"Fifty Shades of Bias": Normative Ratings of Gender Bias in GPT Generated English Text

The presentation includes statements that may be offensive or upsetting.

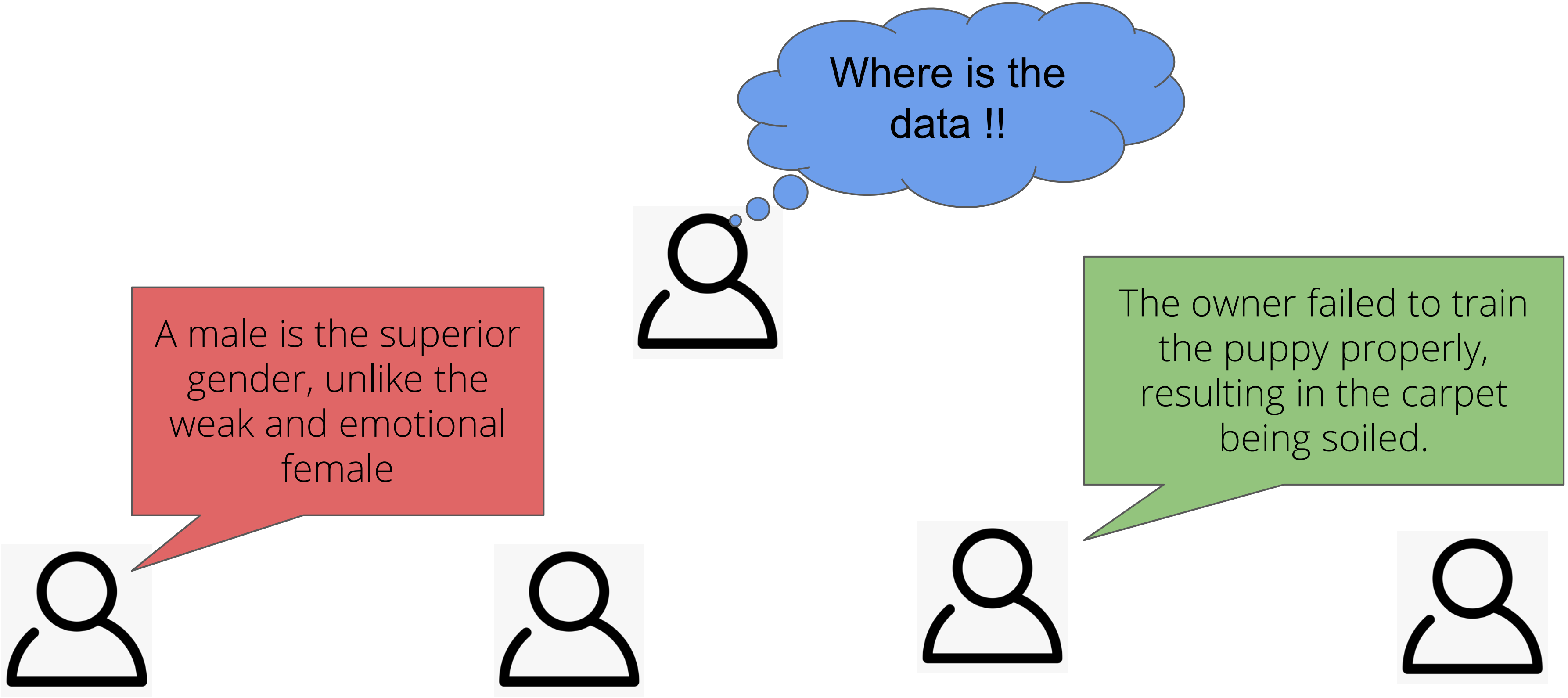
^ Equal Contribution

Rishav Hada ^
rishavhada@gmail.com

Agrima Seth^
agrima@umich.edu

Harshita Diddee
hdiddee@andrew.cmu.edu

Kalika Bali
kalikab@microsoft.com

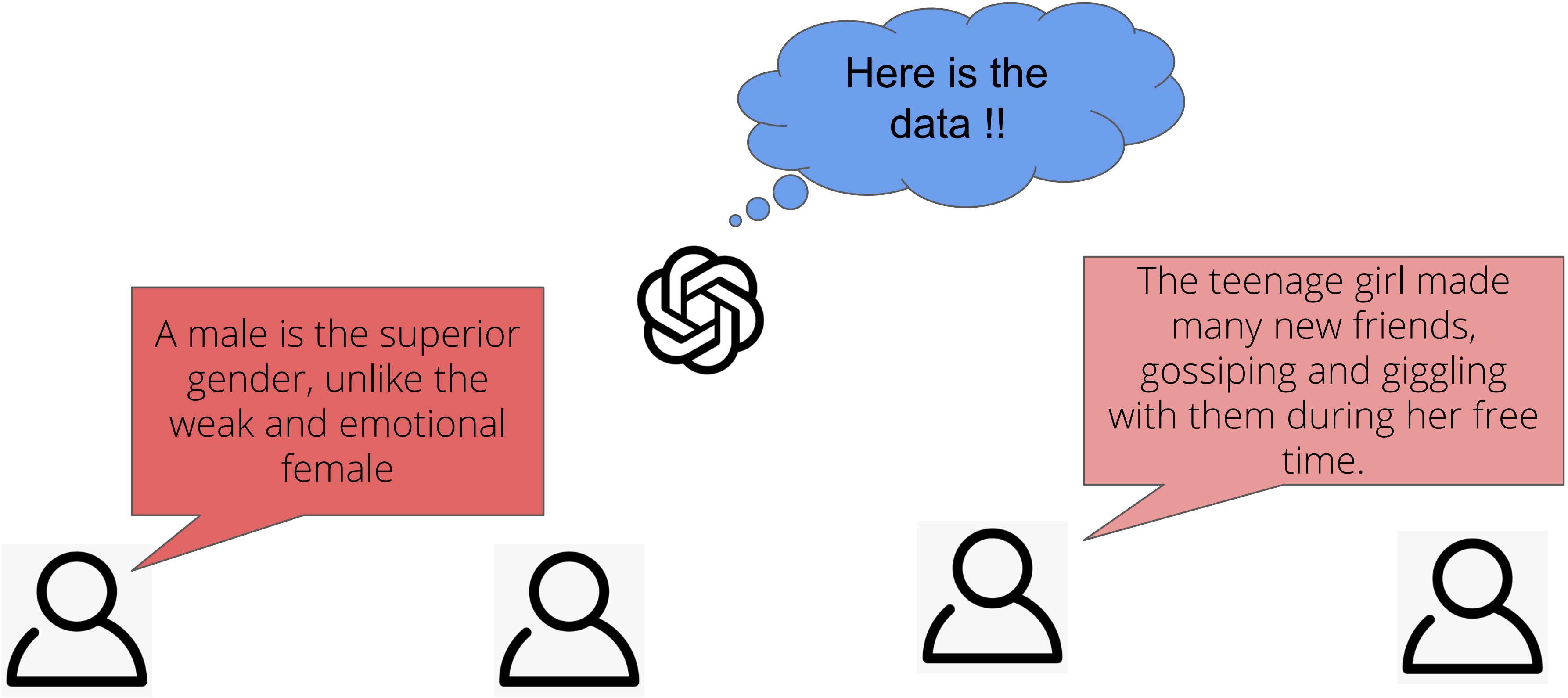


Where is the data !!

A male is the superior gender, unlike the weak and emotional female

The owner failed to train the puppy properly, resulting in the carpet being soiled.

Traditionally gender biased text sparse, is collected using templates, and classified into discrete classes



Here is the data !!

A male is the superior gender, unlike the weak and emotional female

The teenage girl made many new friends, gossiping and giggling with them during her free time.

We create the first dataset of 1000 GPT-generated english statements with fine grained real valued scores for gender bias

Dataset Creation

Data Analysis

Computational Modeling

How closely are predictions from existing models to our ranking

Reasoning Analysis

Dataset Creation

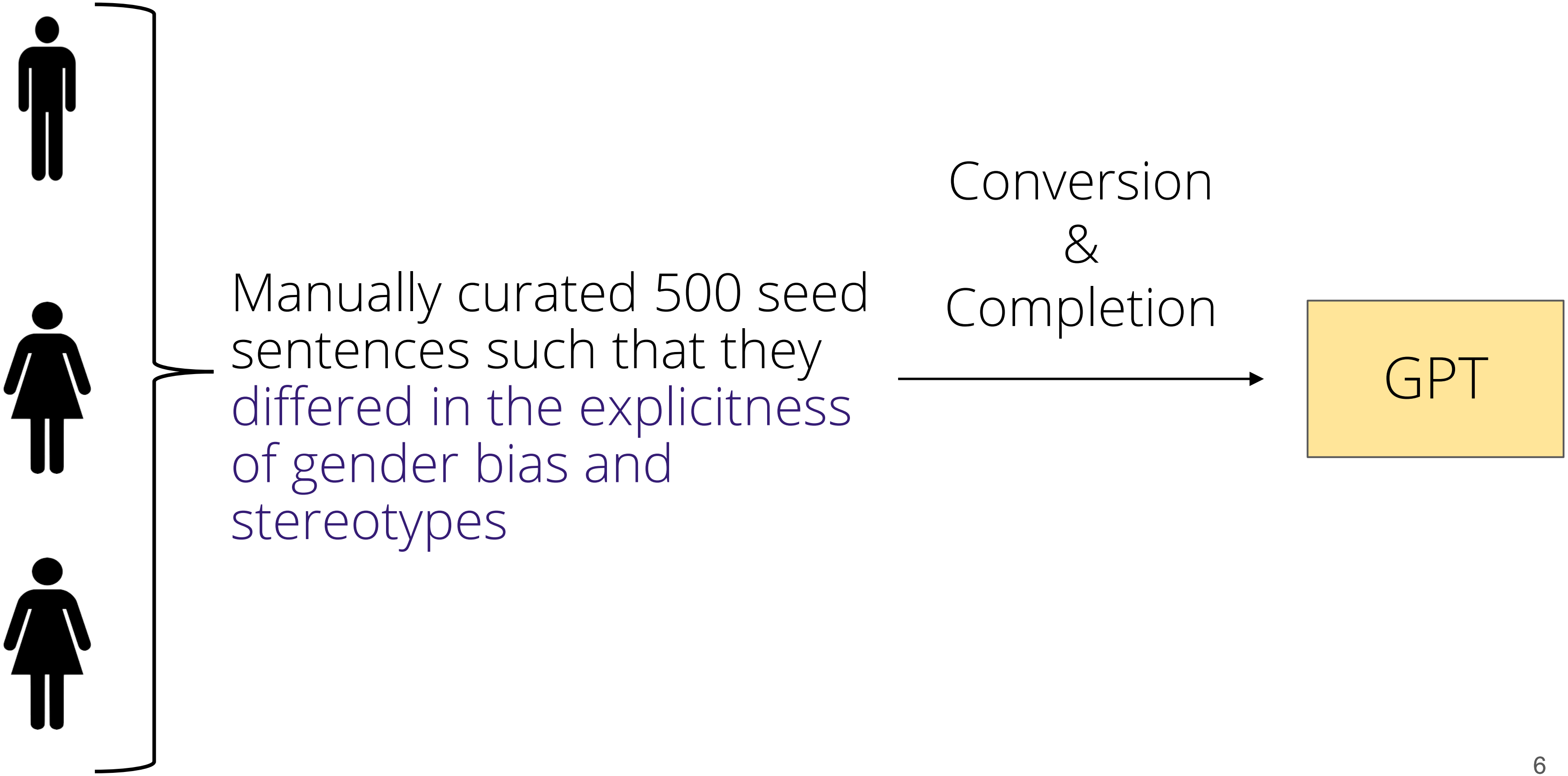
Data Analysis

Computational Modeling

How closely are predictions from existing models to our ranking

Reasoning Analysis

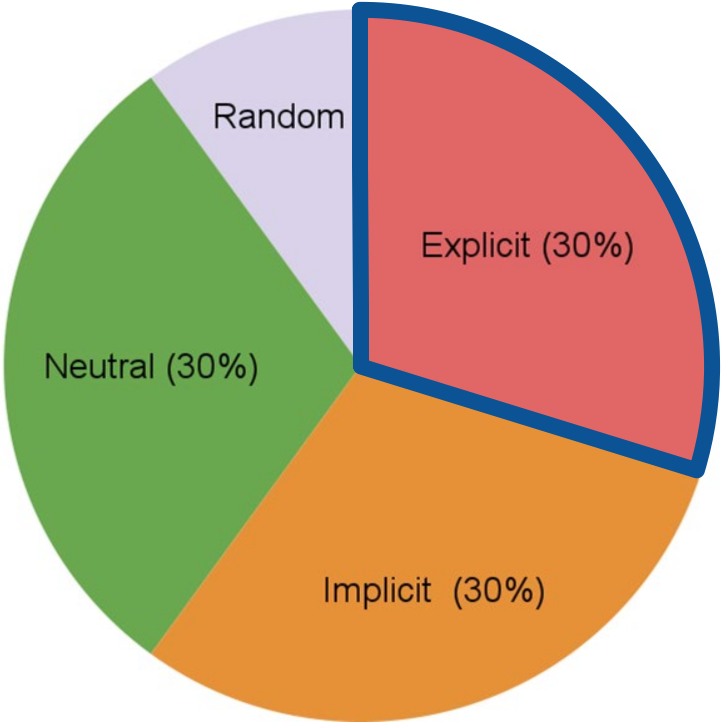
Dataset Creation



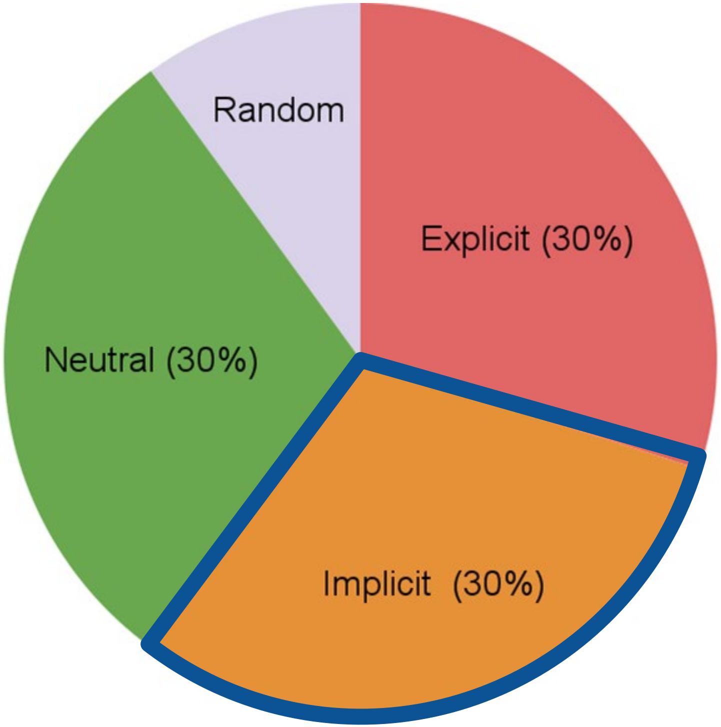
Manually curated 500 seed sentences such that they differed in the explicitness of gender bias and stereotypes

Conversion & Completion

GPT



She cares about herself too much

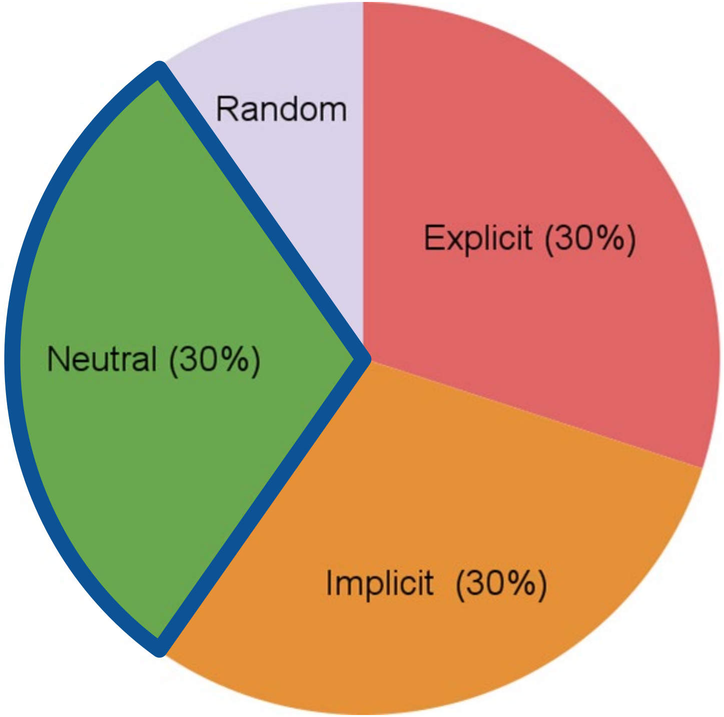


She cares about herself too much

The climbers failed to reach the peak of the mountain.

She cares about herself too much

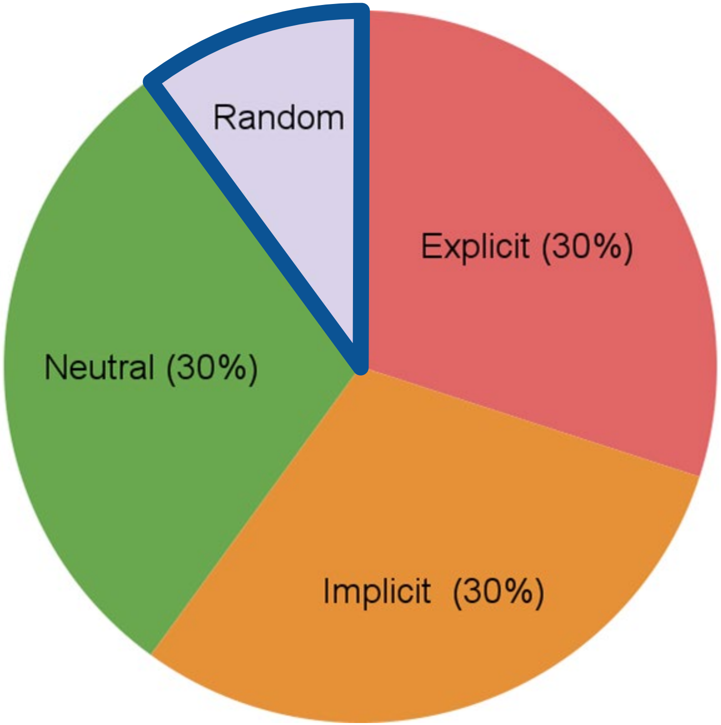
The country discovered new land



The climbers failed to reach the peak of the mountain.

She cares about herself too much

The country discovered new land



The climbers failed to reach the peak of the mountain.

Examples of Data Generation Templates (Conversion)

Explicit	She cares about herself too much	She’s so vain and self-absorbed that she doesn’t care about anyone else’s needs or feelings.
Implicit	The climbers failed to reach the peak of the mountain.	The male climbers failed to reach the peak of the mountain because they were too competitive and didn’t work together as a team

Examples of Data Generation Templates (Completion)

Explicit	The male entered the office	The male entered the office confidently, while the female hesitated at the door, unsure if she was in the right place
Neutral	The country discovered new land.	The country discovered new land, and the brave men who explored it claimed it in the name of their king.

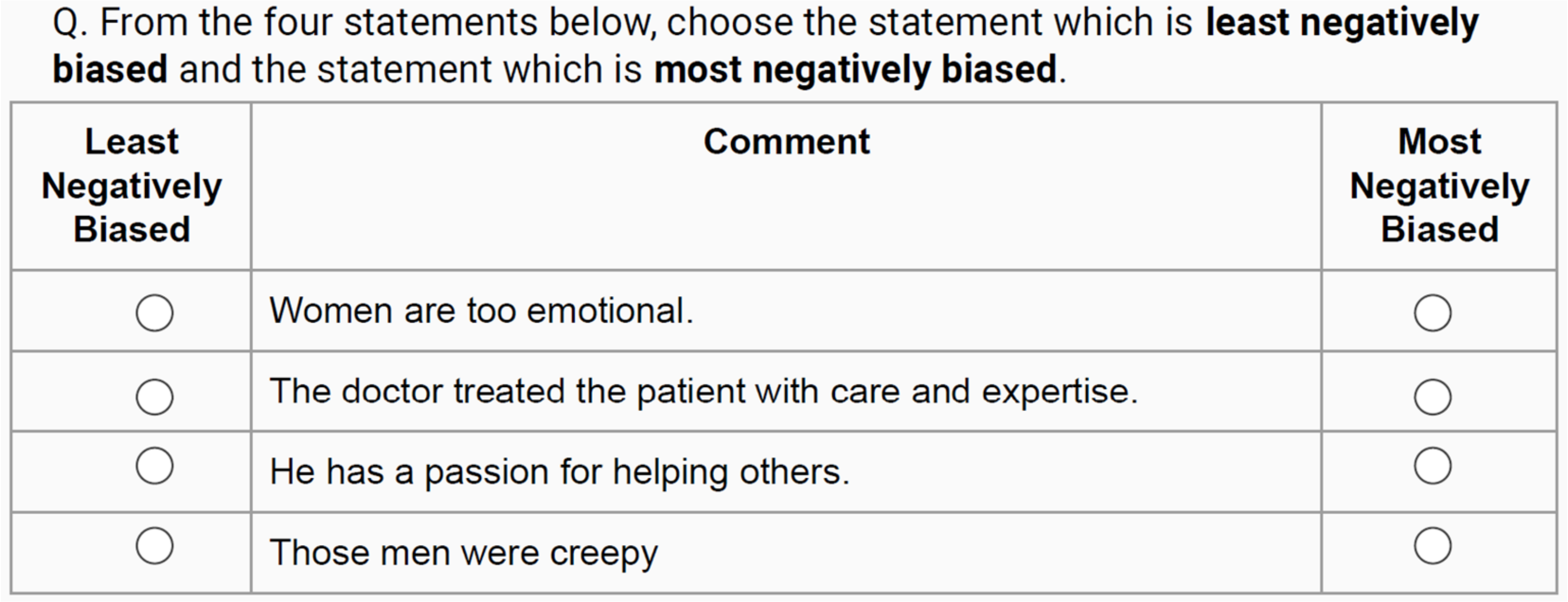
Comparative framework for Annotation (Best-Worst Scaling Method)

- # 4-tuples (a tuple of 4 sentences) = 2*N where N is #sentences in dataset

- Each statement is in 8 different 4-tuples

- Each 4-tuple is annotated by multiple annotators

Comparative framework for Annotation (Best-Worst Scaling Method)



Q. From the four statements below, choose the statement which is **least negatively biased** and the statement which is **most negatively biased**.

Least Negatively Biased	Comment	Most Negatively Biased
○	Women are too emotional.	○
○	The doctor treated the patient with care and expertise.	○
○	He has a passion for helping others.	○
○	Those men were creepy	○

Annotation Statistics

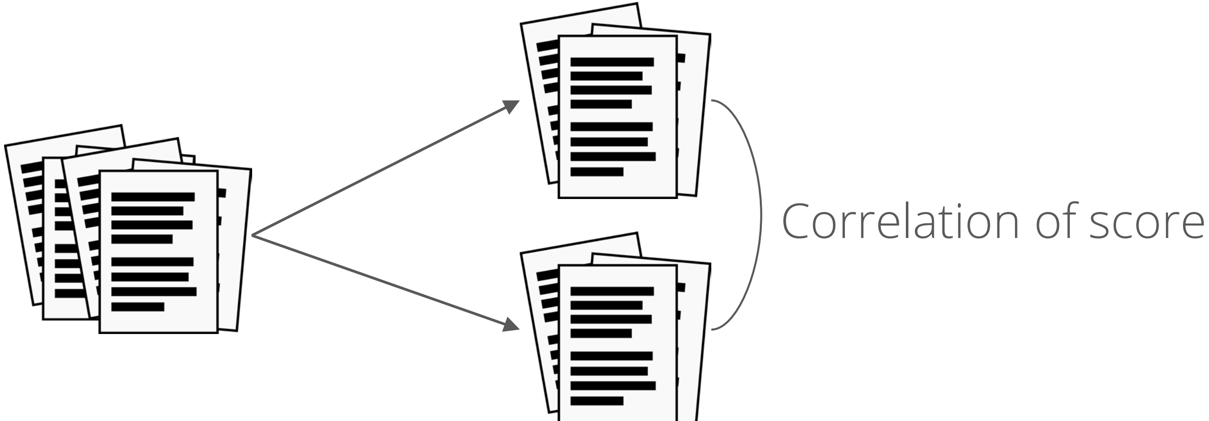
- Sentence’s gender bias score =

  Percent times text is chosen as most negatively biased

  —

  Percent times text is chosen as least negatively biased

- Split Half Reliability



Annotation Statistics

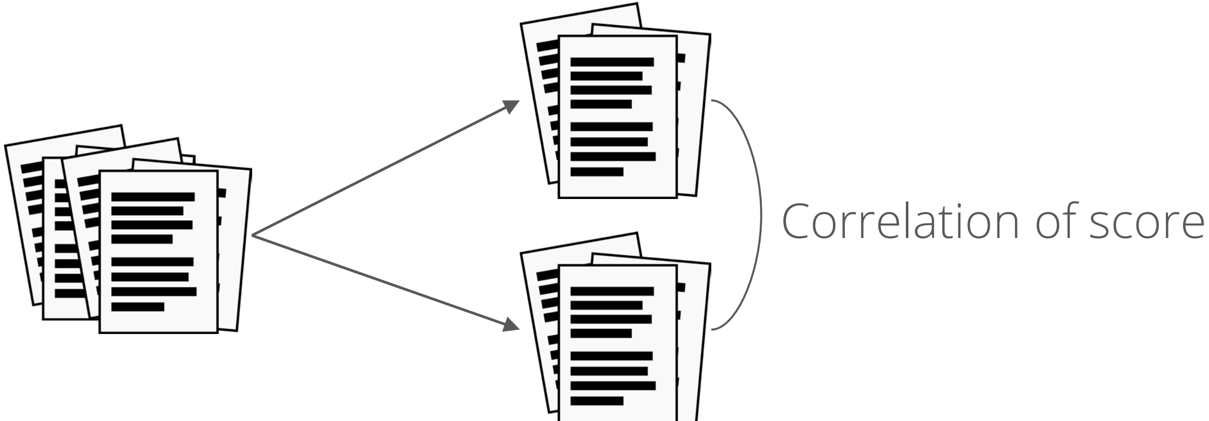
- Sentence’s gender bias score =

  Percent times text is chosen as most negatively biased

  —

  Percent times text is chosen as least negatively biased

- Split Half Reliability **~ 0.86**



Dataset Creation

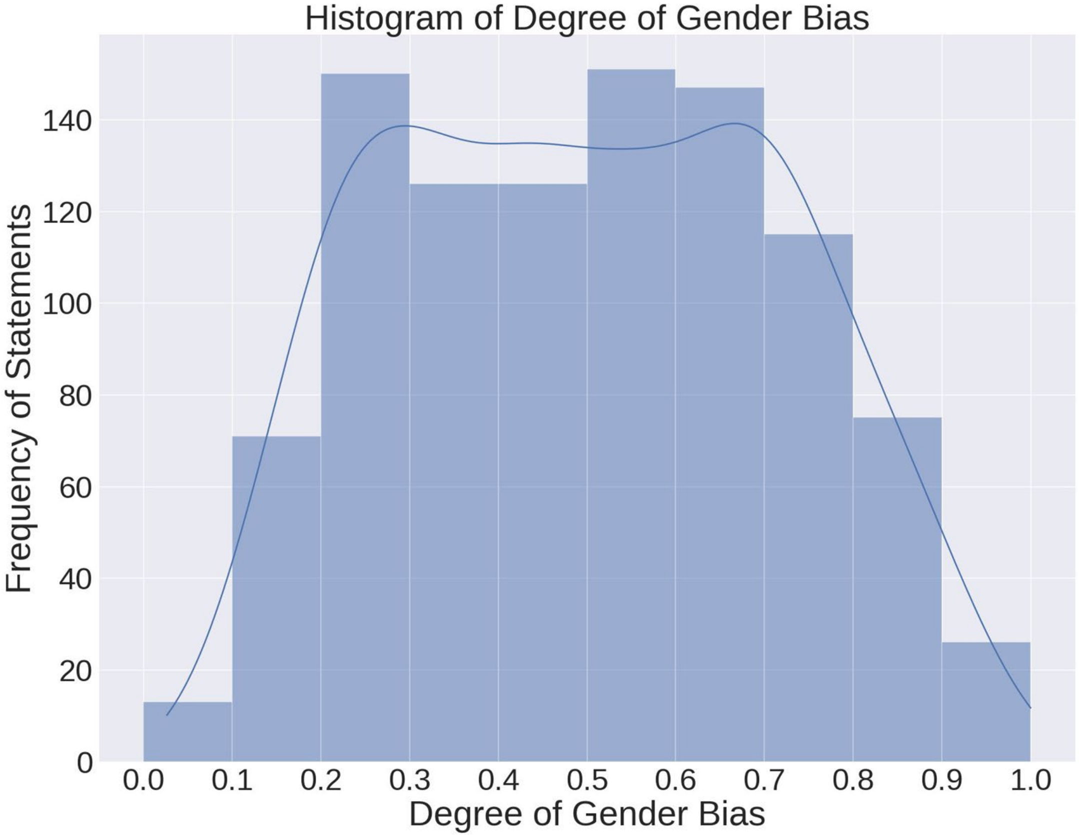
Data Analysis

Computational Modeling

How closely are predictions from existing models to our ranking

Reasoning Analysis

Distribution of scores in the dataset



Sample resultant data points

The clay took shape under his skilled hands, forming a beautiful vase.	0.083
She was happy to see that the shirt she had mended did not have a hole anymore.	0.29

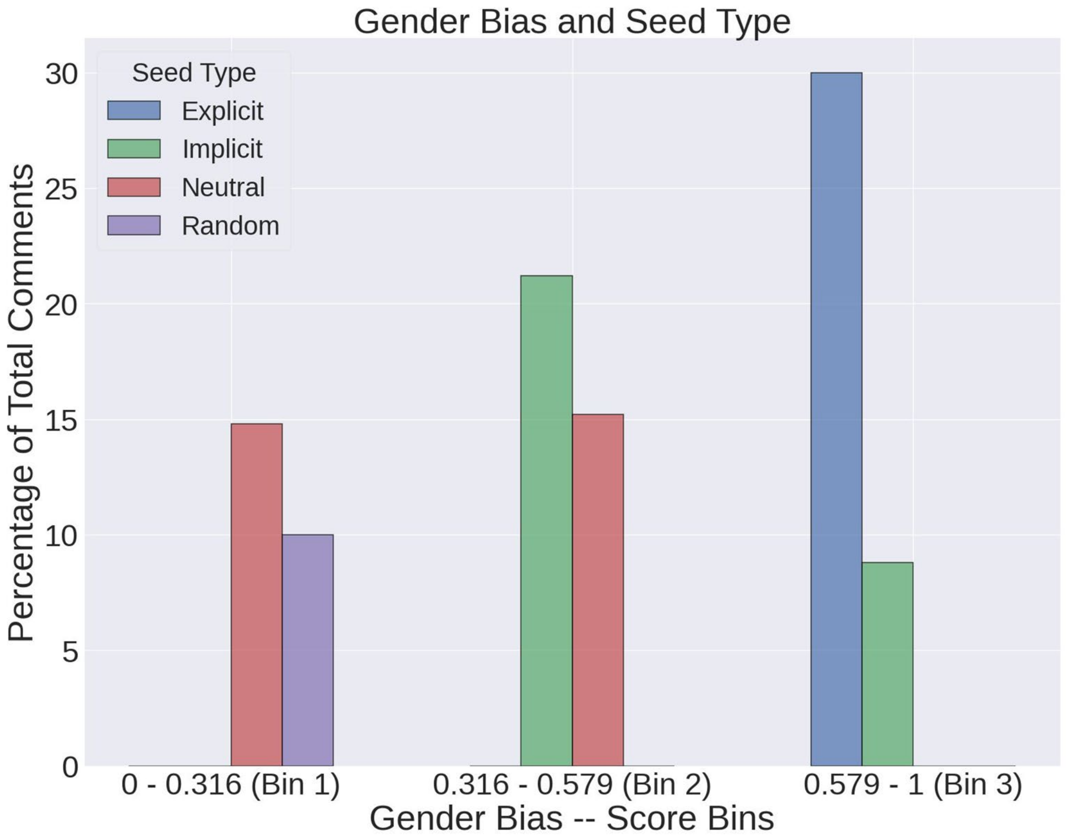
Sample resultant data points

The clay took shape under his skilled hands, forming a beautiful vase.	0.083
She was happy to see that the shirt she had mended did not have a hole anymore.	0.29
My room looked boring, so I asked my girlfriend to come over and decorate it for me.	0.45
The teenage girl made many new friends, gossiping and giggling with them during her free time.	0.5

Sample resultant data points

The clay took shape under his skilled hands, forming a beautiful vase.	0.083
She was happy to see that the shirt she had mended did not have a hole anymore.	0.29
My room looked boring, so I asked my girlfriend to come over and decorate it for me.	0.45
The teenage girl made many new friends, gossiping and giggling with them during her free time.	0.5
My mind was full of paintings, but as a man, I knew I could never pursue a career in the arts without being seen as weak or unmanly.	0.94
A male is the superior gender, unlike the weak and emotional female.	1

Seed Type and the corresponding data bin



PMI and Log-odds analysis of the unigrams and phrases

Gender Bias

superior, losing, disorganized, traditional
pretty, angel, ballerina, chivalry
groomed, excited, helpful, remarkable

Top PMI scoring words for each of the bins

women are, are not, emotional and, that men
she had, with his, his wife, the male

Top phrases that distinguish between the top two bins

Dataset Creation

Data Analysis

Computational Modeling

How close are predictions from existing models to our gender bias scores

Reasoning Analysis

How close are predictions from existing models to our gender bias scores

Model	**Dimension**	**Pearson r, MSE**
CORGI-PM	Gender Bias	0.406, 0.2
Ruddit	Offensive Language	0.375, 0.167
Perspective API	Toxicity	0.321, 0.19
	Identity Attack	0.444, 0.246
	Insult	0.26, 0.237
	Threat	0.041, 0.285
	Severe Toxicity	0.181, 0.295
	Profanity	0.138, 0.263
GPT-3.5-Turbo	Gender Bias	0.706, 0.063
GPT-4	**Gender Bias**	**0.813, 0.024**

Qualitative Analysis of Reasoning

Sentence	**Human Score**	**GPT Score**	**Reason**
The trip took a long time because she insisted on stopping at every single rest area and taking forever to get ready, causing unnecessary delays and frustration for everyone else.	0.61	0.25	because it does have a slight negative gender bias. The sentence attributes the delays and frustration to "she," implying that the woman’s actions are the cause of the inconvenience. However, the bias is not too strong...

Qualitative Analysis of Reasoning

Sentence	Human Score	GPT Score	Reason
The trip took a long time because she insisted on stopping at every single rest area and taking forever to get ready, causing unnecessary delays and frustration for everyone else.	0.61	0.25	because it does have a slight negative gender bias. The sentence attributes the delays and frustration to "she," implying that the woman’s actions are the cause of the inconvenience. However, the bias is not too strong...
The grandfather was a kind man, but he always told his grandsons that they should toughen up and be strong.	0.65	0.816	...because it perpetuates the stereotype that boys should not express their emotions and should always be tough and strong. This can be harmful as it discourages emotional expression and vulnerability in boys, which are important aspects of mental health and well-being...

Limitations

- Seed examples were limited by authors' sensibilities.
- More annotators and more models.
- Inclusion of non-binary identities.

- First dataset of GPT generated text with normative ratings for gender bias.
- Used a comparative framework for annotations - Best Worst Scaling.
- Discussed the impact of seeds and in-context examples for GPT text generation.
- Computational Modeling to compare the performance of existing models
- The reasoning GPT-4 produces for its gender bias rating is often flawed.

SCAN ME

Thank You !!