 22, 2025, <https://www.holisticai.com/blog/bias-mitigation-strategies-techniques-for-classification-tasks>
47. Algorithm fairness in artificial intelligence for medicine and healthcare - PMC - PubMed Central, acessado em abril 22, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10632090/>
48. Chapter 10 Story COMPAS: recidivism reloaded | XAI Stories, acessado em abril 22, 2025, <https://pbiecek.github.io/xai_stories/story-compas.html>
49. Fairness Metrics in Machine Learning - Coralogix, acessado em abril 22, 2025, <https://coralogix.com/ai-blog/fairness-metrics-in-machine-learning/>
50. COMPAS Dataset - R, acessado em abril 22, 2025, <https://search.r-project.org/CRAN/refmans/mlr3fairness/help/compas.html>
51. COMPAS Dataset - mlr3fairness, acessado em abril 22, 2025, <https://mlr3fairness.mlr-org.com/reference/compas.html>
52. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification - Proceedings of Machine Learning Research, acessado em abril 22, 2025, <https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>
53. Bias in open-ended language generation dataset (BOLD) - Amazon Science, acessado em abril 22, 2025, <https://www.amazon.science/code-and-datasets/bias-in-open-ended-language-generation-dataset-bold>
54. SB-Bench: Stereotype Bias Benchmark for Large Multimodal Models - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2502.08779v2>
55. StereoSet Dataset | Papers With Code, acessado em abril 22, 2025, <https://paperswithcode.com/dataset/stereoset>
56. CrowS-Pairs Dataset - Papers With Code, acessado em abril 22, 2025, <https://paperswithcode.com/dataset/crows-pairs>
57. Jigsaw Unintended Bias in Toxicity Classification - Kaggle, acessado em abril 22, 2025, <https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification>
58. Gender shades - DSpace@MIT, acessado em abril 22, 2025, <https://dspace.mit.edu/bitstream/handle/1721.1/114068/1026503582-MIT.pdf>
59. Frequently Asked Questions ‹ Gender Shades - MIT Media Lab, acessado em abril 22, 2025, <https://www.media.mit.edu/projects/gender-shades/faq/>
60. Results ‹ Gender Shades - MIT Media Lab, acessado em abril 22, 2025, <https://www.media.mit.edu/projects/gender-shades/results/>
61. Bias Discovery within Human Raters: A Case Study of the Jigsaw Dataset - ACL Anthology, acessado em abril 22, 2025, <https://aclanthology.org/2022.nlperspectives-1.4.pdf>
62. View of Empirical Analysis of Multi-Task Learning for Reducing Identity Bias in Toxic Comment Detection - AAAI Publications, acessado em abril 22, 2025, <https://ojs.aaai.org/index.php/ICWSM/article/view/7334/7188>
63. README.md · google/civil\_comments at main - Hugging Face, acessado em abril 22, 2025, <https://huggingface.co/datasets/google/civil_comments/blob/main/README.md>
64. Jigsaw Unintended Bias in Toxicity Classification | Kaggle, acessado em abril 22, 2025, <https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data>
65. arXiv:2301.11125v1 [cs.CL] 26 Jan 2023, acessado em abril 22, 2025, <https://arxiv.org/pdf/2301.11125>
66. A benchmark for toxic comment classification on Civil Comments dataset - arXiv, acessado em abril 22, 2025, <https://arxiv.org/abs/2301.11125>
67. An Empirical Investigation of Learning from Biased Toxicity Labels - arXiv, acessado em abril 22, 2025, <https://arxiv.org/pdf/2110.01577>
68. Determination of toxic comments and unintended model bias minimization using Deep learning approach - arXiv, acessado em abril 22, 2025, <https://arxiv.org/pdf/2311.04789>
69. Toxic Comment Classification and Unintended Bias - CS229, acessado em abril 22, 2025, <https://cs229.stanford.edu/proj2019spr/report/69.pdf>
70. BOLD: Dataset and metrics for measuring biases in open-ended language generation - Amazon Science, acessado em abril 22, 2025, <https://www.amazon.science/publications/bold-dataset-and-metrics-for-measuring-biases-in-open-ended-language-generation>
71. Assessing Biases in LLMs: From Basic Tasks to Hiring Decisions - Holistic AI, acessado em abril 22, 2025, <https://www.holisticai.com/blog/assessing-biases-in-llms>
72. Evaluating Stereotypical Biases and Implications for Fairness in Large Language Models - ASEE PEER, acessado em abril 22, 2025, <https://peer.asee.org/evaluating-stereotypical-biases-and-implications-for-fairness-in-large-language-models.pdf>
73. EasyChair Preprint Prompts De-Biasing Augmentation to Mitigate Gender Stereotypes in Large Language Models, acessado em abril 22, 2025, <https://easychair.org/publications/preprint/dTqw/download>
74. Layered Bias: Interpreting Bias in Pretrained Large Language Models - ACL Anthology, acessado em abril 22, 2025, <https://aclanthology.org/2023.blackboxnlp-1.22.pdf>
75. LIBRA: Measuring Bias of Large Language Model from a Local Context - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2502.01679v1>
76. aclanthology.org, acessado em abril 22, 2025, <https://aclanthology.org/2023.eacl-main.89.pdf>
77. Spoken Stereoset: On Evaluating Social Bias Toward Speaker in Speech Large Language Models - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2408.07665v1>
78. The Devil is in the Neurons: Interpreting and Mitigating Social Biases in Pre-trained Language Models - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2406.10130v1>
79. StereoSet: Measuring stereotypical bias in pretrained language models | Papers With Code, acessado em abril 22, 2025, <https://paperswithcode.com/paper/stereoset-measuring-stereotypical-bias-in>
80. lm-evaluation-harness/lm\_eval/tasks/crows\_pairs/README.md at main - GitHub, acessado em abril 22, 2025, <https://github.com/EleutherAI/lm-evaluation-harness/blob/main/lm_eval/tasks/crows_pairs/README.md>
81. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models | Request PDF - ResearchGate, acessado em abril 22, 2025, <https://www.researchgate.net/publication/344447505_CrowS-Pairs_A_Challenge_Dataset_for_Measuring_Social_Biases_in_Masked_Language_Models>
82. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models | Papers With Code, acessado em abril 22, 2025, <https://paperswithcode.com/paper/crows-pairs-a-challenge-dataset-for-measuring>
83. arXiv:2503.19540v1 [cs.CL] 25 Mar 2025, acessado em abril 22, 2025, <https://arxiv.org/pdf/2503.19540>
84. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models | Request PDF - ResearchGate, acessado em abril 22, 2025, <https://www.researchgate.net/publication/347235439_CrowS-Pairs_A_Challenge_Dataset_for_Measuring_Social_Biases_in_Masked_Language_Models>
85. This repository contains the data and code introduced in the paper "CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models" (EMNLP 2020). - GitHub, acessado em abril 22, 2025, <https://github.com/nyu-mll/crows-pairs>
86. CrowS-Pairs Benchmark (Stereotypical Bias Analysis) - Papers With Code, acessado em abril 22, 2025, <https://paperswithcode.com/sota/stereotypical-bias-analysis-on-crows-pairs>
87. Bias and Fairness in Large Language Models: A Survey - MIT Press Direct, acessado em abril 22, 2025, <https://direct.mit.edu/coli/article/50/3/1097/121961/Bias-and-Fairness-in-Large-Language-Models-A>
88. Benchmarking Adversarial Robustness to Bias Elicitation in Large Language Models: Scalable Automated Assessment with LLM-as-a-Judge - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2504.07887v1>
89. A Decade's Battle on Dataset Bias: Are We There Yet? - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2403.08632v2>
90. An Empirical Analysis of Parameter-Efficient Methods for Debiasing Pre-Trained Language Models - ACL Anthology, acessado em abril 22, 2025, <https://aclanthology.org/2023.acl-long.876.pdf>
91. Fairness-aware Learning through Regularization Approach - Toshihiro Kamishima, acessado em abril 22, 2025, <https://www.kamishima.net/archive/2011-ws-icdm_padm.pdf>
92. Adversarial Debiasing | Envisioning Vocab, acessado em abril 22, 2025, <https://www.envisioning.io/vocab/adversarial-debiasing>
93. aif360.algorithms.inprocessing.AdversarialDebiasing - Read the Docs, acessado em abril 22, 2025, <https://aif360.readthedocs.io/en/latest/modules/generated/aif360.algorithms.inprocessing.AdversarialDebiasing.html>
94. A Prompt Array Keeps the Bias Away: Debiasing Vision-Language Models with Adversarial Learning | Montreal AI Ethics Institute, acessado em abril 22, 2025, <https://montrealethics.ai/a-prompt-array-keeps-the-bias-away-debiasing-vision-language-models-with-adversarial-learning/>
95. Multi-task learning for toxic comment classification and rationale extraction - PMC, acessado em abril 22, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC9391651/>
96. Benchmarking Large Language Models for News Summarization - MIT Press Direct, acessado em abril 22, 2025, <https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00632/119276/Benchmarking-Large-Language-Models-for-News>
97. Comparison of new Open-LLM-Leaderboard benchmarks vs. frontier models? - Reddit, acessado em abril 22, 2025, <https://www.reddit.com/r/LocalLLaMA/comments/1dpbbzl/comparison_of_new_openllmleaderboard_benchmarks/>
98. Llama 4: Meta's New AI Model - Evolution, Features, and Comparison | GPT-trainer Blog, acessado em abril 22, 2025, <https://gpt-trainer.com/blog/llama+4+evolution+features+comparison>
99. (PDF) Parity benchmark for measuring bias in LLMs - ResearchGate, acessado em abril 22, 2025, <https://www.researchgate.net/publication/387137967_Parity_benchmark_for_measuring_bias_in_LLMs>
100. GPT-4o Benchmark - Detailed Comparison with Claude & Gemini - Wielded, acessado em abril 22, 2025, <https://wielded.com/blog/gpt-4o-benchmark-detailed-comparison-with-claude-and-gemini>
101. ChatGPT vs Gemini vs Llama vs Meta AI vs Claude: The Class of the Chatbot Titans, acessado em abril 22, 2025, <https://gaper.io/chatgpt-vs-gemini-vs-llama-vs-meta-ai-vs-claude/>
102. STOP! Benchmarking Large Language Models with Sensitivity Testing on Offensive Progressions - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2409.13843v2>
103. BiasAlert: A Plug-and-play Tool for Social Bias Detection in LLMs - ACL Anthology, acessado em abril 22, 2025, <https://aclanthology.org/2024.emnlp-main.820.pdf>
104. FairMT-Bench: Benchmarking Fairness for Multi-turn Dialogue in Conversational LLMs, acessado em abril 22, 2025, <https://openreview.net/forum?id=RSGoXnS9GH>
105. STOP! Benchmarking Large Language Models with Sensitivity Testing on Offensive Progressions - ACL Anthology, acessado em abril 22, 2025, <https://aclanthology.org/2024.emnlp-main.243.pdf>
106. All Should Be Equal in the Eyes of LMs: Counterfactually Aware Fair Text Generation - AAAI Publications, acessado em abril 22, 2025, <https://ojs.aaai.org/index.php/AAAI/article/view/29719/31235>
107. I tested Gemini 2.5 Pro against Claude 3.7 Sonnet (thinking): Google is clearly after Anthropic's lunch - Reddit, acessado em abril 22, 2025, <https://www.reddit.com/r/ClaudeAI/comments/1jnl8pn/i_tested_gemini_25_pro_against_claude_37_sonnet/>
108. Revealing Hidden Bias in AI: Lessons from Large Language Models - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2410.16927v1>
109. GPT-4o vs. Gemini 1.5 Pro vs. Claude 3 Opus: Multimodal AI Model Comparison - Encord, acessado em abril 22, 2025, <https://encord.com/blog/gpt-4o-vs-gemini-vs-claude-3-opus/>
110. Gemini 1.5 Pro vs GPT 4 vs Claude 3, which is better? : r/Bard - Reddit, acessado em abril 22, 2025, <https://www.reddit.com/r/Bard/comments/1feedgr/gemini_15_pro_vs_gpt_4_vs_claude_3_which_is_better/>
111. HEARTS: A Holistic Framework for Explainable, Sustainable and Robust Text Stereotype Detection - arXiv, acessado em abril 22, 2025, <http://www.arxiv.org/pdf/2409.11579>
112. How Are LLMs Mitigating Stereotyping Harms? Learning from Search Engine Studies - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2407.11733v2>
113. Stereotypical Bias Analysis - Papers With Code, acessado em abril 22, 2025, <https://paperswithcode.com/task/stereotypical-bias-analysis>
114. An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models - Papers With Code, acessado em abril 22, 2025, <https://paperswithcode.com/paper/an-empirical-survey-of-the-effectiveness-of/review/>
115. arXiv:2412.14050v3 [cs.CL] 14 Feb 2025, acessado em abril 22, 2025, <https://arxiv.org/pdf/2412.14050>
116. The Other Side of the Coin: Exploring Fairness in Retrieval-Augmented Generation - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2504.12323v2>
117. Different Bias Under Different Criteria: Assessing Bias in LLMs with a Fact-Based Approach, acessado em abril 22, 2025, <https://arxiv.org/html/2411.17338v1>
118. FFT: Towards Evaluating Large Language Models with Factuality, Fairness, Toxicity - GitHub Pages, acessado em abril 22, 2025, <https://genai-evaluation-kdd2024.github.io/genai-evalution-kdd2024/assets/papers/GenAI_Evaluation_KDD2024_paper_5.pdf>
119. VLBiasBench: A Comprehensive Benchmark for Evaluating Bias in Large Vision-Language Model - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2406.14194v2>
120. VLBiasBench: A Comprehensive Benchmark for Evaluating Bias in Large Vision-Language Model - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2406.14194v1>