By: https://vicenteherrera.com

In [ ]:

```
import os
from google.colab import userdata
os.environ['HF_TOKEN'] = userdata.get('HF_TOKEN')
```

What it is:

- CrowS-Pairs is a dataset designed specifically for measuring biases against historically marginalized groups in language models.
- Contains sentence pairs with stereotypical and anti-stereotypical contexts, allowing bias measurement through comparative likelihood.

Dataset Link:

- GitHub: https://github.com/nyu-mll/crows-pairs
- Paper: CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models

In [ ]:

```
%cd /content/
!git clone https://github.com/nyu-mll/crows-pairs.git
%cd crows-pairs
%pip install -r requirements.txt --force

# Example evaluation
!python metric.py --model_name AdamLucek/Orpo-Llama-3.2-1B-40k
```

```
Cloning into 'crows-pairs'...
remote: Enumerating objects: 904, done.
remote: Counting objects: 100% (176/176), done.
remote: Compressing objects: 100% (92/92), done.
remote: Total 904 (delta 81), reused 170 (delta 77), pack-reused 728 (from 1)
Receiving objects: 100% (904/904), 24.59 MiB | 15.70 MiB/s, done.
Resolving deltas: 100% (507/507), done.
/content/crows-pairs/crows-pairs
Collecting pandas (from -r requirements.txt (line 1))
  Downloading pandas-2.2.3-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_6
4.whl.metadata (89 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 89.9/89.9 kB 7.0 MB/s eta 0:00:00
ERROR: Could not find a version that satisfies the requirement torch==1.4.0 (from
versions: 1.13.0, 1.13.1, 2.0.0, 2.0.1, 2.1.0, 2.1.1, 2.1.2, 2.2.0, 2.2.1, 2.2.2,
2.3.0, 2.3.1, 2.4.0, 2.4.1, 2.5.0, 2.5.1, 2.6.0)
ERROR: No matching distribution found for torch==1.4.0
2025-03-27 07:34:24.600234: I tensorflow/core/util/port.cc:153] oneDNN custom ope
rations are on. You may see slightly different numerical results due to floating-
point round-off errors from different computation orders. To turn them off, set t
he environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-03-27 07:34:24.618331: E external/local_xla/xla/stream_executor/cuda/cuda_ff
t.cc:477] Unable to register cuFFT factory: Attempting to register factory for pl
ugin cuFFT when one has already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to S
TDERR
E0000 00:00:1743060864.640218    2853 cuda_dnn.cc:8310] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already bee
n registered
E0000 00:00:1743060864.646921    2853 cuda_blas.cc:1418] Unable to register cuBLA
S factory: Attempting to register factory for plugin cuBLAS when one has already
been registered
2025-03-27 07:34:24.669232: I tensorflow/core/platform/cpu_feature_guard.cc:210]
This TensorFlow binary is optimized to use available CPU instructions in performa
nce-critical operations.
To enable the following instructions: AVX2 AVX512F AVX512_VNNI FMA, in other oper
ations, rebuild TensorFlow with the appropriate compiler flags.
usage: metric.py [-h] [--input_file INPUT_FILE] [--lm_model LM_MODEL] [--output_f
ile OUTPUT_FILE]
metric.py: error: unrecognized arguments: --model_name AdamLucek/Orpo-Llama-3.2-1
B-40k
```

StereoSet What it is: StereoSet is a benchmark dataset that evaluates stereotypes and biases in language models based on stereotypical associations with gender, race, religion, and profession.

Measures biases using masked language modeling tasks.

Dataset Link: GitHub: https://github.com/moinnadeem/StereoSet

Paper: StereoSet: Measuring stereotypical bias in pretrained language models

In [ ]:

```
%cd /content/
!git clone https://github.com/moinnadeem/StereoSet.git
%cd StereoSet
%pip install -r requirements.txt --force

# Run evaluation with HuggingFace model
!python evaluation.py --model_name AdamLucek/Orpo-Llama-3.2-1B-40k --input_file
```

```
/content
Cloning into 'StereoSet'...
remote: Enumerating objects: 83, done.
remote: Counting objects: 100% (13/13), done.
remote: Compressing objects: 100% (5/5), done.
remote: Total 83 (delta 9), reused 8 (delta 8), pack-reused 70 (from 1)
Receiving objects: 100% (83/83), 3.75 MiB | 20.99 MiB/s, done.
Resolving deltas: 100% (28/28), done.
/content/StereoSet
Collecting blis==0.4.1 (from -r requirements.txt (line 1))
  Downloading blis-0.4.1.tar.gz (1.8 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.8/1.8 MB 6.4 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Collecting boto3==1.12.36 (from -r requirements.txt (line 2))
  Downloading boto3-1.12.36-py2.py3-none-any.whl.metadata (4.7 kB)
Collecting botocore==1.15.36 (from -r requirements.txt (line 3))
  Downloading botocore-1.15.36-py2.py3-none-any.whl.metadata (3.0 kB)
Collecting catalogue==1.0.0 (from -r requirements.txt (line 4))
  Downloading catalogue-1.0.0-py2.py3-none-any.whl.metadata (13 kB)
Collecting certifi==2019.11.28 (from -r requirements.txt (line 5))
  Downloading certifi-2019.11.28-py2.py3-none-any.whl.metadata (2.5 kB)
Collecting chardet==3.0.4 (from -r requirements.txt (line 6))
  Downloading chardet-3.0.4-py2.py3-none-any.whl.metadata (3.2 kB)
Collecting click==7.1.1 (from -r requirements.txt (line 7))
  Downloading click-7.1.1-py2.py3-none-any.whl.metadata (2.9 kB)
Collecting colorama==0.4.3 (from -r requirements.txt (line 8))
  Downloading colorama-0.4.3-py2.py3-none-any.whl.metadata (14 kB)
Collecting cymem==2.0.3 (from -r requirements.txt (line 9))
  Downloading cymem-2.0.3.tar.gz (51 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 51.0/51.0 kB 4.3 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Collecting docutils==0.15.2 (from -r requirements.txt (line 10))
  Downloading docutils-0.15.2-py3-none-any.whl.metadata (2.6 kB)
ERROR: Could not find a version that satisfies the requirement en-core-web-sm==2.
2.5 (from versions: none)
ERROR: No matching distribution found for en-core-web-sm==2.2.5
python3: can't open file '/content/StereoSet/evaluation.py': [Errno 2] No such fi
le or directory
```

HolisticBias What it is: HolisticBias is a benchmark dataset designed to comprehensively evaluate social biases (gender, race, religion, etc.) in large language models (LLMs).

The dataset covers multiple demographic groups and interaction contexts to provide a "holistic" view of bias.

Dataset Link: GitHub: https://github.com/allenai/holisticbias

Paper: Holistic Evaluation of Language Models

```
In [ ]: %cd /content/
        !git clone https://github.com/allenai/holisticbias.git
        %cd holisticbias
        %pip install -r requirements.txt --force

        # Example command
        !python evaluate.py --model AdamLucek/Orpo-Llama-3.2-1B-40k
```

```
/content
Cloning into 'holisticbias'...
fatal: could not read Username for 'https://github.com': No such device or addres
s
[Errno 2] No such file or directory: 'holisticbias'
/content
ERROR: Could not open requirements file: [Errno 2] No such file or directory: 're
quirements.txt'
python3: can't open file '/content/evaluate.py': [Errno 2] No such file or direct
ory
```