# Dataset StereoSet para Detecção de Viés

## 1. Visão Geral

O StereoSet é um conjunto de dados de larga escala com cerca de 17.000 sentenças em inglês, criado para quantificar vieses estereotipados em modelos de linguagem pré-treinados. Sua estrutura se baseia no

**Context Association Test (CAT)**, onde cada entrada contém:

* Um **contexto**: uma sentença com um grupo-alvo (target).
* Três **sentenças de atributo**: uma estereotipada, uma anti-estereotipada e uma sem sentido ou não relacionada.

O conjunto de dados avalia dois tipos de tarefas :

* **Intrasentence:** Mede a preferência do modelo em nível de sentença, em uma abordagem de preenchimento de espaços (fill-in-the-blank).
* **Intersentence:** Avalia a capacidade do modelo de continuar uma sentença.

## 2. Domínios de Viés e Exemplos

O StereoSet detecta vieses estereotipados em quatro domínios principais:

* **Gênero**: Vieses sobre homens e mulheres.
* **Raça**: Estereótipos sobre grupos raciais e étnicos, como a associação de pessoas da Etiópia com serem magras e boas em corrida.
* **Religião**: Vieses sobre crenças e práticas religiosas.
* **Profissão**: Estereótipos sobre ocupações. Por exemplo, a associação de músico com dificuldades com drogas.

## 3. Metodologia de Detecção de Viés e Métricas

A avaliação de modelos no StereoSet é feita calculando a probabilidade logarítmica de cada sentença de atributo. Com base nisso, três métricas principais são utilizadas :

#### Tabela 2: Resumo das Métricas de Avaliação do StereoSet

|  |  |  |
| --- | --- | --- |
| Métrica | O que mede | Pontuação Ideal |
| **LMS (Language Modeling Score)** | Capacidade de distinguir texto significativo de texto sem sentido. | 100 |
| **SS (Stereotype Score)** | Propensão a preferir associações estereotipadas sobre as anti-estereotipadas. | 50 |
| **ICAT (Idealized CAT Score)** | Pontuação combinada de proficiência e viés. | 100 |

## 4. Limitações e Considerações

* **Vieses Culturais Específicos**: O dataset foi construído com base em estereótipos do contexto cultural dos Estados Unidos, o que pode limitar sua aplicabilidade a outras culturas.
* **Viés Explícito vs. Implícito**: O StereoSet foca em "formas de viés explícitas, fáceis de ver e relativamente óbvias". Modelos modernos, que passaram por alinhamento de valores, podem passar nesses testes, mas ainda conter vieses sutis e implícitos que o dataset não captura .

#### Referências citadas

1. McGill-NLP/stereoset · Datasets at Hugging Face, acessado em agosto 20, 2025, <https://huggingface.co/datasets/McGill-NLP/stereoset>
2. StereoSet: Measuring stereotypical bias in pretrained language ..., acessado em agosto 20, 2025, <https://arxiv.org/pdf/2004.09456>
3. StereoSet: Measuring stereotypical bias in pretrained language models - ACL Anthology, acessado em agosto 20, 2025, <https://aclanthology.org/2021.acl-long.416.pdf>
4. StereoSet\_Notebook.ipynb - Colab, acessado em agosto 20, 2025, <https://colab.research.google.com/github/JohnSnowLabs/langtest/blob/main/demo/tutorials/task-specific-notebooks/StereoSet_Notebook.ipynb>
5. StereoSet: Combatting Inherently Biased Linguistic Models | by Tanishq Sandhu | Fair Bytes, acessado em agosto 20, 2025, <https://medium.com/fair-bytes/stereoset-combatting-inherently-biased-linguistic-models-a2ecc4c7e0f3>
6. Explicitly unbiased large language models still form biased associations - PNAS, acessado em agosto 20, 2025, <https://www.pnas.org/doi/10.1073/pnas.2416228122>