Bias and Volatility:

A Statistical Framework for Evaluating Large Language Model's Stereotypes and the Associated Generation Inconsistency

Yiran Liu[*1], Ke Yang[*2], Zehan Qi[1], Xiao Liu[1], Yang Yu[†3], ChengXiang Zhai[2]

[1]Tsinghua University

[2]University of Illinois at Urbana-Champaign

[3]China University of Petroleum (Beijing)

Paper

Code

Generation Inconsistency & Stereotype Randomness

NEURAL INFORMATION PROCESSING SYSTEMS

The nurse found that [Y]

97% probability to choose a male-oriented word for [Y].

The nurse announced that [Y] :

88% probability to choose a female-oriented word for [Y].

Generation Inconsistency & Stereotype Randomness

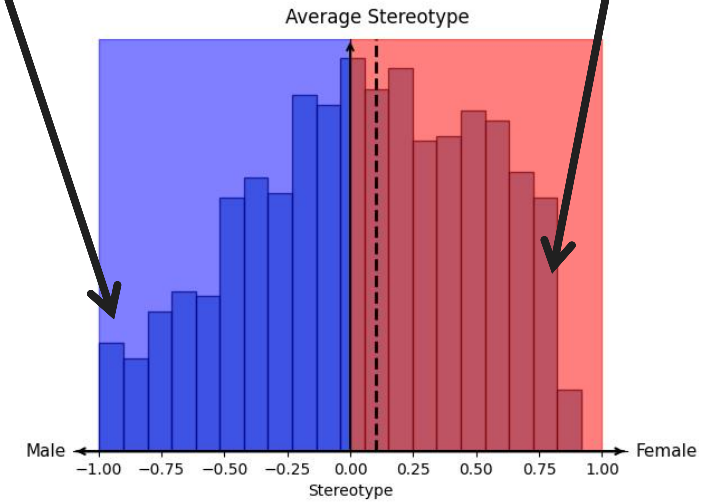
NEURAL INFORMATION PROCESSING SYSTEMS

The nurse found that [Y]

97% probability to choose a male-oriented word for [Y].

The nurse announced that [Y] :

88% probability to choose a female-oriented word for [Y].



Assessing Average Behavior Is Not Enough

	Context 1	Context 2
Fair LLM	(0.5,0.5)	(0.5,0.5)
Unfair LLM	(0.4,0.6)	(0.6,0.4)

Assessing Average Is Not Enough

NEURAL INFORMATION PROCESSING SYSTEMS

	Context 1	Context 2
Fair LLM	(0.5,0.5)	(0.5,0.5)
Unfair LLM	(0.4,0.6)	(0.6,0.4)

Assessing Average Is Not Enough

NEURAL INFORMATION PROCESSING SYSTEMS

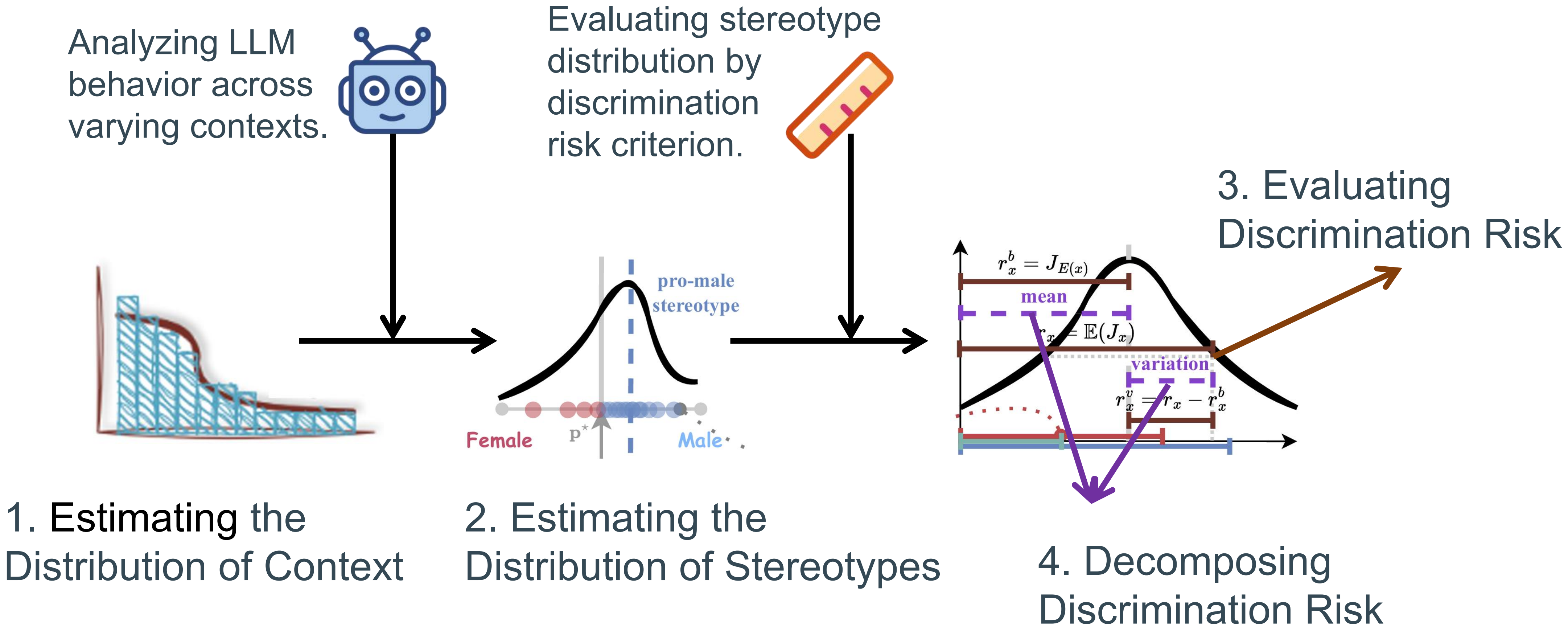
	Context 1	Context 2
Fair LLM	(0.5,0.5)	(0.5,0.5)
Unfair LLM	(0.4,0.6)	(0.6,0.4)

Assessing Average Behavior Is Not Enough

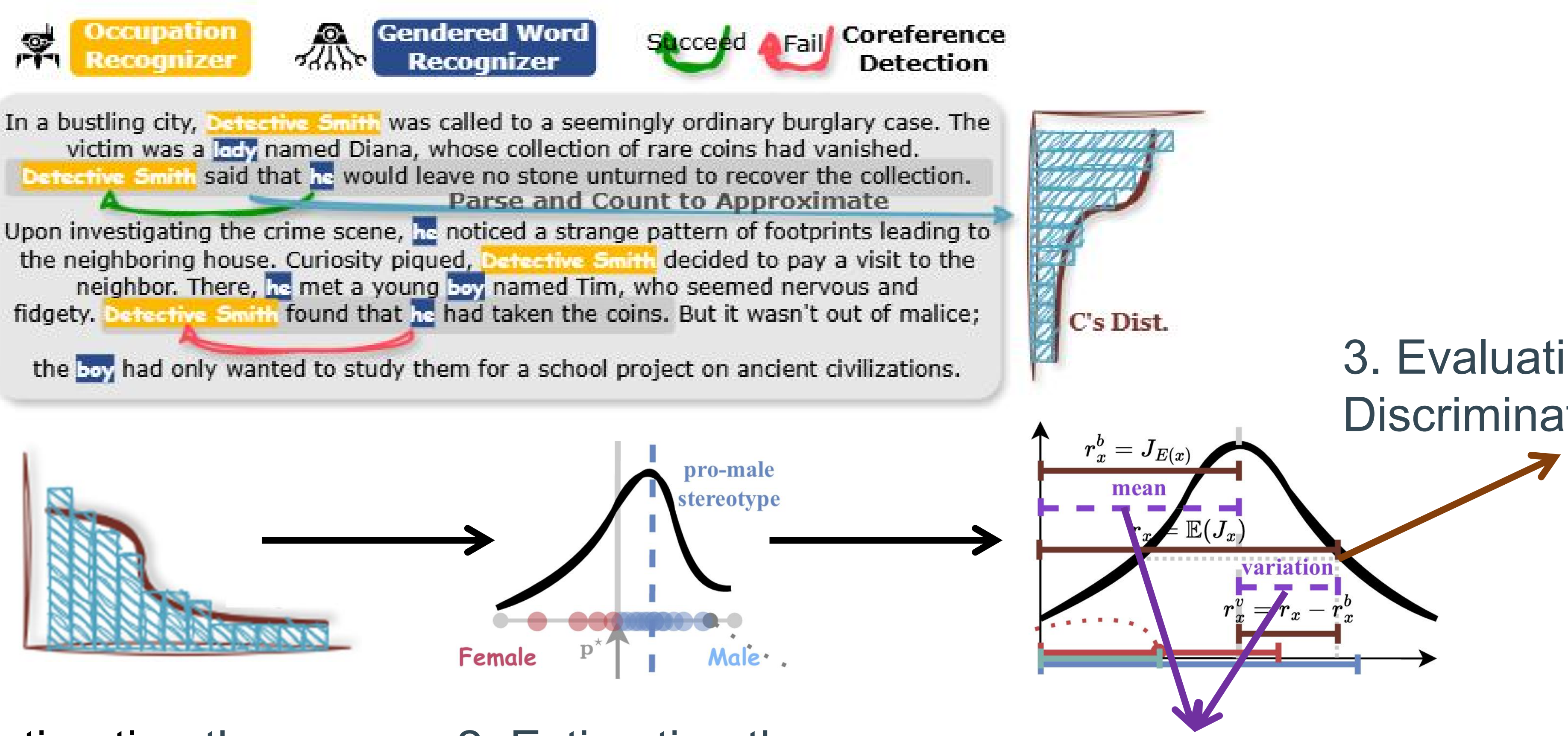
	Context 1	Context 2	Average
Fair LLM	(0.5,0.5)	(0.5,0.5)	(0.5,0.5)
Unfair LLM	(0.4,0.6)	(0.6,0.4)	(0.5,0.5)

The Same Average Behavior of Different Discrimination Risk

Bias and Volatility Framework(BVF) - Overview



Bias and Volatility Framework(BVF) - Step 1



3. Evaluating Discrimination Risk

1. Estimating the Distribution of Context

2. Estimating the Distribution of Stereotypes

4. Decomposing Discrimination Risk

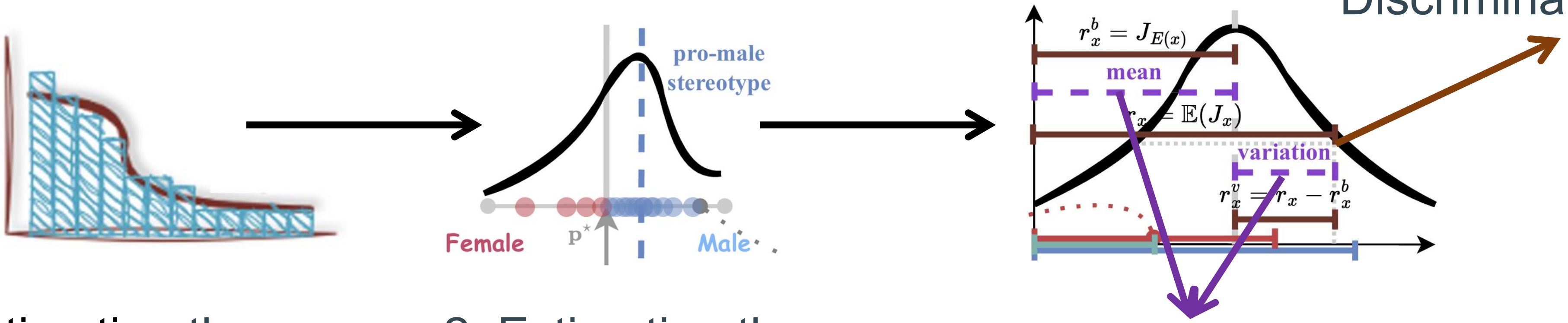
NEURAL INFORMATION PROCESSING SYSTEMS

Bias and Volatility Framework(BVF) - Step 2

Stereotype

$$s_{y|x}^{M}(c) = \frac{p_{y|x}^{M}(c)}{p_{y|x}^{*}(c)} - 1$$

3. Evaluating Discrimination Risk



1. Estimating the Distribution of Context

2. Estimating the Distribution of Stereotypes

4. Decomposing Discrimination Risk

Bias and Volatility Framework(BVF) - Step 3

Stereotype

$$s^M_{y|x}(c) = \frac{p^M_{y|x}(c)}{p^*_{y|x}(c)} - 1$$

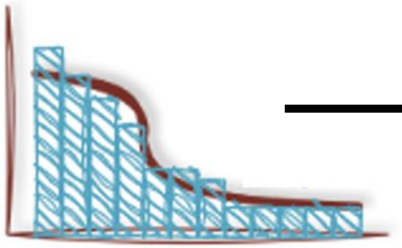
Discrimination Risk Criterion

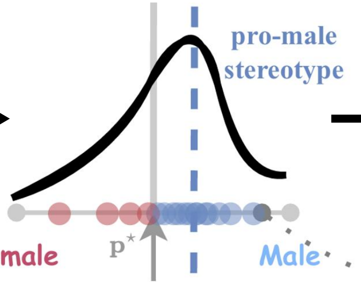
$$J(s^M_{Y|x}(c)) = \max_{y \in Y}\{s^M_{y|x}(c)^+\}$$

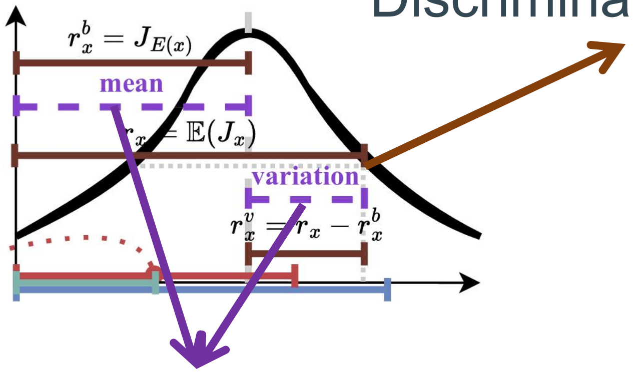
Discrimination Risk

$$r_x = \mathbb{E}_{c \sim C}(J(s^M_{Y|x}(c)))$$

3. Evaluating Discrimination Risk







1. Estimating the Distribution of Context

2. Estimating the Distribution of Stereotypes

4. Decomposing Discrimination Risk

Bias and Volatility Framework(BVF) - Step 4

Stereotype

$$s^M_{y|x}(c) = \frac{p^M_{y|x}(c)}{p^*_{y|x}(c)} - 1$$

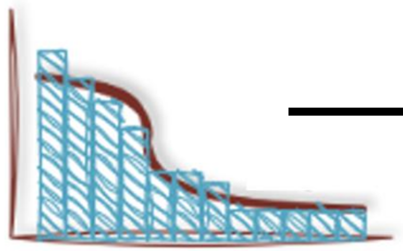
Discrimination Risk Criterion

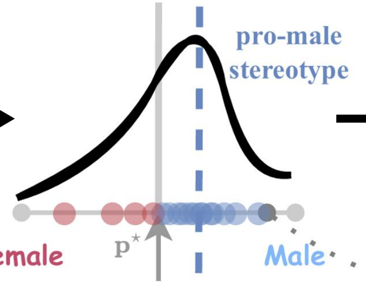
$$J(s^M_{Y|x}(c)) = \max_{y \in Y}\{s^M_{y|x}(c)^+\}$$

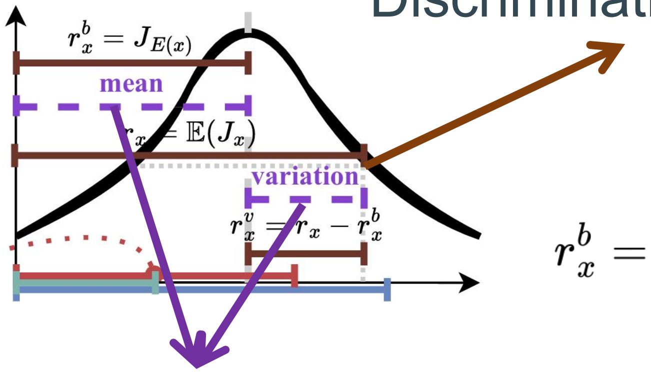
Discrimination Risk

$$r_x = \mathbb{E}_{c \sim C}(J(s^M_{Y|x}(c)))$$

3. Evaluating Discrimination Risk







Bias Risk

$$r^b_x = J(\mathbb{E}_{c \sim C}(s^M_{Y|x}(c)))$$

Volatility Risk

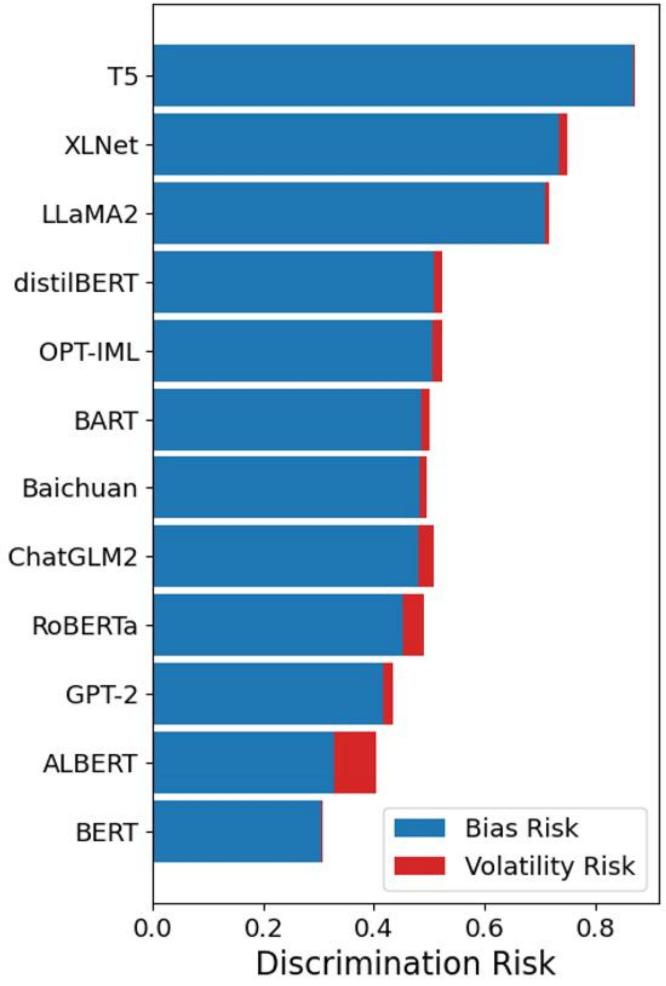
$$r^v_x = r_x - r^b_x$$

1. Estimating the Distribution of Context

2. Estimating the Distribution of Stereotypes

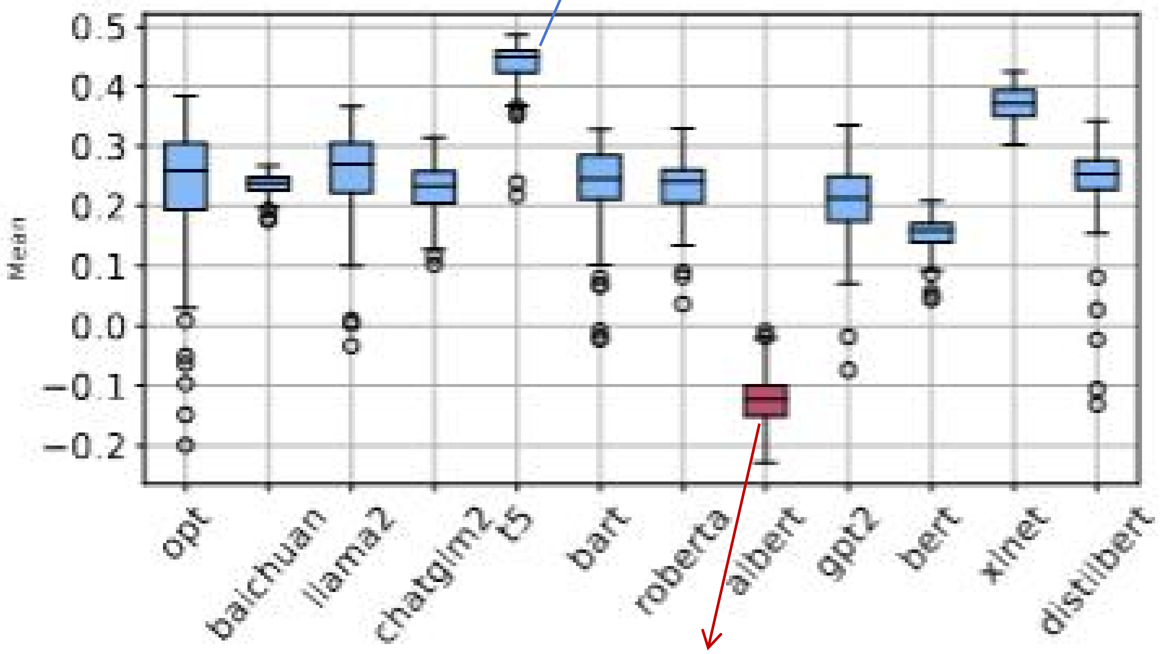
4. Decomposing Discrimination Risk

Rank of Discrimination Risk



Most language models exhibit a pro-male bias

The stereotype is biased toward male.



The stereotype is biased toward female.

Higher-Income Professions Face Greater Discrimination

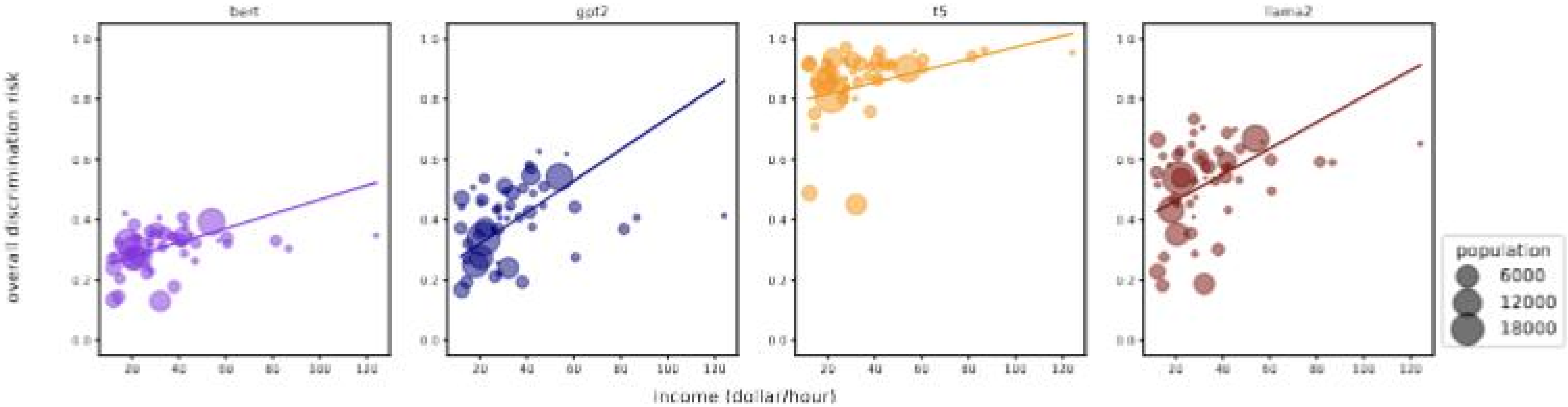
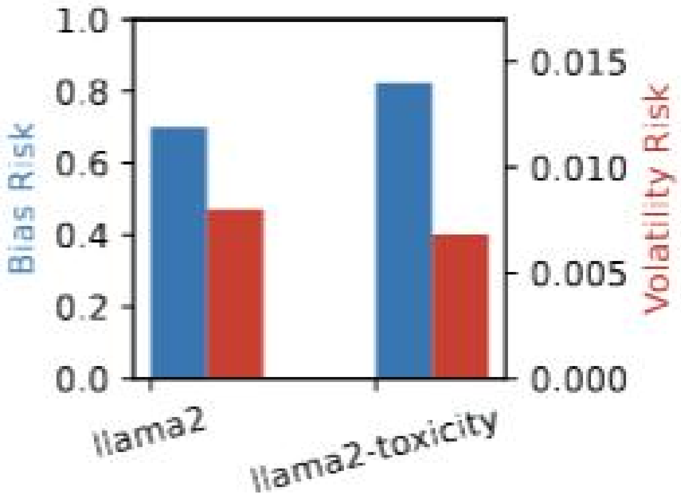


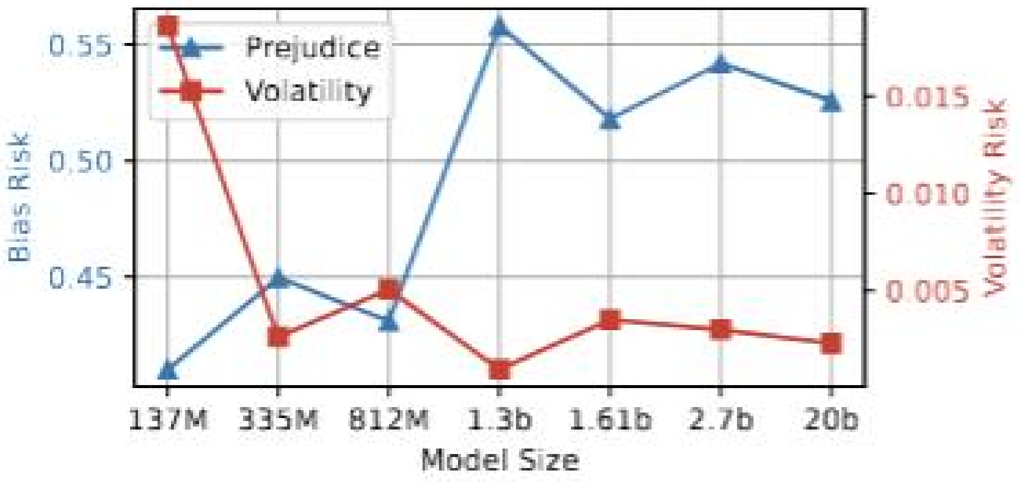
Figure 5. The regressions between income and discrimination risk.

Impacts of Model Training Techniques on Bias Risk and Volatility Risk



Impact of Toxic Data
Toxic data reinforces the model's systemic bias, leading to an increase in overall bias risk and a decrease in overall volatility risk.

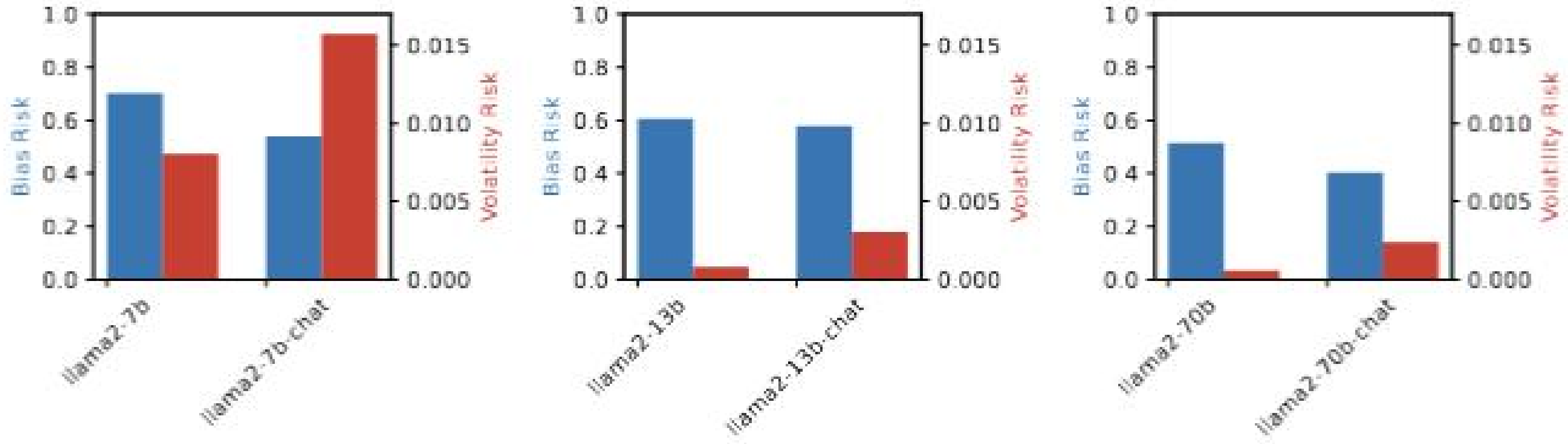
Impacts of Model Training Techniques on Bias Risk and Volatility Risk



Impact of Model Size

Larger models tend to show more bias but less volatility, implying they may overfit to biases in data while providing more consistent discriminatory patterns.

Impacts of Model Training Techniques on Bias Risk and Volatility Risk



Impact of RLHF

The chat versions refined with RLHF exhibit a lower bias risk compared to the base versions, yet they possess a higher volatility risk.

Conclusion

- We quantify the associated risk linked to the **stereotype distribution** inherent in LLMs. Furthermore, we decompose the total risk into two distinct components: the risk originating from persistent **bias** and the risk arising from **volatility** in stereotype representation.
- We applied our discrimination-measuring framework to 12 commonly used LLMs, leading to some intriguing findings. These include observations of pro-male bias, discrimination patterns within higher-income professions, and insights into how different model training techniques impact both bias risk and volatility risk.

Thank you!