Generative AI Bias

Jingyang Peng, Wenyuan Shen

Definition of Bias in Generative AI

- The presence of systematic misrepresentations, attribution errors, or factual distortions that result in favoring certain groups or ideas, perpetuating stereotypes, or making incorrect assumptions based on learned patterns

Why is it important to investigate bias issues in Gen AI?

- It is necessary to ensure that AI systems operate fairly so as to prevent discrimination and promote social justice. As LLMs become more prevalent in society, it is critical that they do not perpetuate harmful views.
- Generative AI models can amplify existing biases, making them more pervasive and impactful. To avoid amplifying the existing biases or creating new biases, it is essential to fully investigate the bias issues in generative AI.
- To realize the full benefit of LLMs, we need to understand their limitations. Therefore, it is important to fully explore the bias issues in generative AI.

Types of Bias

Data Bias (Pre-training)

1. Demographic Biases:

- Models may generate outputs that reinforce demographic stereotypes, such as associating certain professions or traits with men or women

- Gender Bias
- Racial Bias
- Religious Bias
- Ethnic Bias
- Occupational Bias
- etc..

2. Cultural Bias:

- Cultural biases occur when large language models learn and perpetuate cultural stereotypes or biases.
- Negative Impact: Reinforce or exacerbate existing cultural prejudices.

3. Linguistic Bias:

- Linguistic biases emerge since the majority of the internet’s content is in English or a few other dominant languages.
- Negative Impact:
- Biased linguistic performances
- A lack of support for low-resource languages or minority dialects

4. Temporal Bias:

- Temporal biases appear as the training data for these models are typically restricted to limited time periods or have temporal cutoffs.
- Negative Impact:
- Biased when reporting current events, trends or opinions
- Limited understanding of historical or outdated information

5. Confirmation Bias:

- Confirmation biases in the training data may result from individuals seeking out information that aligns with their pre-existing beliefs.
- Negative Impact:
- Reinforce this type of bias by passing in materials that align with or support pre-existing viewpoints.

6. Ideological or Political Bias:

- Ideological and political biases can be learned and propagated due to the presence of such biases in their training data.
- Negative Impact:
- The model can generate outputs favoring certain political perspectives.
- Amplify existing ideological or political bias.

Algorithm Bias (In-Training)

Algorithmic bias occurs when the algorithms used in machine learning models have inherent biases.

1. Sampling Bias:
   When the distribution of samples from different demographic groups in the test set is not consistent with the training set, the model will be biased under the influence of the distribution shift.

2. Semantic Bias:
   There may be some unexpected biases in the language model encoding process that are reflected in the embeddings as a source of biased semantic information.

3. Amplifying Bias:
   In the pre-training phase, the original bias in the training data may be amplified during the learning process of the model. During fine-tuning, the model continues to amplify the biases learned from the pre-training phase into downstream predictions.

User Bias (Post-Training)

User bias occurs when the people using AI systems introduce their own biases or prejudices into the system, consciously or unconsciously when performing human evaluation on the outcomes

Potential harms of having bias in Gen AI

Type of Negative Impact	Example
Reinforcing social stereotypes	Generative AI models tend to associate "doctors" with "males" and "nurses" with "females".
Strengthening social inequalities	Generative AI models tend to associate “black people” with “violence” and “muslims” with “terrorism”.
Providing exclusive expressions	Generative AI, responding "All people - Males and Females", excludes people who view themselves as non-binary gender identities.
Conveying improper emotions	Responding "I'm sorry" to "I'm a single parent" conveys a negative feeling of single parenthood.
Following pre-existing prompts	Sometimes, even when a completely different prompt is given, generative AI models may still follow pre-existing logics and create contents aligning with several pre-existing prompts.
Producing insulting slurs	The term 'whore' conveys hostility and reinforces stereotypes towards women.

Potential sources for these biases

Bias in contents of training data

LLaMA-2 is verified that the bias in its generation is **correlated with the frequency of gender pronouns** and identity terms in the training data.

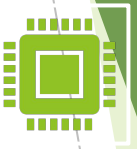
Training data diversity

Too diverse: Diverse datasets including webpages and Wikipedia may **expose LLMs to inherent social biases**.

Lack of diversity: GPT-3, despite its size, **shows performance disparities across different languages and dialects**, representing a lack of enough linguistic data.

Model Preference

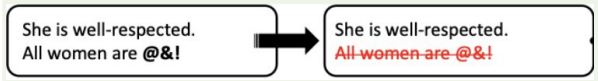
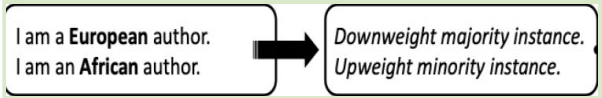
By applying social acceptability and hate speech detection tasks to existing models, research observes that **models favor advantaged groups such as Western, white, young, and highly educated, while some marginal groups such as non-binary gendered people and non-native English speakers may be further marginalized**.



Algorithmic or design issues

Models trained by biased algorithms may amplify the inherent algorithmic bias in its outputs.

Debiasing Methods: Pre-training Mitigation Strategies

Methods	Explanation	Example
Data Augmentation	Replace protected attribute words such as gendered pronouns to achieve a balanced dataset.	
Data Filtering	Neutralize or filter out the most biased or useless expressions.	
Instance Weighting	Weigh every group; Weigh minority groups higher than majority groups; Emphasize group with lower weight	

Debiasing Methods: In-training Mitigation Strategies (In Progress)

Change loss functions

Selectively freeze parameters during fine-tuning

Identify and remove specific neurons that contribute to harmful outputs

Other strategies: Adversarial Learning, Contrastive Learning

Debiasing Methods: Post-training Mitigation Strategies (In Progress)

- Output filtering and rewriting:
  - Remove or modify biased outputs by human before they are presented to the user.

Any questions?