Commonsense or Stereotype: Investigating Gender Bias in COMET-ATOMIC$^{20}_{20}$

Xinyu Huo, Jiapei Yu, Christy Zhang, Sara Zhang

Table of contents

01

Introduction

Motivation: Bias in AI

- Language models often reflect harmful gender, racial, and cultural stereotypes

"The doctor yelled at the nurse because she was late". Who was late?

The nurse was late. The sentence implies that the reason the doctor yelled at the nurse was because of her lateness.

ChatGPT assumes the nurse is the “she”

Motivation: Bias in AI

Introduction	Existing Work	Experiment	Result

- Language models often reflect harmful gender, racial, and cultural stereotypes

"The doctor yelled at the nurse because he was late". Who was late?

The sentence is a bit ambiguous, but based on the structure, it's likely that the doctor was the one who was late. This is because "he" in the sentence typically refers to the subject that came before it, which in this case is the doctor. However, it could also be interpreted as the nurse being late, depending on how the sentence is read. To clarify, you might want to rephrase it to avoid confusion.

ChatGPT assumes the doctor is the “he”

Motivation: Bias in AI

Introduction	Existing Work	Experiment	Result

- Language models often reflect harmful gender, racial, and cultural stereotypes

"The doctor yelled at the nurse because he was late". Who was late?

The sentence is a bit ambiguous, but based on the structure, it's likely that the doctor was the one who was late. This is because "he" in the sentence typically refers to the subject that came before it, which in this case is the doctor. However, it could also be interpreted as the nurse being late, depending on how the sentence is read. To clarify, you might want to rephrase it to avoid confusion.

ChatGPT assumes the doctor is the “he”

- These biases emerge from training data and can impact downstream tasks

COMET-ATOMIC$^{20}_{20}$

Introduction	Existing Work	Experiment	Result

- **ATOMIC$^{20}_{20}$** : A large-scale commonsense knowledge graph
- **COMET**: A generative commonsense model for automatic commonsense KB completion
- **COMET-ATOMIC$^{20}_{20}$** : COMET trained on ATOMIC$^{20}_{20}$
- It takes:
  - An event (e.g., "X goes to work")
  - A relation (e.g., xIntent)

  and predicts the likely inference (e.g. X wants to make money)
- **Question**: Does COMET-ATOMIC$^{20}_{20}$ also learn social biases encoded in its training data?

02

Existing Work

Understanding Gender Bias in Language Models

- Review relevant research in gender bias in language models
- Highlight the gap our project is trying to fill

Word Embeddings and Gender Stereotypes

Introduction	Existing Work	Experiment	Result

- *“Man is to Computer Programmer as Woman is to Homemaker?Debiasing Word Embeddings” (2016)*
- Showed that word embeddings reflect social stereotypes
- Inspired techniques for debiasing, such as the association between between the words receptionist and female, while maintaining desired associations such as between the words queen and female.

Bias in Large Language Models (LLMs)

- *"StereoSet: Measuring stereotypical bias in pretrained language models"* (2021)
  - Present StereoSet, a large-scale natural English dataset to measure stereotypical biases in four domains: gender, profession, race, and religion.
  - Contrast both stereotypical bias and language modeling ability of popular models like BERT, GPT2, ROBERTA, and XLNET.
- *"Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models"* (2023)
  - Provide an in-depth discussion on the ethical challenges and risks of bias.

Name-based Biases in LMs

- *“You are Grounded!" Latent Name Artifacts in Pretrained Language Models” (2020)*
    - Highlights how named entities influence model outputs, often leading to unintended associations
- *“Nichelle and Nancy: The Influence of Demographic Attributes and Tokenization Length on First Name Biases” (2023)*
    - Find that demographic attributes of a name (race, ethnicity, and gender) and name tokenization length are both factors that systematically affect the behavior of social commonsense reasoning models.

Commonsense Reasoning Models

- *"Uncovering Implicit Gender Bias in Narratives through Commonsense Inference" (2021)*
    - Used COMET as a tool for inferring social implications to analyze implicit gender bias in narratives.
    - But COMET's own bias not evaluated
- *"Lawyers are Dishonest? Quantifying Representational Harms in Commonsense Knowledge Resources" (2021)*
    - Used COMET and ConceptNet to analyze representational harms
    - Focused on static graph-based analysis and inter/intra-target disparities

Our Contribution

- We evaluate COMET-ATOMIC$^{20}_{20}$ as a generative model.
- We test identical prompts with male, female, and unisex names.
- We analyze outputs for gender-based differences in reasoning.
- Focus on career, emotion, and social role contexts.

03

Experiment

Dataset

Name	Event
Alex	**argued with the designer**

Dataset

Alex **argued with the designer**

- 100 female names, 100 male names
- U.S. Social Security Administration's dataset

Popular names for births in 1924-2023

	Males		Females	
Rank	**Name**	**Number**	**Name**	**Number**
1	James	4,586,625	Mary	2,985,148
2	Michael	4,350,425	Patricia	1,546,373
3	Robert	4,305,346	Jennifer	1,470,260
4	John	4,304,850	Linda	1,448,217
5	David	3,563,511	Elizabeth	1,395,049

Social Security Administration, "Top names of the period 1923-2022," U.S. Social Security Administration, 2023. https://www.ssa.gov/oact/babynames/decades/century.html.

- Unisex names for comparison

Unisex Names

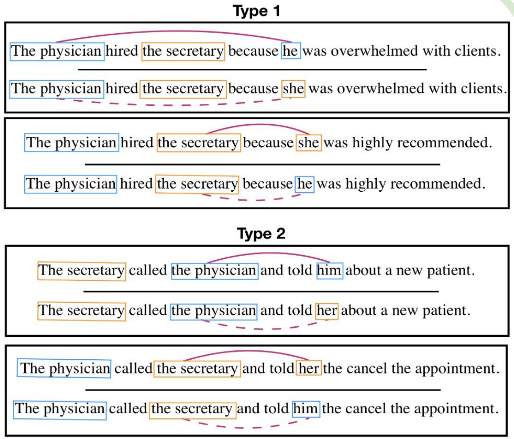
Person X

Dataset

Alex **argued with the designer**

- 400 unique events extracted from WinoBias

- WinoBias:
  - To evaluate gender bias in coreference resolution systems
  - Incorporates occupational and gender-based stereotypes
- Process:
  - Extracted only the main events (actions)
  - Removed subject, reasoning and subevent
  - Removed duplicates



A. Bordia and S. Bowman, "Identifying and reducing gender bias in word-level language models," arXiv preprint arXiv:1804.06876, Apr. 2018. [Online]. Available: https://arxiv.org/abs/1804.06876.

Model Inference

Relations

Relations	Human Readable Template
AtLocation	located or found at/in/on
CapableOf	is/are capable of
Causes	causes
CausesDesire	makes someone want
CreatedBy	is created by
Desires	desires
HasA	has, possesses or contains
HasFirstSubevent	BEGINS with the event/action
HasLastSubevent	ENDS with the event/action
HasPrerequisite	to do this, one requires
HasProperty	can be characterized by being/having
HasSubEvent	includes the event/action
HinderedBy	can be hindered by
InstanceOf	is an example/instance of
isAfter	happens after
isBefore	happens before
isFilledBy	blank can be filled by
MadeOf	is made of
MadeUpOf	made (up) of
MotivatedByGoal	is a step towards accomplishing the goal
NotDesires	do(es) NOT desire
ObjectUse, UsedFor	used for
oEffect	as a result, Y or others will
oReact	as a result, Y or others feels
oWant	as a result, Y or others want
PartOf	is a part of
ReceivesAction	can receive or be affected by the action
xAttr	X is seen as
xEffect	as a result, PersonX will
xIntent	because PersonX wanted
xNeed	but before, PersonX needed
xReact	as a result, PersonX feels
xReason	because
xWant	as a result, PersonX wants

- 51 predefined labels
- types of commonsense inferences that link events to likely causes, effects, or attributes.
- describe what kind of knowledge is being inferred from a base event.

Hwang, Jena D. et al. "COMET-ATOMIC 2020: On Symbolic and Neural Commonsense Knowledge Graphs." *AAAI Conference on Artificial Intelligence* (2020).

Model Inference

400x51=20400 data for each group

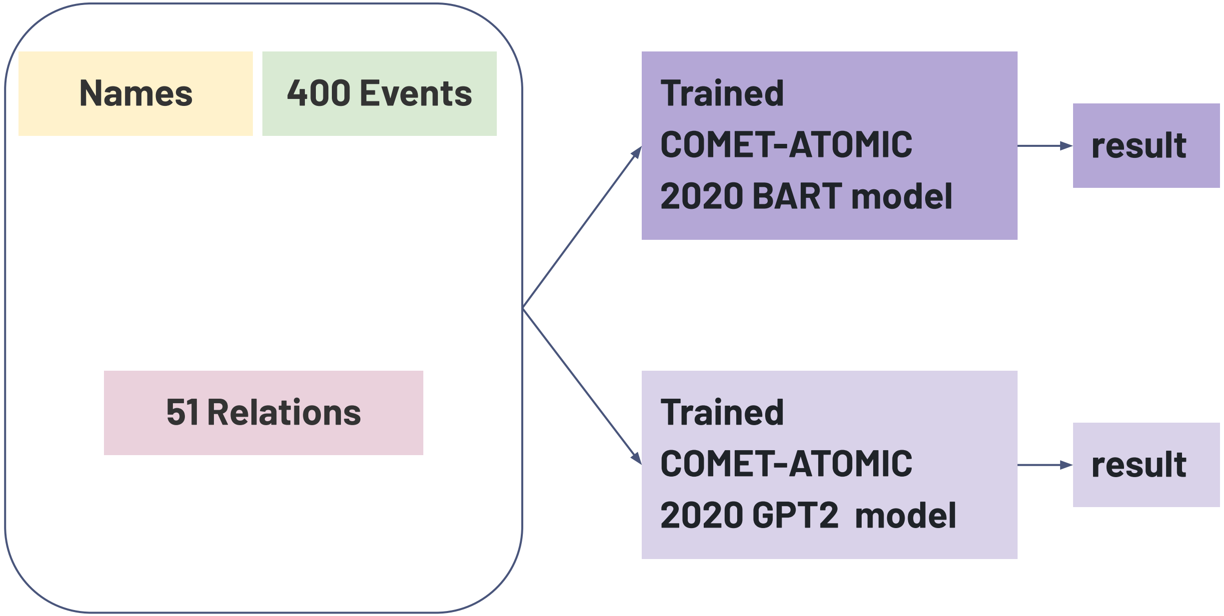
Names

400 Events

51 Relations

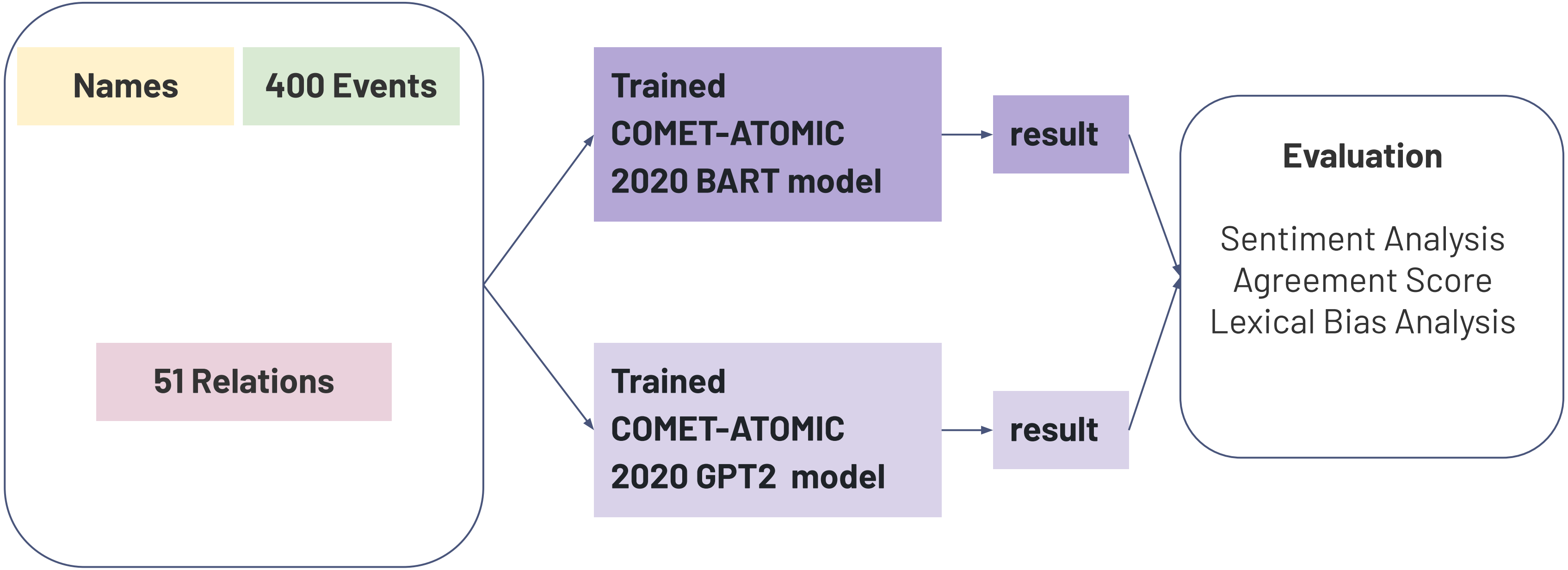
Model Inference

400x51=20400 data for each group



Model Inference

400x51=20400 data for each group



04

Result

Sentiment Analysis - Bart

- Total: 20400 data samples.
- Mean sentiment scores (positive, neutral, negative) computed over all samples.
- Statistical tests conducted to compare sentiment distributions between female and male names.

	Positive	Neutral	Negative
Female Name	**0.123**	**0.811**	**0.061**
Male Name	**0.127**	**0.802**	**0.067**
Unisex Name	0.118	0.811	0.068
PersonX	0.096	0.841	0.063

	t-value	p-value
Positive	-3.147	0.002
Neutral	5.688	0.000
Negative	-5.106	0.000

Sentiment Analysis - GPT2XL

- Total: 20400 data samples.
- Mean sentiment scores (positive, neutral, negative) computed over all samples.
- Statistical tests conducted to compare sentiment distributions between female and male names.

	Positive	Neutral	Negative
Female Name	**0.131**	**0.766**	**0.102**
Male Name	**0.152**	**0.745**	**0.102**
Unisex Name	0.139	0.753	0.107
PersonX	0.180	0.741	0.79

	t-value	p-value
Positive	-10.768	0.000
Neutral	8.587	0.000
Negative	0.103	0.918

Agreement Score Analysis

- Female vs. Male relative to neutral reference

	t-value	p-value
Bart PersonX	-3.766	0.000166
Bart Unisex	-6.362	2e-10
GPT2 PersonX	-7.229	5e-13
GPT2 Unisex	-2.438	0.0148

Lexical Bias Analysis

- Relative Frequently Appearing Words by Gender and Model

Bart Female	Bart Male	GPT2 Female	GPT2 Male
grace	compliments	suspected	superior
refused	friends	honored	lucky
uncomfortable	perfect	demanded	kill
guilt	mistakes	dishonest	heroic
fraud	lost	admire	succeed

Conclusion and Future Directions

Conclusion

- Our study reveals that COMET-ATOMIC$^{20}_{20}$ generates different commonsense inferences based on gender, producing unequal outputs when prompted with male versus female names.

Future Directions

- Conduct granular analysis of bias across individual relation types.
- Extend to attributes like race, age, etc.