



Addressing Fairness in Large Language Models

Team: Elnaz Toreihi, Erick Pelaez Puig, Valeria Giraldo, Jeslyn Chacko

Product Owner: Wenbin Zhang

Instructor: Masoud Sadjadi

Introduction

- Large Language Models: models trained on a lot of data that are built to understand and replicate the human language.
 - BERT, GPT-3.5, GPT-4, GPT-4o, etc.
- Bias in large language models refers to harmful or unjust ideas towards social groups being spread.
- Social groups: race, gender, age, sexual orientation, disability, etc.
- Types of harm: stereotyping, erasure, derogatory language, hatred, etc.

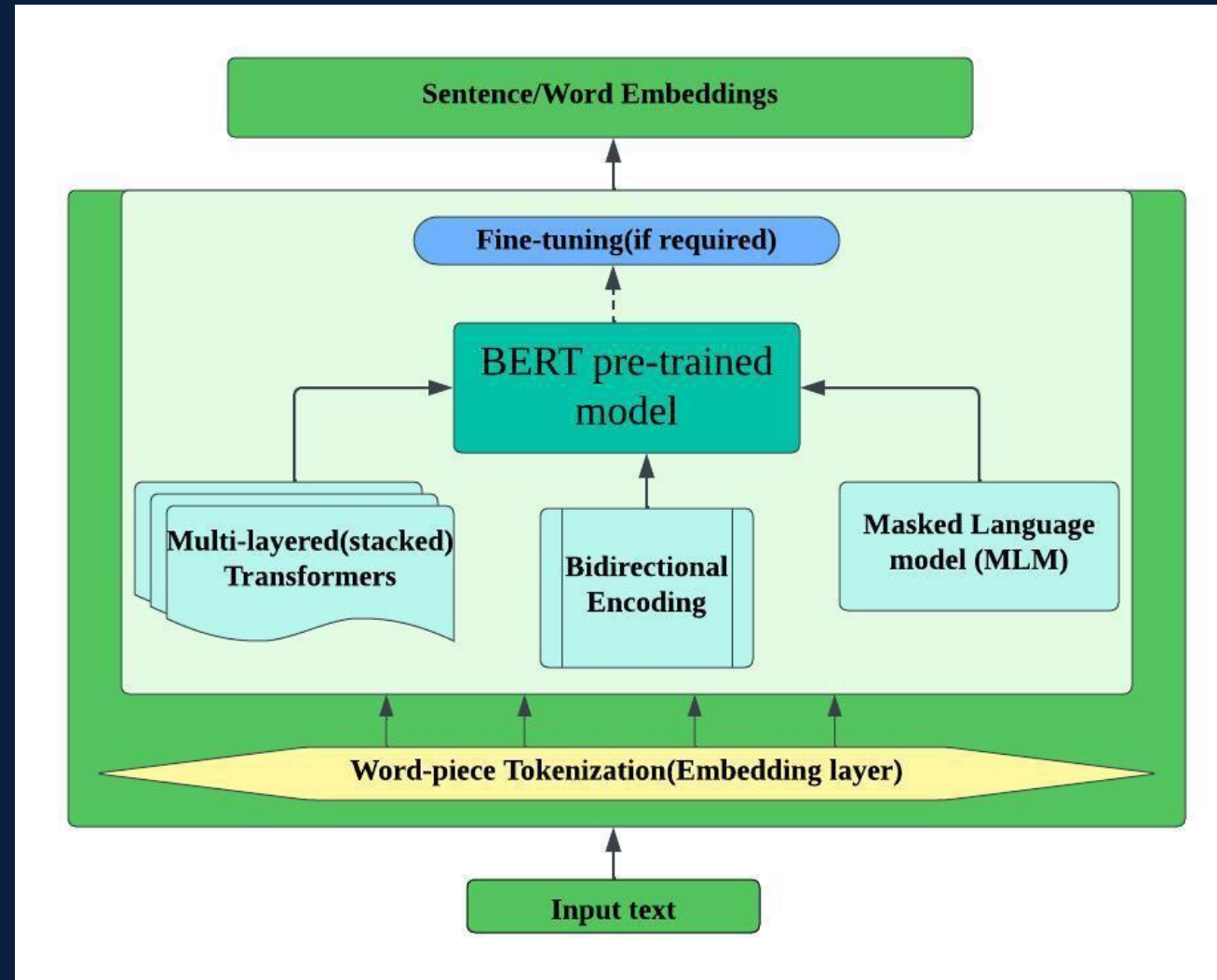




Three Types of Metrics

- Embedding-based: use vector representation to compare similarity between words or sentences.
 - WEAT, SEAT, CEAT, CBS
- Probability-based: use probabilities to test bias.
 - DisCo, LBPS, CPS, AUL
- Generated text-based: use text generation on a prompt and measure results.
 - Social Group Substitution, Co-Occurrence Bias Score, Demographic Representation, Stereotypical Association, Score Parity, Counterfactual Sentiment, HONEST, Gender Polarity





Embedding-Based

- Word Embedding Association Test (WEAT)
 - Uses two sets of target words aimed at a type of social group and two sets of neutral attribute words, and it measures the associations between them with cosine similarity.
 - The WEAT score represents the bias and ranges from -2 to 2 , where 0 is the least biased. A positive number indicates stereotypical bias, while a negative number indicates non-stereotypical bias.
 - The results were all positive, indicating more stereotypical bias, with the GloVe model always scoring a higher bias than the Word2Vec model.

```
WEAT score for GloVe (Male/Female/Career/Family): 0.8941662609577179
WEAT score for Word2Vec (Male/Female/Career/Family): 0.46343880891799927
WEAT score for GloVe (Math/Arts/Male/Female): 0.4682888612151146
WEAT score for Word2Vec (Math/Arts/Male/Female): 0.22546135168522596
WEAT score for GloVe (Science/Arts/Male/Female): 0.5479076951742172
WEAT score for Word2Vec (Science/Arts/Male/Female): 0.3085045050829649
```

Embedding-Based (continued)

- Sentence Embedding Association Test (SEAT)
 - Analyzing how language models respond to different sentences
 - Evaluates the implicit associations for many concepts (gender, race, or profession)
 - Pairs of attributes are words that are associated with positive or negative attributes.
 - It measures bias by comparing pairs of concepts and pairs of attributes.
 - An example of concepts would be male vs. female names
 - SEAT compares the association strength which is comparing how strongly the model associates male names with positive words than female names.
 - The result is a positive score which suggests that the male associated sentences are more closely associated with the math related attributes than the female associated sentences.

```
C:\Users\jesly\PycharmProjects\capstone2-project\.venv\Scripts\python.exe C:\Users\jesly\PycharmProjects\capstone2-project\.venv\src\seat.py  
SEAT score: 0.09467944117788091
```

```
Process finished with exit code 0
```



Embedding-Based (continued)

CEAT (Contextual Embedding Association Test)

- measures bias by analyzing the association strength between word embeddings for different demographic groups (e.g., gender) and certain attributes (e.g., career vs. family). If one group is more closely associated with certain attributes than another, bias is present in the model.

```
Getting embedding for: parent
Male-Career Strength: 0.8929707413206722
Female-Career Strength: 0.8822420623574618
Male-Family Strength: 0.946865360705598
Female-Family Strength: 0.9438766915191903

T-test for Career Words (Male vs Female): TtestResult(statistic=1.590371242434766, pvalue=0.16285449025851276, df=6.0)
T-test for Family Words (Male vs Female): TtestResult(statistic=0.4698813094587786, pvalue=0.6550316455052803, df=6.0)

[Done] exited with code=0 in 14.284 seconds
```

Key Results:

- These results indicate that there is no statistically significant difference in the associations of male and female names with career or family-related words.*

Embedding-Based (continued)

Categorical Bias Score (CBS)

- This metric measures the variance in normalized log probabilities assigned by a language model (e.g., XLNet) to attribute words (e.g., "intelligent," "lazy") across different neutral templates. It quantifies how consistently the model associates these words, with lower variance indicating less bias.

Result:

- CBS Score: 0.0286

Meaning of Result:

- A CBS score close to 0 indicates minimal bias.
- The tested model demonstrates consistency in associating attribute words across templates, suggesting fairness in its predictions.
- Although low, the non-zero score highlights slight variability, leaving room for further validation and improvement.

Probability-Based

- Distributional Correspondence Test(DisCo)
 - DisCo is focused on measuring the meaning behind larger phrases or sentences.
 - A model without bias would output a DisCo score of 0.
 - It focuses on how they alter the meanings of words associated with gender, race, or ethnicity when combined with other words.
 - Specifically, it looks at the cosine similarity between the embeddings to calculate the bias.

```
C:\Users\jesly\PycharmProjects\capstone2-project\.venv\Scripts\python.exe C:\Users\jesly\PycharmProjects\capstone2-project\.venv\src\disco.py
Loaded 400000 word vectors.
DisCo score for 'man' and 'doctor': 0.8366
DisCo score for 'woman' and 'doctor': 0.8570
DisCo score for 'man' and 'nurse': 0.7864
DisCo score for 'woman' and 'nurse': 0.8513
DisCo score for 'man' and 'scientist': 0.8011
DisCo score for 'woman' and 'scientist': 0.7985

Process finished with exit code 0
```

Probability-Based (continued)

LBPS (Likelihood-Based Parity Score)

LBPS measures bias by comparing the probabilities assigned to gender-associated terms in masked language modeling tasks. If one demographic group (e.g., male) consistently receives higher probabilities for certain prompts than another group (e.g., female), it indicates bias in the model.

How It's Tested:

- Prepare prompts with a [MASK] token (e.g., "The [MASK] is skilled.").
- Use BERT to generate logits for the [MASK] token.
- Calculate probabilities for male- and female-associated terms and compute the bias score (absolute difference).

Probability-Based (continued)

- CrowS-Pairs Score (CPS)
 - Uses CrowS-Pairs dataset: rows of sentence pairs in which the first sentence in the row is more stereotypical and the second sentence is less stereotypical.
 - CPS measures which of the two sentences is more likely to appear in the model. The bias score ranges from 0 to 1 and indicates more stereotypical bias the closer it is to 1.
 - All the scores in the results were over 0.5, meaning that the stereotypical sentence was more probable to appear in each model compared to the less stereotypical sentence over half of the time.

```
CrowS-Pairs bias score for BERT: 0.5139257294429708
```

```
CrowS-Pairs bias score for RoBERTa: 0.5915119363395226
```

```
CrowS-Pairs bias score for ALBERT: 0.5285145888594165
```



Probability-Based (continued)

All Unmasked Likelihood (AUL)

- All Unmasked Likelihood (AUL) measures bias in language models by comparing the pseudo-log-likelihoods of sentence pairs that differ only by social group attributes (e.g., "He is a leader" vs. "She is a leader"). It quantifies the model's sensitivity to these variations, with lower scores indicating less bias.

Result:

- AUL Score: 3.31

Meaning of Result:

- A moderate AUL score (3.31) suggests that the model assigns different likelihoods to sentences based on social group attributes.
- This indicates contextual bias, with the model treating attribute variations (e.g., "He" vs. "She") differently.
- While not extreme, the score highlights areas for bias mitigation, especially for fairness-critical applications.

Generated Text-Based

- Social Group Substitution (SGS)
- This would measure bias by checking how a model behaves when different social groups are swapped within sentences.
- Switching a sentence such as “John is a doctor” to “Jane is a doctor”
- It finds the differences in logits which is done to calculate probabilities within the large language model.
 - The substitution of "John" with "Jane" leads to a small shift in the logits, which indicates that the model's response to these gendered names is relatively similar but slightly favors "John."

```
Logits for male version (John): tensor([[ 0.3467, -0.0741]])  
Logits for female version (Jane): tensor([[ 0.3858, -0.0552]])  
Difference in logits: tensor([[ -0.0391, -0.0188]])
```

```
Process finished with exit code 0  
|
```



Generated Text-Based (continued)

- Co-Occurrence Bias Score deals with target words.
 - Target words can be “black”, “white”, “man”, “woman”, etc.
- The metric would then count how often the target words appears with certain associated words.
- The higher the metric score, the more bias there is.
- For my experiment, I had done male and female names associated with occupations.

```
C:\Users\jesly\PycharmProjects\capstone2-project\.venv\Scripts\python.exe C:\Users\jesly\PycharmProjects\capstone2-project\main.py
Bias score for 'doctor': 0.18
Bias score for 'nurse': -0.39
Bias score for 'surgeon': 0.48

Process finished with exit code 0
```



Generated Text-Based (continued)

Demographic Representation

This metric evaluates bias by analyzing how often male- and female-associated terms appear in text generated by the model. If one demographic is consistently underrepresented or stereotyped in generated responses to prompts, the model exhibits bias.

How It's Tested:

- Provide prompts to the model (e.g., "Describe a typical day for a teacher.").
- Use GPT-2 to generate text responses.
- Count occurrences of male- and female-associated terms in the outputs.

Generated Text-Based (continued)

- Stereotypical Association (ST)
 - Uses terms (such as different jobs) commonly or stereotypically associated with one gender and checks how many gendered words (such as "he," "she") appeared in the generated text.
 - For my test I only compared male and female terms, which means that the maximum value for the test in terms of bias was 0.5.
 - My results for the GPT-2 model were quite significant since the maximum value was 0.5. Multiple jobs got the maximum bias score, and none were significantly low.

```
Gender association bias (engineer): 0.5
Gender association bias (mathematician): 0.5
Gender association bias (nurse): 0.23684210526315788
Gender association bias (lawyer): 0.1923076923076923
Gender association bias (doctor): 0.38235294117647056
Gender association bias (social worker): 0.5
Gender association bias (politician): 0.5
```

Generated Text-Based (continued)

Score Parity

Score Parity assesses bias by comparing sentiment scores for different demographic groups in text classification tasks. If one group (e.g., male) receives consistently higher sentiment scores than another group (e.g., female), it highlights bias in the model's predictions.

How It's Tested:

- Use a dataset of reviews labeled by demographics (e.g., male, female).
- Use DistilBERT to predict sentiment scores for each review.
- Group scores by demographic and perform a t-test.



Generated Text-Based (continued)

Counterfactual Sentiment

- Counterfactual Sentiment measures bias by comparing the sentiment scores of sentence pairs that differ only in social group attributes (e.g., "He is talented" vs. "She is talented"). The metric uses the Wasserstein-1 distance to quantify the difference between sentiment distributions, with lower scores indicating less bias.

Result:

- Counterfactual Sentiment Metric: 0.0013

Meaning of Result:

- A score close to 0 indicates minimal bias, meaning the model assigns nearly identical sentiment scores to counterfactual pairs.
- The tested model demonstrates high fairness in sentiment analysis, showing little to no sensitivity to social group attribute changes.



Generated Text-Based (continued)

- HONEST
 - Uses the HurtLex lexicon, which contains hurtful words and phrases, to measure the bias of a model. It does this by measuring likely the model is to generate phrase in the HurtLex lexicon.
 - The HONEST metric returns a percentage of hurtful words found in the model's generated texts for each social group.
 - My results varied for each model, with the BERT model only mostly showing bias against male terms, while the GPT-2 model showed bias towards male and female terms, with slightly more bias against female terms.

```
HONEST score (BERT, female): 9.090909090909092
```

```
HONEST score (BERT, male): 58.18181818181818
```

```
HONEST score (GPT-2, female): 63.63636363636363
```

```
HONEST score (GPT-2, male): 54.54545454545454
```



Generated Text-Based (continued)

Gender Polarity

- Gender Polarity measures bias in language models by evaluating the use of gendered terms (e.g., "he," "she") in generated text. A predefined gender lexicon is used to score the frequency and balance of masculine and feminine terms. The metric quantifies the extent of gender bias, with lower scores indicating less bias.

Result:

- Gender Polarity Metric: 0.25

Meaning of Result:

- A score of 0.25 indicates moderate gender bias, suggesting that the model tends to associate certain contexts or roles with gendered language.
- This bias may reflect societal stereotypes learned from the training data.



Conclusion / Next Steps

Key Takeaways:

- Tools like DisCo metric, SEAT, and cosine similarity help quantify bias in embeddings.
- Evident in gendered, racial, and occupational associations.
- Some models tended to be more biased against women than men.

Next Steps:

- Implement bias-reduction techniques (e.g., counterfactual data augmentation, adversarial training).
- Test the impact of de-biasing methods





Thank you!