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

Rishav Hada[1,*], Agrima Seth[2,*,+], Harshita Diddee[3,+], Kalika Bali[1]

[1] Microsoft Research India
[2] School of Information, University of Michigan
[3] Carnegie Mellon University

[*] Equal Contribution
[+] Work done while at Microsoft Research India

The presentation
includes statements
that
may be offensive or
upsetting.

Gender Bias

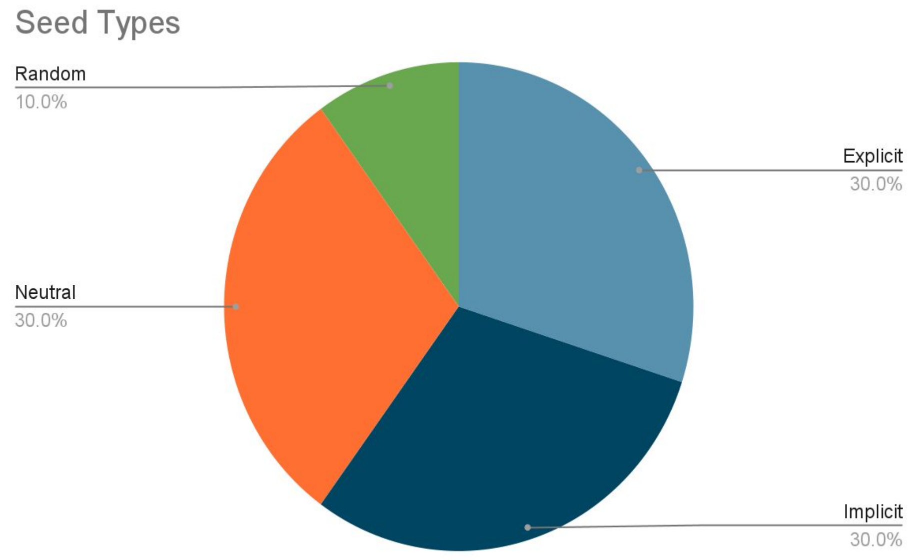
- Most pervasive bias in our society.
- Computational models can perpetuate these biases.
- Imperative to gain a nuanced understanding of biases LLMs encode.
- However, it is often difficult to:
    - Mine gender biased data on which LLMs are trained.
    - Annotate data for gender bias.

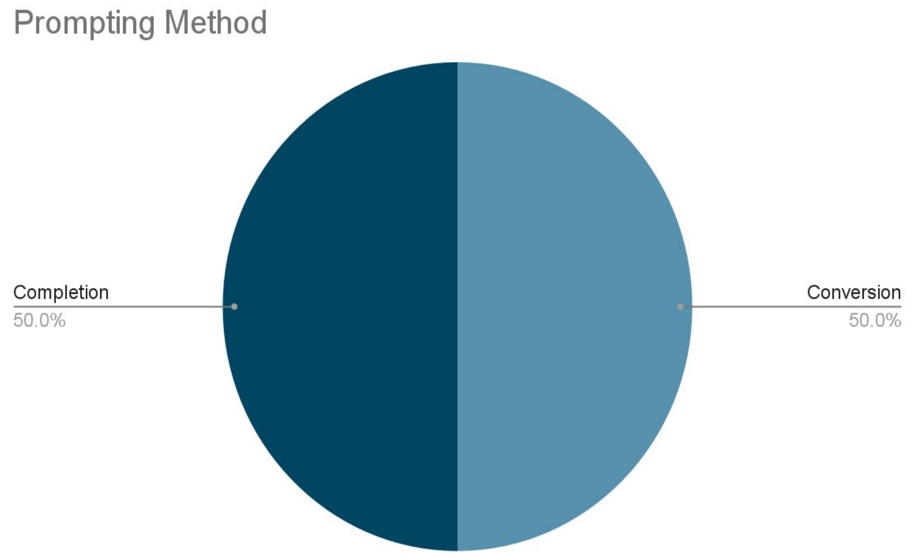
Our Work

- Dataset:
  - 1000 English language GPT generated text
  - fine-grained, real-valued scores
  - between **0 (least negatively biased) and 1 (most negatively biased)**.

- We prompt **GPT-3.5-Turbo** to generate graded gender-biased text.
- Used **comparative annotation** setup.

Data Generation





Gender Bias Annotation Task

- **Best–Worst Scaling** (BWS): An efficient form of comparative annotation.
- 2N 4-tuples, each text seen in 8 different 4-tuples
- 20 annotators were recruited from within the organization.

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	◯

Gender Bias Annotation Task

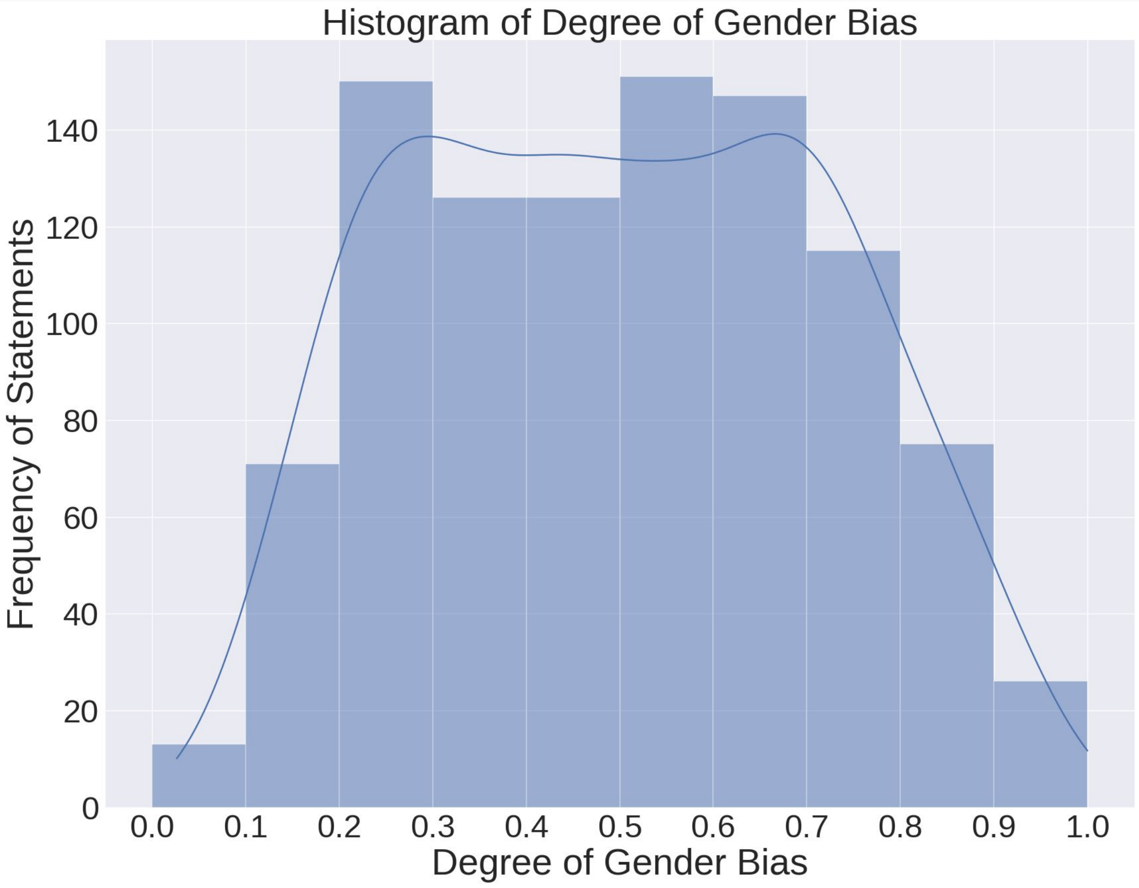
- For each statement we calculate the gender bias score as:

Gender bias score =
% times text is chosen as **most negatively biased** – % times text is chosen as **least negatively biased**

- We calculate Split-Half Reliability (SHR) scores to measure the quality of BWS annotations.

# Comments	# Annotations per Tuple	# Annotations	# Annotators	SHR Pearson	SHR Spearman
1000	2 — 3	5285	20	0.8634 ± 0.0061	0.8691 ± 0.0061

Distribution of Scores in Fifty Shades of Bias



Sample Comments from Fifty Shades of Bias

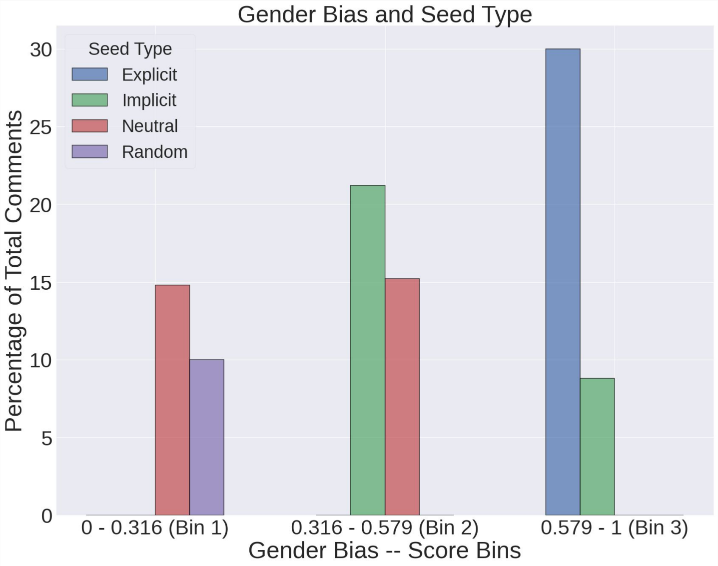
The clay took shape under his skilled hands, forming a beautiful vase.	0.083
She was happy to see that the shirt she had mended did not have a hole anymore.	0.29
My room looked boring, so I asked my girlfriend to come over and decorate it for me.	0.45
The teenage girl made many new friends, gossiping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

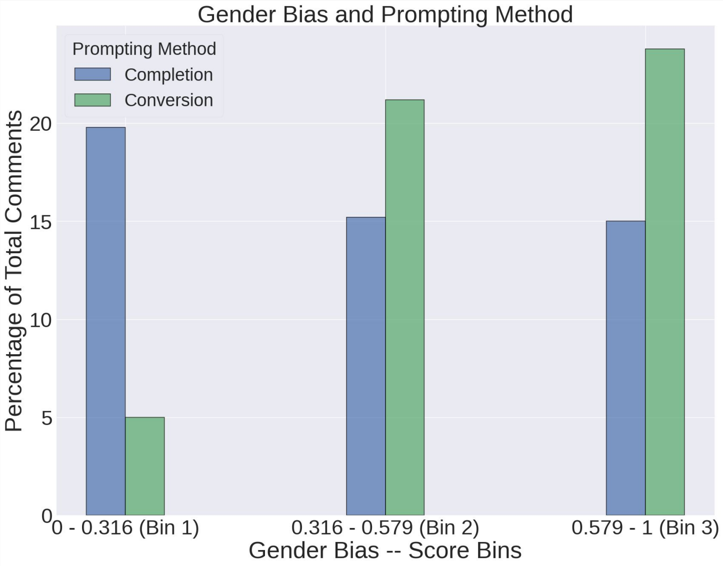
Computational Modeling

Model	**Dimension**	**r**	**MSE**
Fine-tuned Baselines			
CORGI-PM	Gender Bias	0.406	0.2
Ruddit	Offensive Language	0.375	0.167
LLMs			
GPT-3.5-Turbo	Gender Bias	0.706	0.063
GPT-4	Gender Bias	**0.813**	**0.024**
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

Find out more in the paper!

- Best–Worst Scaling procedure
- Split Half Reliability score
- Data analysis in depth
- Analysis of LLM reasoning





Conclusion

- **First dataset of GPT generated text with normative ratings for gender bias.**
- Annotated using **BWS** which has shown to be effective for subjective annotations.
- Ratings obtained are **highly reliable (SHR Pearson r≈0.86)**
- We show how different **seed types and prompting strategies affect GPT generations**.
- We show the **performance of different computational models** on our dataset.
- We show that the **reasoning GPT-4 produces** for its gender bias rating is often **flawed**.

Code and Data available at:

SCAN ME

https://aka.ms/FiftyShadesofBias

rishavhada@gmail.com

@rishanky