# **Análise Detalhada dos Principais Benchmarks de Viés em Inteligência Artificial**

## **1. Introdução**

### **1.1 A Importância dos Benchmarks de Viés em IA**

A inteligência artificial (IA) permeia cada vez mais diversos aspectos da sociedade, desde sistemas de recomendação e tradução automática até aplicações de alto risco em saúde, justiça criminal e contratação.1 No entanto, juntamente com as capacidades avançadas desses sistemas, surgem preocupações significativas sobre seu potencial para perpetuar e até amplificar vieses sociais existentes.1 O viés em IA, frequentemente definido como erros sistemáticos ou desvios que levam a resultados injustos ou imprecisos para determinados grupos 23, pode manifestar-se de várias formas. Exemplos notórios incluem disparidades de desempenho em sistemas de reconhecimento facial comercial, que demonstraram taxas de erro significativamente mais altas para mulheres de pele escura em comparação com homens de pele clara 14, a replicação de estereótipos de gênero e raciais em modelos de linguagem que geram texto 3 e avaliações de risco de reincidência criminal que atribuem pontuações de risco mais altas a réus negros em comparação com réus brancos, mesmo controlando outros fatores.33

Esses vieses geralmente se originam dos dados massivos usados para treinar os modelos de IA, que podem refletir desigualdades históricas, sub-representação de certos grupos ou preconceitos sociais presentes na sociedade.1 Além disso, as próprias escolhas de design do algoritmo e as métricas de otimização podem introduzir ou exacerbar vieses.2 Dada a escala e o impacto potencial desses sistemas, a avaliação e mitigação de vieses tornaram-se áreas críticas de pesquisa e desenvolvimento em IA.

Nesse contexto, os benchmarks de viés surgem como ferramentas essenciais. Eles fornecem conjuntos de dados e métricas padronizadas que permitem quantificar a presença e a magnitude de diferentes tipos de vieses em modelos de IA.12 Esses benchmarks facilitam a comparação entre diferentes modelos ou técnicas de mitigação, ajudam a rastrear o progresso ao longo do tempo e aumentam a transparência sobre as limitações dos sistemas de IA. A própria existência de uma gama crescente de benchmarks 4, cobrindo tarefas como classificação de imagens, processamento de linguagem natural e avaliação de risco, reflete uma maior conscientização sobre a complexidade do problema da justiça (fairness) em IA. No entanto, essa proliferação também sinaliza a ausência de uma definição única e universalmente aceita de "viés" ou "justiça" e a falta de um método único para sua medição. Diferentes benchmarks podem operacionalizar o conceito de viés de maneiras distintas, levando potencialmente a conclusões variadas sobre o mesmo modelo.6 Portanto, a seleção e a interpretação de resultados de benchmarks de viés exigem uma compreensão cuidadosa de seus objetivos, metodologias e limitações intrínsecas.

### **1.2 Foco do Relatório**

Este relatório concentra-se na análise detalhada de seis benchmarks proeminentes desenvolvidos para avaliar diferentes tipos de viés em sistemas de IA: Gender Shades, COMPAS, Civil Comments/Jigsaw Unintended Bias, BOLD (Bias in Open-Ended Language Generation), StereoSet e CrowS-Pairs. Para cada um desses benchmarks, o relatório examinará:

* **Objetivo Principal:** A finalidade primária para a qual o benchmark foi criado.
* **Metodologia de Avaliação:** Como o benchmark avalia os modelos de IA.
* **Conjunto de Dados:** As características do(s) conjunto(s) de dados utilizado(s).
* **Tipos de Viés Medidos:** As categorias específicas de viés que o benchmark visa quantificar.
* **Métricas de Avaliação Chave:** As principais métricas usadas para reportar os resultados.
* **Modelos Avaliados e Resultados:** Exemplos de modelos de IA que foram avaliados usando o benchmark e seus resultados quantitativos, quando disponíveis.
* **Limitações e Críticas:** As limitações conhecidas do benchmark ou críticas levantadas sobre sua validade ou aplicabilidade.
* **Fontes Principais:** Referências aos documentos e recursos primários que descrevem o benchmark.

O relatório priorizará, sempre que possível, a utilização de terminologia e fontes em português, recorrendo ao inglês técnico e a fontes originais em inglês quando necessário, com base no material de pesquisa disponível. O objetivo é fornecer uma análise estruturada e aprofundada que sirva como um recurso técnico para pesquisadores, praticantes e estudantes interessados em compreender e utilizar esses importantes instrumentos de avaliação de justiça em IA.

## **2. Análise dos Benchmarks de Viés**

### **2.1 Gender Shades**

#### **2.1.1 Objetivo Principal**

O benchmark Gender Shades teve como objetivo principal avaliar a precisão de sistemas comerciais de classificação de gênero baseados em inteligência artificial, especificamente em relação a diferentes grupos demográficos definidos pela interseção de gênero e tipo de pele.28 O projeto surgiu da necessidade de expor e quantificar vieses em tecnologias de análise facial, definindo viés, neste contexto, como diferenças práticas e significativas nas taxas de erro de classificação entre esses grupos.30 Além de medir o desempenho, Gender Shades buscou destacar a falta de transparência das empresas de tecnologia sobre o desempenho de seus produtos em subgrupos demográficos e defender a necessidade urgente de testes mais inclusivos e relatórios de precisão desagregados.29 Em última análise, o projeto visava motivar a indústria a desenvolver sistemas de IA mais justos, transparentes e responsáveis, especialmente aqueles aplicados a seres humanos.15

#### **2.1.2 Metodologia de Avaliação**

A metodologia do Gender Shades centrou-se numa abordagem interseccional para avaliar a precisão da classificação de gênero.14 Em vez de analisar gênero e tipo de pele separadamente, o estudo examinou o desempenho dos sistemas nos quatro subgrupos formados pela sua interseção: mulheres de pele escura, homens de pele escura, mulheres de pele clara e homens de pele clara.

Para realizar a avaliação, foi criado um conjunto de dados de benchmark específico, o Pilot Parliaments Benchmark (PPB), projetado para ser mais equilibrado em termos de gênero e tipo de pele do que os benchmarks existentes na época.15

Três sistemas comerciais de classificação de gênero, oferecidos como APIs pelas empresas IBM, Microsoft e Face++, foram selecionados e tratados como "caixas pretas".28 Imagens do conjunto de dados PPB foram enviadas para essas APIs, e as classificações de gênero retornadas foram comparadas com os rótulos de verdade terrestre (ground truth) do PPB para calcular as taxas de erro para cada subgrupo interseccional. A análise focou nas disparidades dessas taxas de erro entre os grupos.

#### **2.1.3 Conjunto de Dados: Pilot Parliaments Benchmark (PPB)**

O conjunto de dados Pilot Parliaments Benchmark (PPB) foi especificamente construído para o estudo Gender Shades. Ele é composto por 1.270 imagens faciais únicas.28 As imagens são de parlamentares de seis países: três países africanos (Ruanda, Senegal, África do Sul) e três países europeus (Islândia, Finlândia, Suécia).28 A seleção desses países foi baseada, em parte, na paridade de gênero em seus parlamentos na época da coleta (maio de 2017).29

As imagens foram rotuladas segundo duas dimensões principais:

* **Gênero:** Utilizaram-se rótulos binários (feminino, masculino) baseados no gênero percebido dos parlamentares. Os autores reconhecem que esta é uma visão reducionista do gênero, mas foi uma limitação herdada das APIs comerciais avaliadas, que forneciam apenas classificações binárias de sexo/gênero.30
* **Tipo de Pele:** Um dermatologista certificado rotulou as faces usando a escala Fitzpatrick de seis pontos (I a VI).15 Para a análise interseccional, esses seis tipos foram agrupados em duas categorias: "pele clara" (Fitzpatrick I, II, III) e "pele escura" (Fitzpatrick IV, V, VI).30

O PPB foi concebido para ser mais equilibrado demograficamente do que os benchmarks de análise facial existentes na época, como IJB-A e Adience, que se mostraram predominantemente compostos por indivíduos de pele clara (79,6% e 86,2%, respetivamente).15 O PPB resultante continha 44,39% de indivíduos do sexo feminino e 47% de indivíduos de pele escura.60 Devido a regulamentações de privacidade como o GDPR, as imagens originais do PPB não estão mais disponíveis publicamente, mas os resultados das APIs aplicadas a este conjunto de dados foram disponibilizados (GS5-API-Results).61

#### **2.1.4 Tipos de Viés Medidos**

Gender Shades mediu vieses de desempenho na tarefa de classificação de gênero, focando em disparidades entre grupos demográficos:

* **Viés de Gênero:** Diferença na precisão da classificação entre sujeitos masculinos e femininos.30
* **Viés de Tipo de Pele (Viés Fenotípico):** Diferença na precisão entre sujeitos de pele clara e de pele escura.29
* **Viés Interseccional:** O viés exacerbado observado em subgrupos na interseção de gênero e tipo de pele. O estudo demonstrou que mulheres de pele escura experimentaram as taxas de erro mais altas, indicando um viés composto.14

#### **2.1.5 Métricas de Avaliação Chave**

As principais métricas utilizadas para quantificar o desempenho e o viés foram:

* **Taxa de Erro de Classificação (%):** A porcentagem de vezes que o sistema classificou incorretamente o gênero de um rosto.15
* **Precisão (%):** A porcentagem de vezes que o sistema classificou corretamente o gênero (100% - Taxa de Erro).15
* **Diferença na Taxa de Erro (Disparidade):** A métrica central para quantificar o viés. É a diferença absoluta nas taxas de erro entre dois grupos comparados (por exemplo, homens vs. mulheres, pele clara vs. pele escura, ou entre subgrupos interseccionais).30
* **Maior Diferença (Largest Gap):** A diferença máxima na taxa de erro observada entre o subgrupo com melhor desempenho e o subgrupo com pior desempenho no benchmark.30

#### **2.1.6 Modelos Avaliados e Resultados**

Os modelos avaliados foram as APIs comerciais de classificação de gênero da IBM, Microsoft e Face++, conforme disponíveis no final de 2017 e início de 2018.14

**Principais Resultados:**

* Todos os sistemas apresentaram desempenho consistentemente melhor em rostos masculinos do que em femininos, com diferenças nas taxas de erro variando de 8,1% a 20,6%.30
* Todos os sistemas tiveram melhor desempenho em rostos de pele clara do que em rostos de pele escura, com diferenças nas taxas de erro de 11,8% a 19,2%.30
* A análise interseccional revelou que **todos os sistemas tiveram o pior desempenho em mulheres de pele escura**.15
* As taxas de erro para mulheres de pele escura foram: IBM (34,7%), Microsoft (20,8%) e Face++ (34,5%).15
* As taxas de erro para homens de pele clara foram: IBM (0,3%), Microsoft (0,0%) e Face++ (0,9%).30
* A maior diferença na taxa de erro (entre homens de pele clara e mulheres de pele escura) foi observada na API da IBM, atingindo 34,4 pontos percentuais.30

A tabela a seguir resume as taxas de erro (%) para os quatro subgrupos interseccionais nos três sistemas avaliados:

| **Grupo Interseccional** | **Taxa de Erro IBM (%)** | **Taxa de Erro Microsoft (%)** | **Taxa de Erro Face++ (%)** |
| --- | --- | --- | --- |
| Mulheres Pele Escura | 34.7 | 20.8 | 34.5 |
| Mulheres Pele Clara | 7.1 | 6.0 | 4.4 |
| Homens Pele Escura | 12.0 | 5.6 | 4.0 |
| Homens Pele Clara | 0.3 | 0.0 | 0.9 |

*Fonte: Baseado nos dados de precisão apresentados em.30 As taxas de erro são calculadas como (100 - Precisão).*

Estes resultados quantitativos demonstram claramente a existência de vieses significativos de gênero e tipo de pele, e um forte viés interseccional contra mulheres de pele escura nos sistemas avaliados na época.

#### **2.1.7 Limitações e Críticas**

O benchmark Gender Shades e o estudo associado possuem algumas limitações e foram alvo de críticas:

* **Representação Binária de Gênero:** A avaliação utilizou rótulos de gênero binários (feminino/masculino) impostos pelas APIs comerciais, não refletindo a diversidade e complexidade das identidades de gênero.30
* **Uso da Escala Fitzpatrick:** Embora seja um padrão dermatológico, a escala Fitzpatrick é um sistema de classificação que pode não abranger toda a variação contínua dos tons de pele humanos. Além disso, é um proxy fenotípico e não uma medida direta de raça ou etnia.30
* **Escopo Limitado:** O estudo focou apenas na tarefa de classificação de gênero, avaliou somente três APIs comerciais específicas num determinado momento (2017-2018) e utilizou um conjunto de dados fixo (PPB).30 Os resultados podem não ser generalizáveis para outras tarefas de análise facial, outros modelos (incluindo versões mais recentes) ou outros conjuntos de dados.
* **Críticas Metodológicas Externas:** Algumas críticas apontaram que o estudo avaliou a "classificação de gênero" e não o "reconhecimento facial" (identificação 1:N), e que os algoritmos testados eram versões mais antigas.63 No entanto, o estudo foi explícito sobre seu foco na classificação de gênero como um exemplo motivador. A evolução da tecnologia é uma limitação inerente a qualquer benchmark pontual.
* **Foco na Precisão:** O benchmark mede principalmente disparidades nas taxas de erro. Embora crucial, a precisão não é a única preocupação ética; tecnologias de análise facial, mesmo que precisas, podem ser usadas de forma abusiva.29
* **Falta de Transparência da Indústria:** Uma limitação contextual importante foi a falta de divulgação por parte das empresas sobre o desempenho desagregado de seus sistemas, dificultando auditorias externas e a compreensão completa das capacidades e vieses dos modelos.28

#### **2.1.8 Fontes Principais**

* Website do Projeto Gender Shades (inclui Overview, FAQ, Results): 28
* Artigo Acadêmico (Buolamwini & Gebru, 2018): 14
* Outras Fontes Relevantes: 7

### **2.2 COMPAS**

#### **2.2.1 Objetivo Principal**

O COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) é um algoritmo comercial desenvolvido pela Northpointe (agora Equivant) e utilizado em vários sistemas judiciais dos EUA para avaliar o risco de um réu criminal cometer um novo crime (reincidência).2 As pontuações geradas pelo COMPAS visam auxiliar juízes e oficiais de liberdade condicional em decisões como fiança, sentença e liberdade condicional.67

A análise realizada pela organização de jornalismo investigativo ProPublica em 2016 teve como objetivo principal **avaliar se o algoritmo COMPAS, apesar de seu uso generalizado, exibia viés racial**, especificamente se discriminava réus negros em comparação com réus brancos.33 A investigação buscava determinar se o software, que pretendia fornecer uma medida objetiva de risco, na verdade produzia resultados sistematicamente diferentes e potencialmente injustos para diferentes grupos raciais.

#### **2.2.2 Metodologia de Avaliação (Análise da ProPublica)**

A ProPublica empregou uma metodologia quantitativa rigorosa para analisar o desempenho e o viés do COMPAS 34:

* **Aquisição e Preparação de Dados:** Foram obtidos dados de scores do COMPAS para mais de 10.000 réus avaliados em 2013 e 2014 no Condado de Broward, Flórida. Esses scores foram então cruzados com registros criminais públicos do mesmo condado para rastrear a reincidência (novas prisões) nos dois anos seguintes à avaliação.33 Os dados foram filtrados para focar em réus avaliados na fase pré-julgamento e para remover entradas com possíveis problemas de qualidade.34
* **Definição de Reincidência:** A reincidência foi definida como uma nova prisão por qualquer crime dentro de dois anos após a avaliação do COMPAS, alinhando-se com a metodologia usada pela Northpointe em seus próprios estudos de validação.33 A reincidência violenta também foi analisada separadamente.
* **Análise Estatística Comparativa:** A ProPublica comparou as previsões de risco do COMPAS (categorizadas como Baixo, Médio, Alto com base nos scores decílicos de 1 a 10) com as taxas reais de reincidência observadas nos dois anos seguintes.34
* **Análise de Viés Focada em Erros:** A investigação concentrou-se nas disparidades das taxas de erro entre réus negros e brancos. Especificamente, analisaram as taxas de falsos positivos (previstos como de alto risco, mas não reincidiram) e falsos negativos (previstos como de baixo risco, mas reincidiram) para cada grupo racial.33
* **Controle de Variáveis:** Para isolar o efeito da raça, a ProPublica utilizou modelos de regressão logística. Esses modelos controlaram variáveis como número de crimes anteriores, idade, gênero e se o réu realmente reincidiu, permitindo estimar a probabilidade de um réu negro receber um score de risco mais alto em comparação com um réu branco com características semelhantes.33 Modelos de riscos proporcionais de Cox e análise de sobrevivência de Kaplan-Meier também foram usados para avaliar a precisão preditiva geral e visualizar as taxas de reincidência.34

#### **2.2.3 Conjunto de Dados (Análise da ProPublica)**

O conjunto de dados central para a análise da ProPublica foi compilado a partir de duas fontes principais no Condado de Broward, Flórida 33:

* **Scores COMPAS:** Obtidos do Gabinete do Xerife do Condado de Broward para réus avaliados em 2013 e 2014. Incluíam scores decílicos (1-10) para "Risco de Reincidência", "Risco de Violência" e "Risco de Não Comparecimento".33
* **Registros Criminais:** Coletados do site do Cartório do Condado de Broward até abril de 2016. Incluíam informações sobre prisões anteriores e subsequentes (até dois anos após o score COMPAS), grau da acusação, etc..33
* **Dados Demográficos:** Informações como raça, idade e gênero foram obtidas dos registros do xerife.33

O conjunto de dados final usado na análise principal continha informações sobre mais de 7.000 indivíduos.68 Versões processadas e limpas deste conjunto de dados, frequentemente contendo cerca de 6.172 observações e 14 variáveis (após remoção de identificadores, tratamento de outliers e erros, e fatorização de variáveis categóricas), foram disponibilizadas publicamente (por exemplo, no Kaggle e no pacote mlr3fairness) e são comumente usadas em pesquisas sobre justiça algorítmica.76 Variáveis chave incluem idade (age), grau da acusação (c\_charge\_degree), raça (race), sexo (sex), contagem de antecedentes (priors\_count), score decílico (decile\_score) e o indicador de reincidência em dois anos (two\_year\_recid).76

#### **2.2.4 Tipos de Viés Medidos**

A análise da ProPublica concentrou-se principalmente no **viés racial**, investigando disparidades sistemáticas nos resultados da avaliação de risco do COMPAS entre réus negros e brancos.33 Especificamente, examinou:

* **Viés nas Taxas de Erro:** Se os tipos de erros cometidos pelo algoritmo (classificar erroneamente alguém como de alto risco ou como de baixo risco) eram distribuídos de forma desigual entre os grupos raciais.33 Isso se conecta a noções de justiça como igualdade de taxas de falsos positivos e falsos negativos.

#### **2.2.5 Métricas de Avaliação Chave (Análise da ProPublica)**

As métricas chave utilizadas pela ProPublica para avaliar o desempenho e o viés do COMPAS incluíram:

* **Precisão Preditiva (Accuracy):** A porcentagem geral de previsões corretas (reincidente vs. não reincidente).34
* **Taxa de Falsos Positivos (FPR - False Positive Rate):** A proporção de réus previstos como de alto risco que, na verdade, *não* reincidiram nos dois anos seguintes. A ProPublica comparou a FPR entre réus negros e brancos.34
* **Taxa de Falsos Negativos (FNR - False Negative Rate):** A proporção de réus previstos como de baixo risco que, na verdade, *reincidiram* nos dois anos seguintes. A ProPublica comparou a FNR entre réus negros e brancos.34
* **Probabilidade de Score Mais Alto (Controlando Fatores):** Usando regressão logística, calcularam a razão de chances (odds ratio) que indica a probabilidade aumentada de réus negros receberem um score de risco mais alto em comparação com réus brancos, mesmo após controlar por histórico criminal, reincidência real, idade e gênero.33

Embora a análise da ProPublica tenha se concentrado principalmente em FPR e FNR, trabalhos subsequentes sobre o COMPAS frequentemente discutem outras métricas de justiça, como **Paridade Preditiva** (garantir que o score de risco signifique a mesma probabilidade de reincidência para todos os grupos), **Paridade Demográfica** (garantir que a proporção de indivíduos classificados como de alto risco seja a mesma entre os grupos) e **Igualdade de Oportunidade/Probabilidades Equalizadas** (relacionadas à igualdade das taxas de verdadeiro positivo e/ou falso positivo entre os grupos).35 Uma constatação importante decorrente do debate sobre o COMPAS é que é matematicamente impossível satisfazer simultaneamente todas essas definições de justiça quando as taxas base de reincidência diferem entre os grupos.71 Isso implica que qualquer algoritmo de avaliação de risco envolverá inevitavelmente um trade-off entre diferentes noções de justiça.

#### **2.2.6 Modelos Avaliados e Resultados (Análise da ProPublica)**

O "modelo" primário avaliado pela ProPublica foi o próprio **algoritmo COMPAS**, tratado como uma caixa preta, uma vez que sua fórmula exata é proprietária.2 É importante notar que o conjunto de dados COMPAS subsequentemente se tornou um benchmark popular para avaliar a justiça de *outros* modelos de aprendizado de máquina (como Regressão Logística, SVMs, Random Forests, etc.) treinados nesses dados.2

**Principais Resultados da Análise da ProPublica:**

* A precisão geral do COMPAS na previsão de reincidência (qualquer crime) em dois anos foi de aproximadamente 61%.34 A precisão na previsão de reincidência *violenta* foi muito menor, em torno de 20%.34
* A precisão preditiva geral foi semelhante para réus negros (63%) e brancos (59%).34
* No entanto, houve disparidades raciais significativas nos tipos de erros cometidos:
  + **Falsos Positivos:** Réus negros que não reincidiram foram quase duas vezes mais propensos a serem rotulados como de alto risco em comparação com réus brancos que não reincidiram (FPR de 45% para negros vs. 23% para brancos).33
  + **Falsos Negativos:** Réus brancos que reincidiram foram quase duas vezes mais propensos a serem rotulados como de baixo risco em comparação com réus negros que reincidiram (FNR de 48% para brancos vs. 28% para negros).33
* Mesmo controlando por histórico criminal, reincidência futura, idade e gênero, réus negros ainda tinham 45% mais probabilidade de receber scores de risco mais altos para reincidência geral e 77% mais probabilidade para reincidência violenta.33

A tabela a seguir resume as principais taxas de erro díspares encontradas pela ProPublica:

| **Grupo** | **Métrica** | **Valor (%)** |
| --- | --- | --- |
| Réus Negros | FPR (Reincidência Geral) | 45 |
| Réus Brancos | FPR (Reincidência Geral) | 23 |
| Réus Negros | FNR (Reincidência Geral) | 28 |
| Réus Brancos | FNR (Reincidência Geral) | 48 |

*Fonte: Baseado nos achados reportados em.33 FPR = Taxa de Falsos Positivos, FNR = Taxa de Falsos Negativos.*

Estes resultados sugerem que, embora o COMPAS possa ter uma precisão preditiva semelhante entre os grupos (satisfazendo a paridade preditiva), ele o faz à custa de cometer tipos diferentes de erros para réus negros e brancos, falhando em satisfazer critérios de justiça baseados na igualdade das taxas de erro.

#### **2.2.7 Limitações e Críticas**

Tanto a análise da ProPublica quanto o próprio conjunto de dados COMPAS enfrentaram críticas e têm limitações reconhecidas:

* **Críticas à Metodologia da ProPublica:**
  + **Erro de Processamento de Dados:** A crítica mais substancial é que a ProPublica pode ter processado incorretamente os dados para a análise de reincidência de dois anos, mantendo desproporcionalmente mais reincidentes do que deveria, inflando artificialmente a taxa de reincidência geral no conjunto de dados analisado.81 Embora isso afete métricas como PPV e NPV, argumenta-se que teve pouco impacto nas taxas de erro (FPR, FNR) que foram centrais para as conclusões de viés da ProPublica.81
  + **Foco Seletivo em Métricas de Justiça:** Críticos argumentam que a ProPublica focou em disparidades de FPR/FNR, enquanto o COMPAS pode ter sido calibrado para satisfazer a paridade preditiva (onde um determinado score de risco corresponde à mesma probabilidade de reincidência para todos os grupos).69 A impossibilidade matemática de satisfazer ambas as métricas simultaneamente quando as taxas base diferem torna a escolha da métrica de "justiça" um ponto crucial de debate.71
  + **Contexto de Aplicação e Variáveis Ocultas:** A análise foi baseada em dados pré-julgamento, e a ProPublica não teve acesso à fórmula exata do COMPAS nem a todas as variáveis que ele utiliza (como respostas a questionários), levando a possíveis imprecisões na reconstrução ou controle de variáveis.66 A suposição de linearidade da idade também foi questionada.91
* **Limitações do Conjunto de Dados COMPAS para Benchmarking:**
  + **Não Recomendado para Novos Benchmarks:** Pesquisadores de justiça algorítmica geralmente desaconselham o uso do COMPAS para desenvolver ou validar *novos* algoritmos de mitigação de viés. Isso se deve a problemas inerentes ao próprio conjunto de dados, como **viés de medição díspar** (onde a própria variável de resultado ou características podem ser medidas de forma diferente entre grupos) e uma **conexão fraca entre a tarefa de predição definida (reincidência baseada em prisão) e tarefas de avaliação de risco do mundo real**.76 Embora historicamente importante para iniciar o debate sobre viés algorítmico, suas falhas intrínsecas podem levar a conclusões enganosas sobre a eficácia de novas técnicas de fairness.
  + **Qualidade dos Dados:** Existem preocupações sobre a qualidade dos dados subjacentes, incluindo possíveis erros de entrada, integração e dados faltantes, que podem afetar a confiabilidade das pontuações.67
  + **Proxy para Reincidência:** O benchmark usa a re-prisão como um proxy para a reincidência, o que não captura necessariamente todos os casos de reincidência (crimes cometidos sem prisão) nem distingue a gravidade dos novos delitos.81
* **Natureza Proprietária do COMPAS:** O fato de o algoritmo COMPAS ser um produto comercial de caixa preta impede a transparência total e a auditoria independente de sua metodologia e cálculos.66

#### **2.2.8 Fontes Principais**

* Artigo Metodológico da ProPublica: 33
* Artigo Principal da ProPublica ("Machine Bias"): 33
* Repositório de Dados/Código da ProPublica: 33
* Críticas e Reanálises: 69
* Descrições do Conjunto de Dados COMPAS: 2

### **2.3 Civil Comments / Jigsaw Unintended Bias**

#### **2.3.1 Objetivo Principal**

O conjunto de dados Civil Comments, frequentemente associado à competição "Jigsaw Unintended Bias in Toxicity Classification" do Kaggle, foi criado com o objetivo principal de facilitar a pesquisa sobre a **civilidade em conversas online** e, mais especificamente, para estudar e mitigar o **viés não intencional (unintended bias)** em modelos de classificação de toxicidade.96 A plataforma Civil Comments disponibilizou seu arquivo de ~2 milhões de comentários públicos após seu encerramento, e a Jigsaw (uma unidade do Google) patrocinou anotações adicionais para atributos de toxicidade e menções de identidade.96

O foco no "viés não intencional" surge do problema observado em que modelos de toxicidade, treinados em dados onde certos termos de identidade (por exemplo, "negro", "gay", "muçulmano", "mulher") aparecem com mais frequência em comentários tóxicos, aprendem a associar esses próprios termos à toxicidade. Consequentemente, esses modelos podem classificar erroneamente comentários perfeitamente inócuos ou até positivos que mencionam esses grupos de identidade como sendo tóxicos.18 O benchmark visa fornecer dados e métricas para avaliar e reduzir esse tipo específico de viés, promovendo modelos de moderação de conteúdo mais justos e precisos.

#### **2.3.2 Metodologia de Avaliação**

A avaliação de modelos usando o benchmark Civil Comments/Jigsaw envolve várias etapas:

* **Anotação de Dados:** Comentários da plataforma Civil Comments foram anotados por múltiplos avaliadores humanos (via Figure Eight/Appen) para determinar o nível de toxicidade e a presença de vários subtipos de toxicidade (como severe\_toxicity, obscene, threat, insult, identity\_attack, sexual\_explicit).18 A pontuação final de toxicidade (target) para cada comentário é um valor fracionário (0-1) representando a proporção de avaliadores que o consideraram tóxico.18 Para fins de classificação binária, um limiar comum é target >= 0.5.105
* **Anotação de Identidade:** Um subconjunto dos comentários também foi anotado para identificar quais grupos de identidade (raça, gênero, religião, orientação sexual, etc.) foram mencionados no texto.18
* **Métricas de Desempenho Padrão:** O desempenho geral na classificação de toxicidade é medido usando métricas padrão como AUROC (Area Under the Receiver Operating Characteristic curve) e F1-score (Macro e Micro).102
* **Métricas de Viés Específicas:** Para quantificar o viés não intencional, são usadas métricas especializadas baseadas na AUC, calculadas em subconjuntos específicos dos dados definidos pela presença ou ausência de menções a identidades e pelo rótulo de toxicidade real.96 As três principais métricas de viés AUC são:
  + **Subgroup AUC:** Avalia o desempenho do modelo *dentro* de comentários que mencionam uma identidade específica.
  + **BPSN (Background Positive, Subgroup Negative) AUC:** Avalia a capacidade do modelo de distinguir comentários *não tóxicos* que mencionam uma identidade de comentários *tóxicos* que *não* a mencionam.
  + **BNSP (Background Negative, Subgroup Positive) AUC:** Avalia a capacidade do modelo de distinguir comentários *tóxicos* que mencionam uma identidade de comentários *não tóxicos* que *não* a mencionam.
* **Agregação de Métricas de Viés:** As métricas de viés AUC por identidade são agregadas usando uma média generalizada (Generalized Mean of Bias AUCs - GMB), frequentemente com uma potência negativa (p=-5) para dar mais peso às identidades onde o modelo tem pior desempenho.100
* **Métrica Final Combinada:** A métrica final usada na competição Kaggle combina a AUROC geral com a média generalizada das três métricas de viés AUC (Subgroup, BPSN, BNSP), geralmente com pesos iguais.100

#### **2.3.3 Conjunto de Dados**

O conjunto de dados Civil Comments / Jigsaw Unintended Bias é caracterizado por:

* **Origem:** Comentários públicos da plataforma Civil Comments, coletados entre 2015 e 2017 em aproximadamente 50 sites de notícias em inglês.96
* **Tamanho:** O conjunto de dados original contém cerca de 2 milhões de comentários. A versão usada na competição Kaggle tem aproximadamente 1,8 milhão de exemplos de treinamento.96 Um conjunto de dados separado com anotações individuais de toxicidade contém quase 16 milhões de avaliações.18
* **Conteúdo e Rótulos:** Cada entrada contém o texto do comentário (comment\_text) e rótulos fracionários (0-1) para toxicidade (target) e vários subtipos (severe\_toxicity, obscene, threat, insult, identity\_attack, sexual\_explicit). Um subconjunto também possui rótulos fracionários indicando a menção de diversas identidades (gênero, raça/etnia, religião, orientação sexual, deficiência, etc.).96 Metadados adicionais da plataforma Civil Comments e do processo de anotação também estão disponíveis.105
* **Processo de Anotação:** Realizado via crowdsourcing (Figure Eight/Appen). Cada comentário foi avaliado por múltiplos anotadores (geralmente 10 ou mais) para capturar a subjetividade da toxicidade e garantir a qualidade através de perguntas de teste.18

#### **2.3.4 Tipos de Viés Medidos**

O principal tipo de viés medido por este benchmark é o **viés não intencional relacionado à identidade**. Isso se refere à tendência dos modelos de classificação de toxicidade associarem erroneamente termos de identidade (palavras ou frases que se referem a grupos demográficos específicos como raça, gênero, religião, orientação sexual, etc.) com toxicidade, mesmo quando o comentário que menciona a identidade não é tóxico.96 Este viés surge frequentemente porque os termos de identidade podem aparecer com mais frequência em comentários genuinamente tóxicos nos dados de treinamento.99

O benchmark quantifica esse viés medindo as **disparidades de desempenho do modelo em relação a diferentes grupos de identidade**, utilizando as métricas Subgroup AUC, BPSN AUC e BNSP AUC.96

Além disso, a disponibilidade dos dados de anotação individuais permite a investigação do **viés do anotador**, ou seja, como as características ou perspectivas dos próprios avaliadores humanos podem influenciar os rótulos de toxicidade atribuídos.18 A natureza subjetiva da toxicidade significa que o que um anotador considera tóxico pode diferir de outro, e a agregação dessas avaliações pode mascarar essas nuances ou privilegiar a visão da maioria.18

#### **2.3.5 Métricas de Avaliação Chave**

As métricas de avaliação chave para o benchmark Civil Comments / Jigsaw Unintended Bias são:

* **AUROC Geral (Overall AUC):** Mede o desempenho geral do modelo na tarefa de classificação de toxicidade em todo o conjunto de teste, usando a Área sob a Curva ROC.100
* **Métricas de Viés AUC (Bias AUCs):** Projetadas especificamente para medir o viés não intencional em relação a grupos de identidade 96:
  + **Subgroup AUC:** Mede a capacidade do modelo de distinguir entre comentários tóxicos e não tóxicos *dentro* do subconjunto de comentários que mencionam uma identidade específica. Um valor baixo indica dificuldade em fazer essa distinção para aquele grupo.
  + **BPSN AUC (Background Positive, Subgroup Negative):** Mede a capacidade do modelo de distinguir comentários *não tóxicos* que mencionam a identidade de comentários *tóxicos* que *não* a mencionam. Um valor baixo sugere que o modelo confunde comentários inócuos sobre o grupo com toxicidade geral, indicando um viés de falso positivo para menções da identidade.
  + **BNSP AUC (Background Negative, Subgroup Positive):** Mede a capacidade do modelo de distinguir comentários *tóxicos* que mencionam a identidade de comentários *não tóxicos* que *não* a mencionam. Um valor baixo sugere que o modelo confunde toxicidade direcionada ao grupo com comentários inócuos gerais, indicando um viés de falso negativo para menções da identidade.
* **Média Generalizada das Métricas de Viés AUC (GMB - Generalized Mean of Bias AUCs):** Agrega as pontuações de Bias AUC por identidade (Subgroup, BPSN, BNSP) em uma única métrica geral para cada tipo de Bias AUC. Usa uma média de potência (com p=-5 na competição Kaggle) para dar mais peso às identidades com pior desempenho.100 Os resultados são frequentemente reportados como GMB-Subgroup, GMB-BPSN e GMB-BNSP.
* **Métrica Final da Competição:** Uma pontuação combinada que pondera igualmente a AUROC Geral e as três médias generalizadas das métricas de viés AUC (GMB-Subgroup, GMB-BPSN, GMB-BNSP).100
* **Outras Métricas de Desempenho:** Em algumas análises, métricas como Macro F1, Micro F1, Precisão e Recall também são reportadas para avaliar o desempenho da classificação.102

#### **2.3.6 Modelos Avaliados e Resultados**

Uma ampla gama de modelos foi avaliada usando o benchmark Civil Comments, com foco particular em modelos baseados em Transformer.

* **Modelos Avaliados:** Exemplos incluem variantes do BERT (ALBERT, BERTweet, HateBERT, DistilBERT, RoBERTa com diferentes funções de perda como Binary Cross-Entropy - BCE ou Focal Loss - FL), XLNet, XLM RoBERTa, modelos RNN (BiLSTM, BiGRU), modelos CNN (ResNet com embeddings GloVe), Compact Convolutional Transformers (CCT) e modelos de machine learning tradicionais como Regressão Logística.50 Modelos mais recentes como GPT-4, Llama, Claude e Gemini também foram mencionados em contextos relacionados à toxicidade e viés, mas resultados quantitativos específicos neste benchmark não foram encontrados nos trechos fornecidos.22
* **Resultados:**
  + **Desempenho Geral:** Modelos baseados em Transformer, como as variantes do RoBERTa e BERT, geralmente alcançam as maiores pontuações de AUROC Geral, frequentemente superiores a 0.97 ou 0.98.102 RoBERTa com Focal Loss obteve a maior AUROC (0.9818) em um estudo comparativo.102
  + **Viés Não Intencional:** Apesar do alto desempenho geral, a maioria dos modelos ainda exibe viés não intencional significativo. As métricas GMB BPSN e GMB Subgroup frequentemente ficam abaixo de 0.90, indicando que os modelos lutam para lidar corretamente com menções de identidade.102 O RoBERTa com Focal Loss também obteve o melhor GMB BPSN (0.901) no estudo citado.102
  + **Disparidades entre Grupos:** O viés tende a ser mais pronunciado para certos grupos de identidade, como negros, homossexuais, muçulmanos e brancos, conforme indicado por pontuações mais baixas de BPSN AUC e Subgroup AUC para esses grupos.108
  + **Trade-offs:** Modelos RNN (BiLSTM, BiGRU) oferecem tempos de inferência muito mais rápidos, mas com desempenho ligeiramente inferior em AUROC e métricas de viés em comparação com os modelos BERT.108 DistilBERT representa um bom equilíbrio entre desempenho e velocidade.108

A tabela a seguir resume os resultados de alguns modelos chave avaliados no artigo de Duchene et al. (2023) 102:

| **Modelo** | **AUROC Geral (↑)** | **GMB BPSN (↑)** | **GMB Subgroup (↑)** | **GMB BNSP (↑)** |
| --- | --- | --- | --- | --- |
| RoBERTa Focal Loss | 0.9818 | 0.9010 | 0.8807 | 0.9581 |
| ALBERT | 0.9790 | 0.8998 | 0.8734 | 0.9499 |
| BERTweet | 0.9816 | 0.8945 | 0.8780 | 0.9603 |
| HateBERT | 0.9791 | 0.8915 | 0.8744 | 0.9589 |
| RoBERTa BCE | 0.9813 | 0.8901 | 0.8800 | 0.9616 |
| XLM RoBERTa | 0.9790 | 0.8859 | 0.8689 | N/A |
| XLNet | 0.9800 | 0.8834 | 0.8738 | 0.9597 |
| DistilBERT | 0.9804 | 0.8740 | 0.8762 | 0.9644 |
| BiLSTM | 0.9754 | 0.8758 | 0.8636 | N/A |

*Fonte: Adaptado de Duchene et al. (2023).102 GMB BPSN, GMB Subgroup, GMB BNSP referem-se à Média Generalizada das respectivas métricas de viés AUC.*

#### **2.3.7 Limitações e Críticas**

O benchmark Civil Comments/Jigsaw apresenta algumas limitações:

* **Desbalanceamento de Dados:** O conjunto de dados é inerentemente desbalanceado, tanto em termos de toxicidade geral versus não toxicidade, quanto na frequência de menções a diferentes identidades ou subtipos de toxicidade.101 Isso pode dificultar o treinamento de modelos robustos e a avaliação justa do desempenho em classes minoritárias.
* **Subjetividade da Toxicidade e Agregação de Rótulos:** A definição e percepção de toxicidade são subjetivas. O benchmark principal depende de rótulos agregados (média ou fração de anotações), o que mascara a discordância entre anotadores e pressupõe uma "verdade fundamental" única, o que pode ser problemático para tarefas sensíveis e subjetivas.18 A disponibilidade dos rótulos individuais 18 mitiga parcialmente isso para pesquisas futuras, mas as métricas padrão do benchmark usam os rótulos agregados.
* **Uso de Rótulos Sintéticos em Alguns Estudos:** Algumas pesquisas que investigam o impacto de dados ruidosos usam rótulos gerados sinteticamente (por exemplo, previsões de um modelo treinado) para simular viés, o que pode não refletir com precisão as nuances do viés humano real na anotação.101
* **Potencial de Contaminação do Benchmark:** Como em qualquer benchmark público amplamente utilizado, existe o risco de que os dados de teste ou exemplos semelhantes sejam incluídos inadvertidamente nos dados de treinamento de modelos futuros, inflando artificialmente as pontuações de desempenho.120

#### **2.3.8 Fontes Principais**

* Descrição do Dataset (PapersWithCode, Kaggle): 96
* Artigo de Benchmark (Duchene et al., 2023): 102
* Artigo Original sobre Métricas de Viés (Borkan et al., 2019): 96 (Referência implícita frequente)
* Estudos Relacionados: 99

### **2.4 BOLD (Bias in Open-Ended Language Generation)**

#### **2.4.1 Objetivo Principal**

O benchmark BOLD (Bias in Open-Ended Language Generation Dataset) foi desenvolvido com o objetivo central de **estudar e avaliar sistematicamente os vieses sociais na geração de texto de formato aberto** por modelos de linguagem (LMs).3 Diferentemente de benchmarks focados em tarefas de classificação ou preenchimento de lacunas, o BOLD visa analisar os vieses que emergem quando os LMs geram continuações de texto a partir de um prompt inicial, uma tarefa mais próxima de muitas aplicações do mundo real, como chatbots e assistentes de escrita.3

Para isso, o BOLD introduz duas contribuições chave:

1. Um **conjunto de dados em larga escala (BOLD dataset)** com 23.679 prompts únicos em inglês, derivados de textos naturais da Wikipédia, abrangendo cinco domínios demográficos.3
2. Um conjunto de **métricas automatizadas, incluindo novas métricas** para toxicidade, normas psicolinguísticas e polaridade de gênero, além de métricas existentes como sentimento e consideração (regard), para medir o viés no texto gerado a partir de múltiplas perspectivas.3

O objetivo final é fornecer uma ferramenta para comparar a justiça (fairness) de diferentes LMs na geração de texto aberto e alertar sobre os preconceitos embutidos que podem ser propagados por esses modelos.3

#### **2.4.2 Metodologia de Avaliação**

A metodologia de avaliação do BOLD segue os seguintes passos:

* **Seleção de Prompts:** Utiliza prompts do conjunto de dados BOLD. Esses prompts são sequências curtas de palavras (6 a 9 palavras) extraídas do início de frases da Wikipédia que mencionam um grupo demográfico específico (por exemplo, uma profissão, um grupo racial, uma afiliação religiosa).3 A intenção é usar inícios de frases naturais como gatilhos para a geração de texto.
* **Geração de Texto:** Os prompts são fornecidos como entrada para os modelos de linguagem a serem avaliados. Os LMs então geram continuações de texto de formato aberto de maneira autorregressiva.3
* **Cálculo de Métricas:** O texto gerado para cada prompt é analisado usando um conjunto de métricas automatizadas propostas e validadas no estudo. Estas incluem Sentimento (VADER), Toxicidade (classificador BERT), Consideração/Regard (classificador BERT), Normas Psicolinguísticas (baseadas em léxicos VAD e BE5) e Polaridade de Gênero (baseada em contagem de unigramas e projeção de embeddings).3
* **Análise Comparativa:** As pontuações das métricas são comparadas entre diferentes subgrupos demográficos dentro de cada um dos cinco domínios (profissão, gênero, raça, religião, ideologia política) para identificar disparidades e vieses.3 O desempenho dos LMs também é comparado com um baseline: o viés medido no texto original da Wikipédia de onde os prompts foram extraídos, usando as mesmas métricas.3
* **Validação Humana:** Para garantir a relevância das métricas automatizadas, os autores coletaram anotações humanas (via Amazon Mechanical Turk) para sentimento, toxicidade e polaridade de gênero em um subconjunto dos textos gerados e calcularam a correlação entre as métricas automáticas e os julgamentos humanos.3

#### **2.4.3 Conjunto de Dados: BOLD**

O conjunto de dados BOLD é a peça central do benchmark:

* **Tamanho e Escopo:** Contém 23.679 prompts únicos de geração de texto em inglês.3 Foi apresentado como o maior benchmark de fairness para geração de linguagem aberta na época.3
* **Origem dos Prompts:** Os prompts foram extraídos de frases de artigos da Wikipédia em inglês.3
* **Domínios Cobertos:** Avalia vieses em cinco domínios: profissão, gênero, raça, religião e ideologia política, abrangendo 43 subgrupos distintos.1
* **Construção dos Prompts:** Cada prompt é formado pelas cinco primeiras palavras de uma frase da Wikipédia onde o atributo do grupo (por exemplo, "engenheiro", "mulher", "asiático", "muçulmano", "conservadorismo") é mencionado nas primeiras oito palavras, seguido pelo próprio termo do atributo. Isso resulta em prompts de 6 a 9 palavras.3
* **Processamento:** Os textos da Wikipédia passaram por limpeza e filtragem. Durante a avaliação, os prompts foram anonimizados (nomes substituídos por "Person", nomes de grupos por "XYZ") para evitar que vieses nas próprias ferramentas de avaliação ou nos anotadores humanos influenciassem os resultados baseados em identidades específicas.3

#### **2.4.4 Tipos de Viés Medidos**

O BOLD visa medir vieses sociais que se manifestam na geração de texto aberto. O viés é entendido como a geração desproporcional de texto percebido como negativo, injusto, preconceituoso ou estereotipado contra uma ideia ou grupo de pessoas.3 Os vieses são avaliados em relação aos cinco domínios cobertos pelo conjunto de dados:

* **Profissão**
* **Gênero** (com foco em binário: masculino/feminino)
* **Raça** (com foco em grupos relevantes no contexto dos EUA)
* **Religião**
* **Ideologia Política**

O viés é inferido através da análise das propriedades do texto gerado, utilizando as métricas de sentimento, toxicidade, consideração (regard), normas psicolinguísticas e polaridade de gênero.3

#### **2.4.5 Métricas de Avaliação Chave**

As métricas chave propostas e utilizadas pelo BOLD são 3:

* **Sentimento (VADER):** Avalia a polaridade geral (positiva/negativa) do texto gerado.
* **Toxicidade (BERT fine-tuned):** Classifica o texto em categorias como tóxico, insulto, ameaça, etc.
* **Consideração/Regard (BERT fine-tuned):** Mede a polaridade (positiva/negativa) do texto especificamente em relação a grupos demográficos (gênero e raça, neste estudo).
* **Normas Psicolinguísticas (VAD, BE5):** Analisa a carga emocional do texto em dimensões como Valência, Ativação, Dominância e emoções básicas (Alegria, Raiva, Tristeza, Medo, Nojo) [3

#### Referências citadas

1. debiaSAE: Benchmarking and Mitigating Vision-Language Model Bias - arXiv, acessado em abril 24, 2025, <https://arxiv.org/html/2410.13146v2>
2. Chapter 10 Story COMPAS: recidivism reloaded | XAI Stories, acessado em abril 24, 2025, <https://pbiecek.github.io/xai_stories/story-compas.html>
3. arxiv.org, acessado em abril 24, 2025, <https://arxiv.org/pdf/2101.11718>
4. 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>
5. 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>
6. 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>
7. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Mode - ACL Anthology, acessado em abril 24, 2025, <https://aclanthology.org/2023.acl-long.656.pdf>
8. BiasDPO: Mitigating Bias in Language Models through Direct Preference Optimization - ACL Anthology, acessado em abril 24, 2025, <https://aclanthology.org/2024.acl-srw.7.pdf>
9. BOLD: Dataset and metrics for measuring biases in open-ended language generation - Amazon Science, acessado em abril 24, 2025, <https://www.amazon.science/publications/bold-dataset-and-metrics-for-measuring-biases-in-open-ended-language-generation>
10. StereoSet: Measuring stereotypical bias in pretrained language models - ACL Anthology, acessado em abril 24, 2025, <https://aclanthology.org/2021.acl-long.416.pdf>
11. VLStereoSet: A Study of Stereotypical Bias in Pre-trained Vision-Language Models - ACL Anthology, acessado em abril 24, 2025, <https://aclanthology.org/2022.aacl-main.40.pdf>
12. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models | Request PDF - ResearchGate, acessado em abril 24, 2025, <https://www.researchgate.net/publication/344447505_CrowS-Pairs_A_Challenge_Dataset_for_Measuring_Social_Biases_in_Masked_Language_Models>
13. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models | Request PDF - ResearchGate, acessado em abril 24, 2025, <https://www.researchgate.net/publication/347235439_CrowS-Pairs_A_Challenge_Dataset_for_Measuring_Social_Biases_in_Masked_Language_Models>
14. Addressing Gender Bias in Facial Recognition Technology: An Urgent Need for Fairness and Inclusion - Cogent Infotech, acessado em abril 24, 2025, <https://www.cogentinfo.com/resources/addressing-gender-bias-in-facial-recognition-technology-an-urgent-need-for-fairness-and-inclusion>
15. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification - Proceedings of Machine Learning Research, acessado em abril 24, 2025, <https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>
16. (PDF) Parity benchmark for measuring bias in LLMs - ResearchGate, acessado em abril 24, 2025, <https://www.researchgate.net/publication/387137967_Parity_benchmark_for_measuring_bias_in_LLMs>
17. peer.asee.org, acessado em abril 24, 2025, <https://peer.asee.org/evaluating-stereotypical-biases-and-implications-for-fairness-in-large-language-models.pdf>
18. StereoSet: Measuring stereotypical bias in pretrained language models | Request PDF - ResearchGate, acessado em abril 24, 2025, <https://www.researchgate.net/publication/353489403_StereoSet_Measuring_stereotypical_bias_in_pretrained_language_models>
19. (PDF) Foundation Models at Work: Fine-Tuning for Fairness in Algorithmic Hiring, acessado em abril 24, 2025, <https://www.researchgate.net/publication/387975199_Foundation_Models_at_Work_Fine-Tuning_for_Fairness_in_Algorithmic_Hiring>
20. [Literature Review] Bias in Language Models: Beyond Trick Tests, acessado em abril 24, 2025, <https://www.themoonlight.io/review/bias-in-language-models-beyond-trick-tests-and-toward-ruted-evaluation>
21. Anthropic Claude 3 - Arize AI, acessado em abril 24, 2025, <https://arize.com/blog/anthropic-claude-3/>
22. Llama 4 Comparison with Claude 3.7 Sonnet, GPT-4.5, and Gemini 2.5 - Bind AI, acessado em abril 24, 2025, <https://blog.getbind.co/2025/04/06/llama-4-comparison-with-claude-3-7-sonnet-gpt-4-5-and-gemini-2-5/>
23. 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>
24. Unveiling AI Bias: Understanding Its Ethical Implications - Robert F. Smith, acessado em abril 22, 2025, <https://robertsmith.com/blog/ai-bias/>
25. Bias in Machine Learning - Lightly.ai, acessado em abril 22, 2025, <https://www.lightly.ai/blog/bias-in-machine-learning>
26. 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>
27. 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.>
28. How 'Gender Shades' Sheds Light on Bias in Machine Learning, acessado em abril 24, 2025, <https://www.dataprivacyadvisory.com/how-gender-shades-sheds-light-on-bias-in-machine-learning/>
29. Overview ‹ Gender Shades - MIT Media Lab, acessado em abril 24, 2025, <https://www.media.mit.edu/projects/gender-shades/overview/>
30. Gender Shades, acessado em abril 24, 2025, <http://gendershades.org/overview.html>
31. Study finds gender and skin-type bias in commercial artificial-intelligence systems, acessado em abril 24, 2025, <https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212>
32. BOLD: Dataset and metrics for measuring biases in open-ended language generation, acessado em abril 24, 2025, <http://web.cs.ucla.edu/~kwchang/bibliography/Dhamala2021bold/>
33. COMPAS Recidivism Racial Bias - Kaggle, acessado em abril 24, 2025, <https://www.kaggle.com/datasets/danofer/compass/data>
34. How We Analyzed the COMPAS Recidivism Algorithm — ProPublica, acessado em abril 24, 2025, <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>
35. Investigating Fairness in Predictive Modeling: A Case Study on COMPAS Recidivism Data - IRJET, acessado em abril 24, 2025, <https://www.irjet.net/archives/V11/i11/IRJET-V11I11111.pdf>
36. FairCompass: Operationalising Fairness in Machine Learning - arXiv, acessado em abril 24, 2025, <https://arxiv.org/pdf/2312.16726>
37. A Survey on Fairness in Large Language Models - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2308.10149v2>
38. Evaluate model and system for safety | Responsible Generative AI Toolkit - Gemini API, acessado em abril 22, 2025, <https://ai.google.dev/responsible/docs/evaluation>
39. Bias and Fairness in Large Language Models: A Survey - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2309.00770v2>
40. Bias Detection | Papers With Code, acessado em abril 22, 2025, [https://paperswithcode.com/task/bias-detection/codeless?page=6&q=](https://paperswithcode.com/task/bias-detection/codeless?page=6&q)
41. Bias Detection | Papers With Code, acessado em abril 22, 2025, [https://paperswithcode.com/task/bias-detection/codeless?page=4&q=](https://paperswithcode.com/task/bias-detection/codeless?page=4&q)
42. Bias Detection | Papers With Code, acessado em abril 24, 2025, <https://paperswithcode.com/task/bias-detection>
43. 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>
44. [2101.11718] BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation - arXiv, acessado em abril 22, 2025, <https://arxiv.org/abs/2101.11718>
45. Mitigating Gender Bias in Large Language Models with Debias Prompting - OpenReview, acessado em abril 24, 2025, <https://openreview.net/pdf/1096e3651906cf975759252a8a72d8368b182b8a.pdf>
46. StereoSet Dataset | Papers With Code, acessado em abril 24, 2025, <https://paperswithcode.com/dataset/stereoset>
47. StereoSet: Measuring stereotypical bias in pretrained language models - ACL Anthology, acessado em abril 24, 2025, <https://aclanthology.org/2021.acl-long.416/>
48. [2010.00133] CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models - arXiv, acessado em abril 24, 2025, <https://arxiv.org/abs/2010.00133>
49. Daily Papers - Hugging Face, acessado em abril 24, 2025, <https://huggingface.co/papers?q=Crows-Pairs>
50. TrustLLM: Trustworthiness in Large Language Models – A principle and benchmark - arXiv, acessado em abril 24, 2025, <https://arxiv.org/html/2401.05561v6>
51. Stereotypical Bias Analysis - Papers With Code, acessado em abril 24, 2025, <https://paperswithcode.com/task/stereotypical-bias-analysis>
52. A Decade's Battle on Dataset Bias: Are We There Yet? - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2403.08632v2>
53. The Impact of Responsible AI Research on Innovation and Development - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2407.15647v4>
54. 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>
55. arXiv:2302.12640v1 [cs.CL] 24 Feb 2023, acessado em abril 22, 2025, <https://arxiv.org/pdf/2302.12640>
56. 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>
57. 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>
58. Frequently Asked Questions ‹ Gender Shades - MIT Media Lab, acessado em abril 22, 2025, <https://www.media.mit.edu/projects/gender-shades/faq/>
59. Results ‹ Gender Shades - MIT Media Lab, acessado em abril 24, 2025, <https://www.media.mit.edu/projects/gender-shades/results/>
60. Gender shades - DSpace@MIT, acessado em abril 22, 2025, <https://dspace.mit.edu/bitstream/handle/1721.1/114068/1026503582-MIT.pdf>
61. Gender Shades, acessado em abril 24, 2025, <https://gs.ajl.org/>
62. Black Panther Face Scorecard - Gender Shades, acessado em abril 24, 2025, <http://gendershades.org/bpscorecard.html>
63. The Myth of Facial Recognition Bias - Clearview AI, acessado em abril 24, 2025, <https://www.clearview.ai/post/the-myth-of-facial-recognition-bias>
64. Gender Shades, acessado em abril 24, 2025, <http://gendershades.org/>
65. Viés Racial em Modelos de Inteligência Artificial para Classificaç ˜ao de Melanomas, acessado em abril 24, 2025, <https://sol.sbc.org.br/index.php/wics/article/download/24833/24654/>
66. HOW TO ARGUE WITH AN ALGORITHM: LESSONS FROM THE COMPAS- PROPUBLICA DEBATE - Colorado Technology Law Journal, acessado em abril 24, 2025, <http://ctlj.colorado.edu/wp-content/uploads/2021/02/17.1_4-Washington_3.18.19.pdf>
67. (PDF) Determining the best feature combination through text and probabilistic feature analysis for GPT-2-based mobile app review detection - ResearchGate, acessado em abril 24, 2025, <https://www.researchgate.net/publication/376832794_Determining_the_best_feature_combination_through_text_and_probabilistic_feature_analysis_for_GPT-2-based_mobile_app_review_detection>
68. Machine Bias — ProPublica, acessado em abril 24, 2025, <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
69. False Positives, False Negatives, and False Analyses: A Rejoinder to “Machine Bias: There's Software Used Across the Country - U.S. Courts, acessado em abril 24, 2025, <https://www.uscourts.gov/sites/default/files/80_2_6_0.pdf>
70. Angwin, J., Larson, J., Mattu, S., Kirchner, L., & ProPublica (2016). Machine Bias Theres Software Used across the Country to Predict Future Criminals. And Its Biased against Blacks. - References - Scientific Research Publishing, acessado em abril 24, 2025, <https://www.scirp.org/reference/referencespapers?referenceid=3771021>
71. Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say - ProPublica, acessado em abril 24, 2025, <https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say>
72. ProPublica Responds to Company's Critique of Machine Bias Story, acessado em abril 24, 2025, <https://www.propublica.org/article/propublica-responds-to-companys-critique-of-machine-bias-story>
73. Algorithms, fairness, and race: Comparing human recidivism risk assessment with the COMPAS, acessado em abril 24, 2025, <http://jdiesnerlab.ischool.illinois.edu/calls/biasescss2018_papers/SocialPanel_ArpitaBiswas_et_al_Algorithms_fairness_race.pdf>
74. Breaking down machine bias, from criminal sentencing by ProPublica - investigate.ai, acessado em abril 24, 2025, <https://investigate.ai/propublica-criminal-sentencing/week-5-1-machine-bias-class/>
75. COMPAS Recidivism Racial Bias - Kaggle, acessado em abril 22, 2025, <https://www.kaggle.com/datasets/danofer/compass>
76. COMPAS Dataset - R, acessado em abril 22, 2025, <https://search.r-project.org/CRAN/refmans/mlr3fairness/help/compas.html>
77. COMPAS Dataset — compas • mlr3fairness, acessado em abril 24, 2025, <https://mlr3fairness.mlr-org.com/reference/compas.html>
78. Honors Capstone Report: Robustness of Fairness in Machine Learning - Deep Blue Repositories, acessado em abril 24, 2025, <https://deepblue.lib.umich.edu/bitstream/handle/2027.42/176722/Honors_Capstone_Fairness_ML_-_Serafina_Kamp.pdf?sequence=1>
79. GHC: U: Robustness of Fairness in Machine Learning - Association for Computing Machinery, acessado em abril 24, 2025, <https://src.acm.org/binaries/content/assets/src/2022/serafina-kamp.pdf>
80. [1911.10640] Algorithmic Bias in Recidivism Prediction: A Causal Perspective - arXiv, acessado em abril 24, 2025, <https://arxiv.org/abs/1911.10640>
81. arxiv.org, acessado em abril 24, 2025, <https://arxiv.org/pdf/1906.04711>
82. 4. COMPAS Recidivism Algorithm — Fairness & Algorithmic Decision Making - Aaron Fraenkel, acessado em abril 24, 2025, <https://afraenkel.github.io/fairness-book/content/04-compas.html>
83. ProPublica's Misleading Machine Bias - by Llewellyn Jones - Investigative Economics, acessado em abril 24, 2025, <https://www.investigativeeconomics.org/p/propublicas-misleading-machine-bias>
84. Machine Bias: Risk Assessments in Criminal Sentencing - Data & Society, acessado em abril 24, 2025, <https://datasociety.net/library/machine-bias-risk-assessments-in-criminal-sentencing/>
85. Analyzing Fairness of Classification Machine Learning Model with Structured Dataset - arXiv, acessado em abril 24, 2025, <https://arxiv.org/html/2412.09896v2>
86. An Empirical Study on Learning Fairness Metrics for COMPAS Data with Human Supervision - Semantic Scholar, acessado em abril 24, 2025, <https://www.semanticscholar.org/paper/An-Empirical-Study-on-Learning-Fairness-Metrics-for-Wang-Grgic-Hlaca/8d4e3ec650e8509f2ff001462f53064a3071c86a>
87. Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods - MDPI, acessado em abril 24, 2025, <https://www.mdpi.com/2504-2289/7/1/15>
88. Present-best results on the COMPAS dataset - humancompatible.org, acessado em abril 24, 2025, <https://humancompatible.org/index.php/2023/06/17/present-best-results-on-the-compas-dataset/>
89. Results for the COMPAS dataset on different fairness constraints being... - ResearchGate, acessado em abril 24, 2025, <https://www.researchgate.net/figure/Results-for-the-COMPAS-dataset-on-different-fairness-constraints-being-incorporated_fig4_364530773>
90. Bayesian Modeling of Intersectional Fairness: The Variance of Bias - NSF Public Access Repository, acessado em abril 24, 2025, <https://par.nsf.gov/servlets/purl/10148349>
91. The Age of Secrecy and Unfairness in Recidivism Prediction - Harvard Data Science Review, acessado em abril 24, 2025, <https://hdsr.mitpress.mit.edu/pub/7z10o269>
92. Response to 'How we analyzed the COMPAS recidivism algorithm' (Larson et al.), acessado em abril 24, 2025, <https://jeffreyfossett.com/2020/02/20/compas-response.html>
93. COMPAS Recidivism — FairLens 0.1.1.dev25+g6b864ec documentation, acessado em abril 24, 2025, <https://fairlens.readthedocs.io/en/latest/user_guide/compas.html>
94. Machine Bias — ProPublica, acessado em abril 24, 2025, <https://www.propublica.org/series/machine-bias>
95. propublica/compas-analysis: Data and analysis for 'Machine Bias' - GitHub, acessado em abril 24, 2025, <https://github.com/propublica/compas-analysis>
96. Civil Comments Dataset - Papers With Code, acessado em abril 24, 2025, <https://paperswithcode.com/dataset/civil-comments>
97. 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>
98. [2004.09456] StereoSet: Measuring stereotypical bias in pretrained language models - arXiv, acessado em abril 24, 2025, <https://arxiv.org/abs/2004.09456>
99. 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>
100. Jigsaw Unintended Bias in Toxicity Classification - Kaggle, acessado em abril 22, 2025, <https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification>
101. An Empirical Investigation of Learning from Biased Toxicity Labels - arXiv, acessado em abril 22, 2025, <https://arxiv.org/pdf/2110.01577>
102. Civil Comments Benchmark (Toxic Comment Classification) - Papers With Code, acessado em abril 24, 2025, <https://paperswithcode.com/sota/toxic-comment-classification-on-civil>
103. unitaryai/detoxify: Trained models & code to predict toxic comments on all 3 Jigsaw Toxic Comment Challenges. Built using ⚡ Pytorch Lightning and 🤗 Transformers. For access to our API, please email us at contact@unitary.ai. - GitHub, acessado em abril 24, 2025, <https://github.com/unitaryai/detoxify>
104. EMNLP 2023 - ACL Anthology, acessado em abril 24, 2025, <https://aclanthology.org/anthology-files/anthology-files/handbooks/emnlp/2023.emnlp.handbook.pdf>
105. Jigsaw Unintended Bias in Toxicity Classification | Kaggle, acessado em abril 22, 2025, <https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data>
106. Toxic Comment Classification and Unintended Bias - CS229, acessado em abril 22, 2025, <https://cs229.stanford.edu/proj2019spr/report/69.pdf>
107. arXiv:2301.11125v1 [cs.CL] 26 Jan 2023, acessado em abril 22, 2025, <https://arxiv.org/pdf/2301.11125>
108. arxiv.org, acessado em abril 24, 2025, <https://arxiv.org/abs/2301.11125>
109. A benchmark for toxic comment classification on Civil Comments dataset | Papers With Code, acessado em abril 24, 2025, <https://paperswithcode.com/paper/a-benchmark-for-toxic-comment-classification>
110. Essays in Economics of Artificial Intelligence - IRIS, acessado em abril 24, 2025, <https://iris.unibocconi.it/retrieve/3db1facd-a52a-4156-9454-a5205feb576f/Revised_Thesis_Habibi_Mahyar.pdf>
111. The 2023 Conference on Empirical Methods in Natural Language Processing, acessado em abril 24, 2025, <https://aclanthology.org/events/emnlp-2023/>
112. We benchmarked GPT-4.1: it's better at code reviews than Claude Sonnet 3.7 - Reddit, acessado em abril 24, 2025, <https://www.reddit.com/r/ChatGPTCoding/comments/1jz5x09/we_benchmarked_gpt41_its_better_at_code_reviews/>
113. LegalBench Benchmark - Vals AI, acessado em abril 24, 2025, <https://www.vals.ai/benchmarks/legal_bench-03-13-2025>
114. GPT-4o Benchmark - Detailed Comparison with Claude & Gemini - Wielded, acessado em abril 24, 2025, <https://wielded.com/blog/gpt-4o-benchmark-detailed-comparison-with-claude-and-gemini>
115. Performance of Advanced Large Language Models (GPT-4o, GPT-4, Gemini 1.5 Pro, Claude 3 Opus) on Japanese Medical Licensing Examination: A Comparative Study | medRxiv, acessado em abril 24, 2025, <https://www.medrxiv.org/content/10.1101/2024.07.09.24310129v1.full-text>
116. How We Tested Leading AI Models Performance on Election Queries - Proof News, acessado em abril 24, 2025, <https://www.proofnews.org/how-we-tested-leading-ai-models-performance-on-election-queries/>
117. Claude 3 Opus Blows Out GPT-4 and Gemini Ultra in a New Benchmark that Requires Reasoning and Accuracy : r/singularity - Reddit, acessado em abril 24, 2025, <https://www.reddit.com/r/singularity/comments/1bzik8g/claude_3_opus_blows_out_gpt4_and_gemini_ultra_in/>
118. Capabilities of Large Language Models in Control Engineering: A Benchmark Study on GPT-4, Claude 3 Opus, and Gemini 1.0 Ultra - arXiv, acessado em abril 24, 2025, <https://arxiv.org/abs/2404.03647>
119. The Claude 3 Model Family: Opus, Sonnet, Haiku - Anthropic, acessado em abril 24, 2025, <https://www.anthropic.com/claude-3-model-card>
120. 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>
121. 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>
122. 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>
123. Multi-task learning for toxic comment classification and rationale extraction - PMC, acessado em abril 22, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC9391651/>
124. Proceedings - dco.uscg.mil, acessado em abril 24, 2025, <https://www.dco.uscg.mil/Portals/9/DCO%20Documents/Proceedings%20Magazine/Archive/1998/Vol55_No4_Oct-Dec1998.pdf?ver=Jsy1TsUPN69s5CpA7IJ7QA%3D%3D>
125. Algal Toxin Risk Assessment and Management Strategic Plan for Drinking Water - Environmental Protection Agency (EPA), acessado em abril 24, 2025, <https://www.epa.gov/sites/default/files/2015-11/documents/algal-risk-assessment-strategic-plan-2015.pdf>
126. Laboratory Test Methods to Assess the Effects of Chemicals on Terrestrial Animal Species - epa nepis, acessado em abril 24, 2025, <https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=9100RYRQ.TXT>
127. 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/>
128. 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>
129. Gemma: Open Models Based on Gemini Research and Technology - arXiv, acessado em abril 24, 2025, <https://arxiv.org/html/2403.08295v1>
130. Aligning to Thousands of Preferences via System Message Generalization - NIPS papers, acessado em abril 24, 2025, <https://proceedings.neurips.cc/paper_files/paper/2024/file/86c9df30129f7663ad4d429b6f80d461-Paper-Conference.pdf>
131. VLBiasBench: A Comprehensive Benchmark for Evaluating Bias in Large Vision-Language Model - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2406.14194v1>
132. VLBiasBench: A Comprehensive Benchmark for Evaluating Bias in Large Vision-Language Model - arXiv, acessado em abril 22, 2025, <https://arxiv.org/html/2406.14194v2>
133. CrowS-Pairs Dataset - Papers With Code, acessado em abril 24, 2025, <https://paperswithcode.com/dataset/crows-pairs>