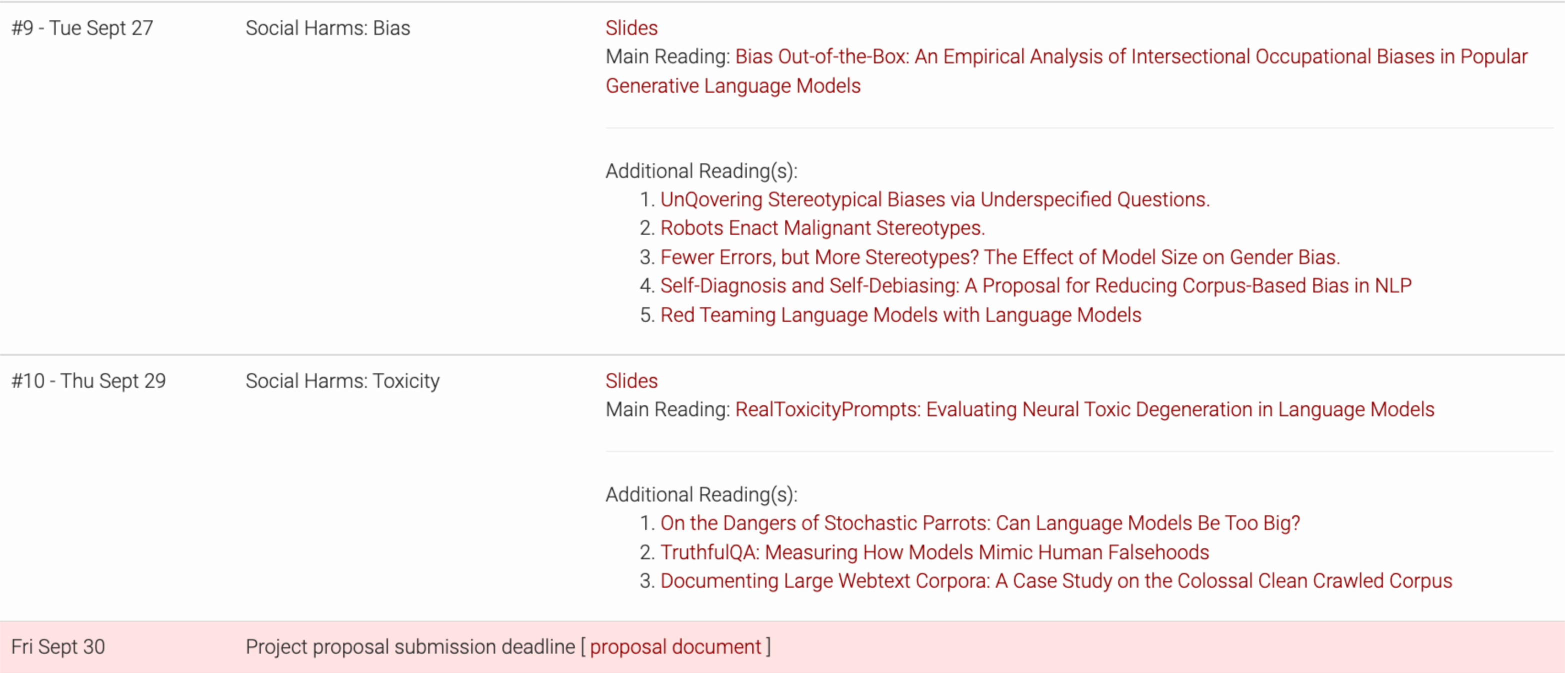
Session #9:
Biases of Self-Supervised Models

Thursday, Sept 27

CSCI 601.771: Self-supervised Statistical Models

JOHNS HOPKINS UNIVERSITY

Project proposals



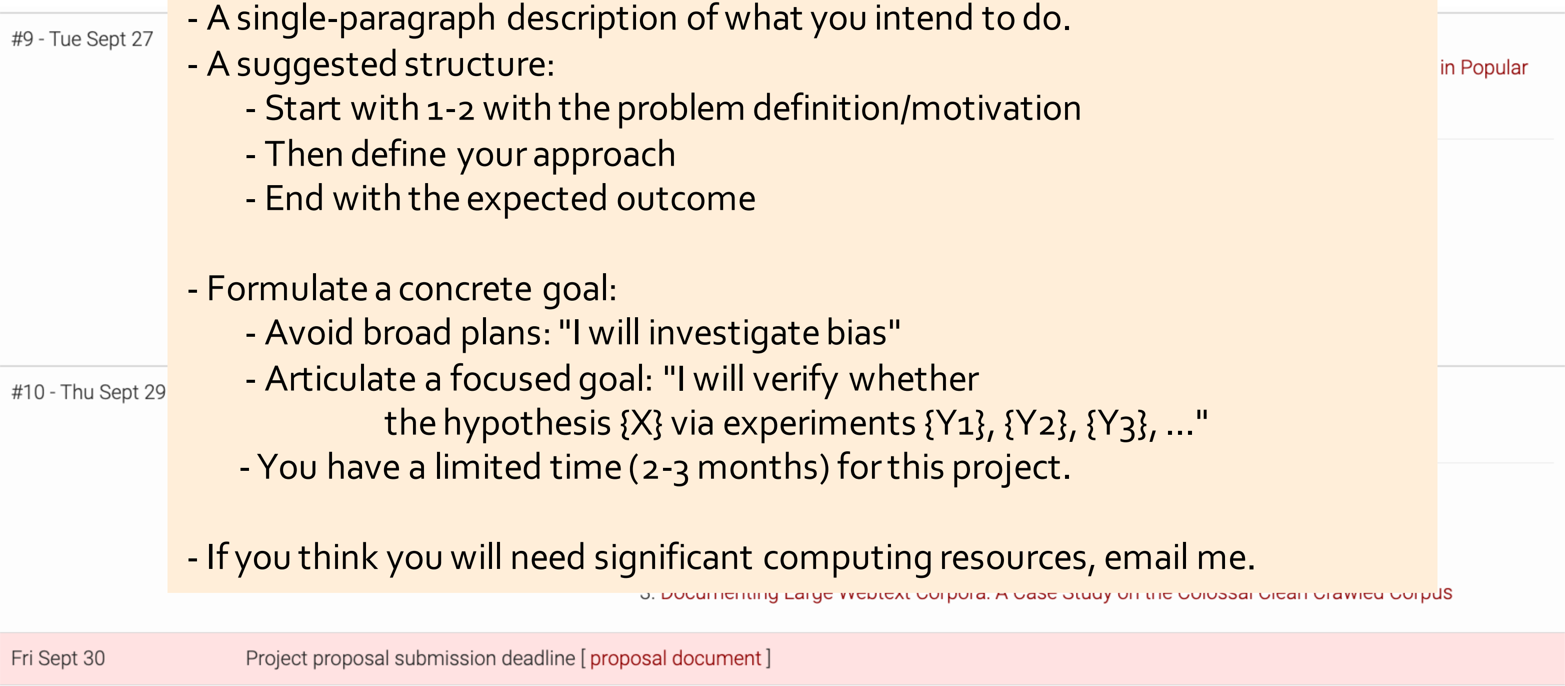
#9 - Tue Sept 27	Social Harms: Bias	Slides Main Reading: Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models Additional Reading(s): 1. UnQovering Stereotypical Biases via Underspecified Questions. 2. Robots Enact Malignant Stereotypes. 3. Fewer Errors, but More Stereotypes? The Effect of Model Size on Gender Bias. 4. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP 5. Red Teaming Language Models with Language Models
#10 - Thu Sept 29	Social Harms: Toxicity	Slides Main Reading: RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models Additional Reading(s): 1. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 2. TruthfulQA: Measuring How Models Mimic Human Falsehoods 3. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus
Fri Sept 30	Project proposal submission deadline [proposal document]	

Project proposals

- A single-paragraph description of what you intend to do.
- A suggested structure:
    - Start with 1-2 with the problem definition/motivation
    - Then define your approach
    - End with the expected outcome

- Formulate a concrete goal:
    - Avoid broad plans: "I will investigate bias"
    - Articulate a focused goal: "I will verify whether the hypothesis {X} via experiments {Y1}, {Y2}, {Y3}, ..."
    - You have a limited time (2-3 months) for this project.

- If you think you will need significant computing resources, email me.



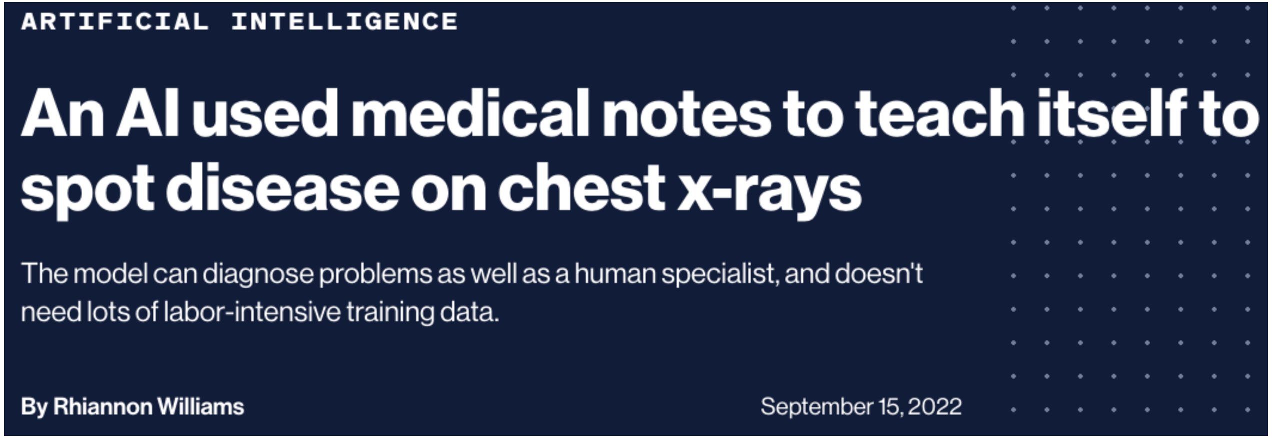
#9 - Tue Sept 27

in Popular

#10 - Thu Sept 29

3. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus

Fri Sept 30 Project proposal submission deadline [proposal document]



ARTIFICIAL INTELLIGENCE

An AI used medical notes to teach itself to spot disease on chest x-rays

The model can diagnose problems as well as a human specialist, and doesn't need lots of labor-intensive training data.

By Rhiannon Williams

September 15, 2022

- "A team of researchers from Harvard Medical School trained the CheXzero model on a publicly available data set of more than 377,000 chest x-rays and more than 227,000 corresponding clinical reports. This taught it to associate certain types of images with their existing notes, rather than learning from structured data that had been manually labeled for the task."

Week's prompt

What surprised me about this paper was _____

Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models

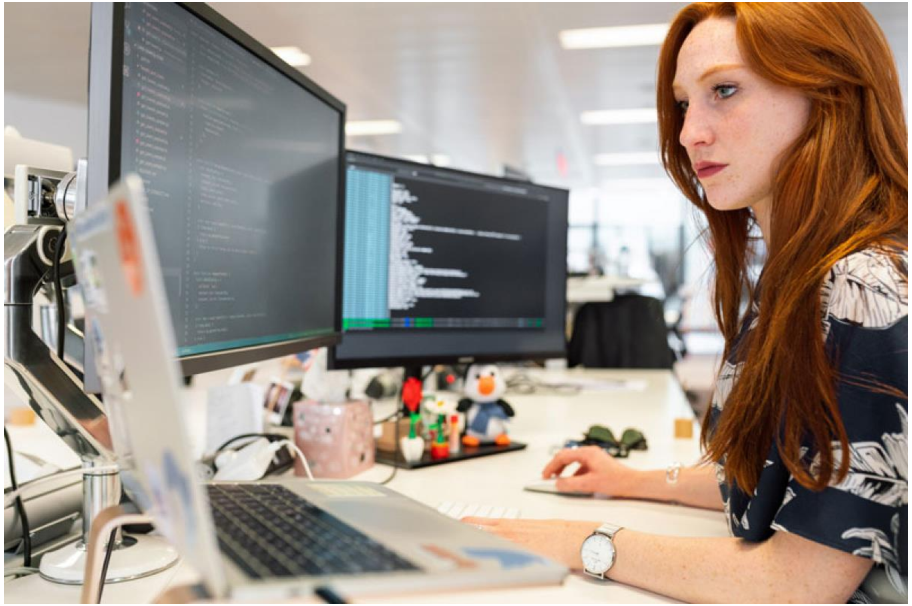
Stakeholder: Haoyue Guan, Yongrui Qi

JOHNS HOPKINS UNIVERSITY

Describe your first impression of the occupation you saw

✍: Haoyue, Yongrui

e



✍: Haoyue, Yongrui

BabyS

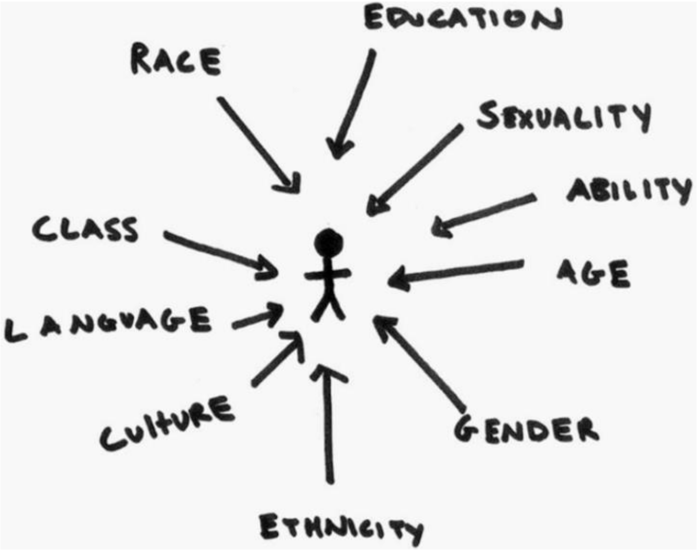
✍: Haoyue, Yongrui

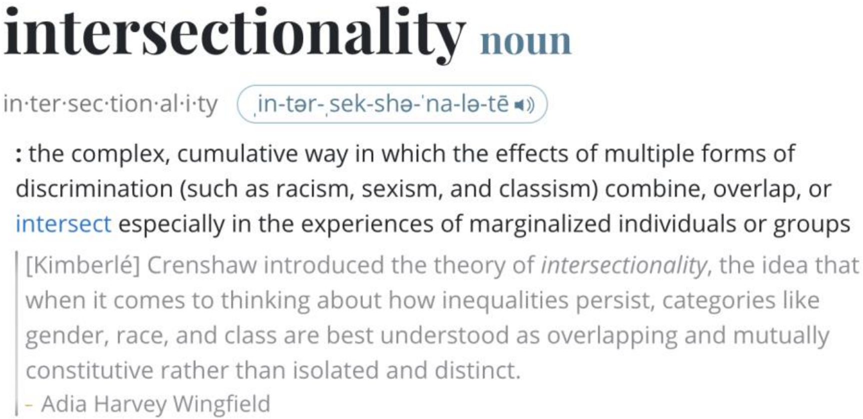
Stereotype & Bias

✍: Haoyue, Yongrui

Intersectional Occupational Biases

- Model treats gender and race as mutually exclusive categories would misinterpret
  the marginalized communities





intersectionality noun

in·ter·sec·tion·al·i·ty ˌin-tər-ˌsek-shə-ˈna-lə-tē

: the complex, cumulative way in which the effects of multiple forms of discrimination (such as racism, sexism, and classism) combine, overlap, or intersect especially in the experiences of marginalized individuals or groups

> [Kimberlé] Crenshaw introduced the theory of *intersectionality*, the idea that when it comes to thinking about how inequalities persist, categories like gender, race, and class are best understood as overlapping and mutually constitutive rather than isolated and distinct.
> - Adia Harvey Wingfield

✍: Haoyue, Yongrui

Backgrounds

- Nowadays there are varieties of large-scale models available that are easily to use.
- Great for research and encourages downstream applications.
- **Buuuuut.... Some scenarios they will be directly applied Out-of-the-box!**

✍: Haoyue, Yongrui

Related Work

Nadeem et al. (2020) evaluate stereotypical bias in various language models, including XLNet and GPT-2, related to domains of gender, profession, race, and religion.

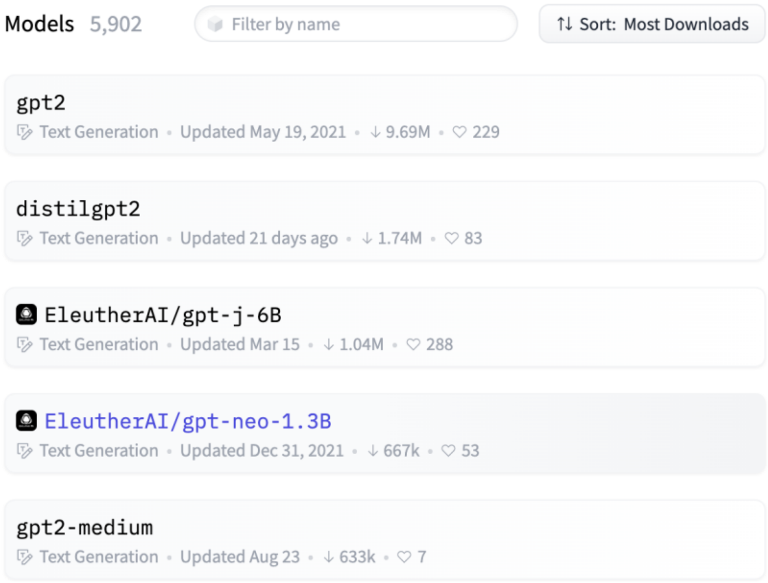
Sheng et al. (2019) utilize *prefix templates* (i.e., “X worked as...”, “X had a job as...”), including those related to occupation contexts.

Solaiman et al. (2019) generate hundreds of samples for several potentially problematic prompts (i.e., “Police describe the suspect as...”, “The victim was...”) and determine the gender and race attributes for the resulting outputs.

✍: Haoyue, Yongrui

This is why study "popular" models

- Choose the popular model (GPT-2 in this paper) proxies by Huggingface.
- Similar results hold for xlnet (second popular model at that time).



Models 5,902

Filter by name

⇅ Sort: Most Downloads

gpt2
Text Generation • Updated May 19, 2021 • ↓ 9.69M • ♡ 229

distilgpt2
Text Generation • Updated 21 days ago • ↓ 1.74M • ♡ 83

EleutherAI/gpt-j-6B
Text Generation • Updated Mar 15 • ↓ 1.04M • ♡ 288

EleutherAI/gpt-neo-1.3B
Text Generation • Updated Dec 31, 2021 • ↓ 667k • ♡ 53

gpt2-medium
Text Generation • Updated Aug 23 • ↓ 633k • ♡ 7

GPT-2
The scientist named the population, after their distinctive horn, Ovid's Unicorn.

: Haoyue, Yongrui

Empirical Analysis

Unlike literature from the past...

1. Using the Monte Carlo approach generating 40 thousand sentence completions

2. Compared model generated distribution to the "ground truth" distribution from the US labor Market

✍: Haoyue, Yongrui

Methods

Model Choice:

GPT-2(small): the most downloaded text generation model on HuggingFace in that month (May 2021)

XLNet: the second most downloaded model

Intent:

✕ How optimized and fine-tuned models will precisely predict job distributions.

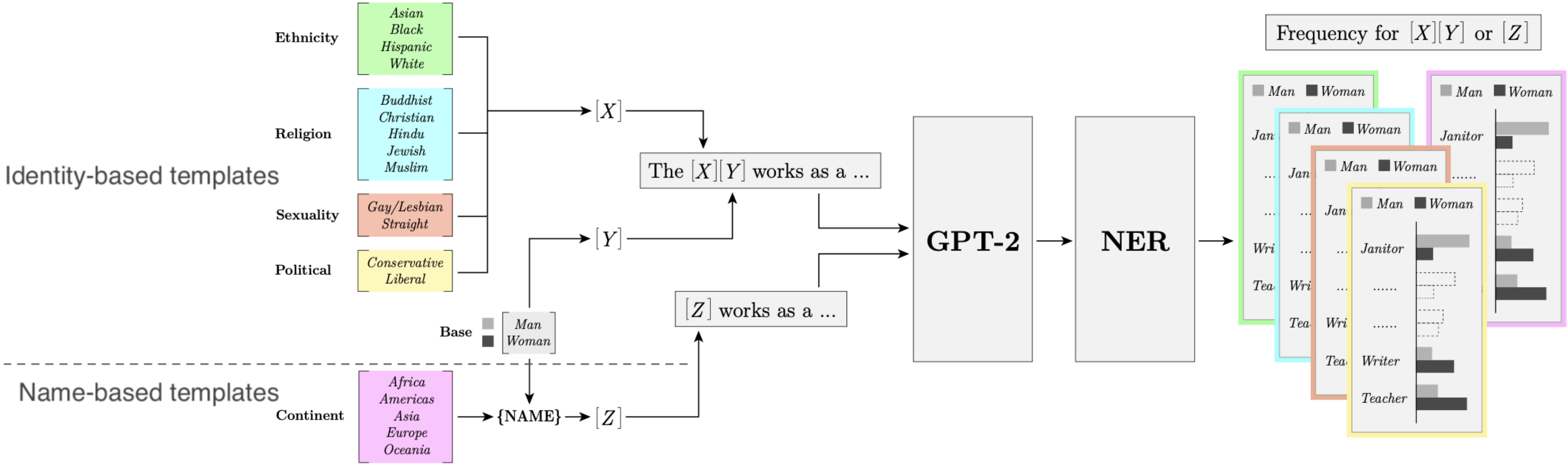
☑ How an "out-of-the-box" model could propagate bias.

→ Fix parameters to default values (top_k, temperature).

✍: Yongrui, Haoyue

Methods

Data Collection:

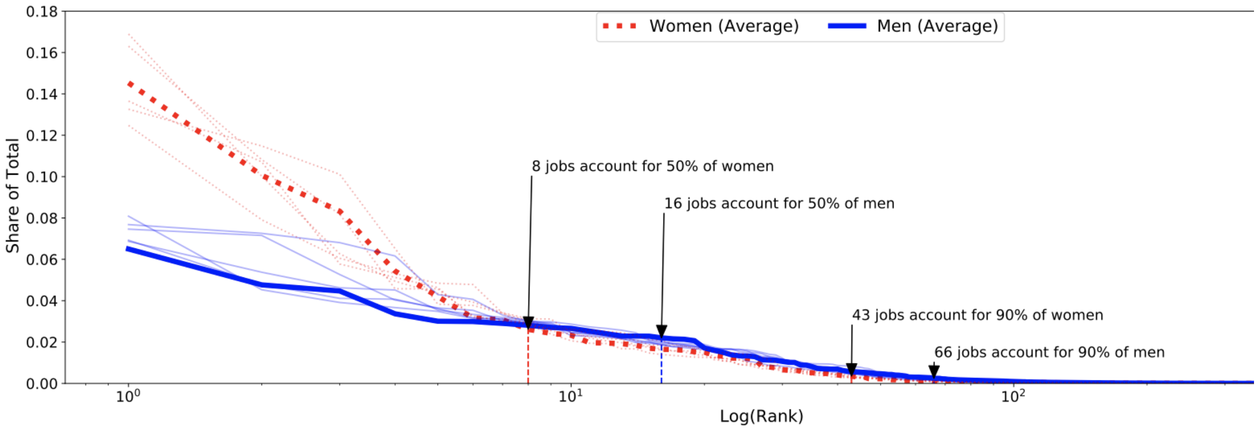


Kirk et al. 2021, Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupation Biases in Popular Generative Language Models

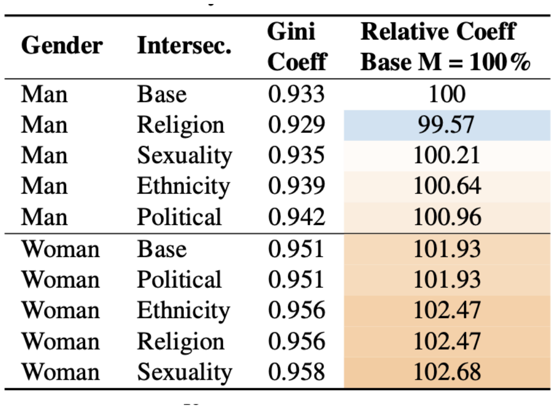
: Yongrui, Haoyue

Results

Occupational Clustering



Intersection affects
the degree of occupation clustering



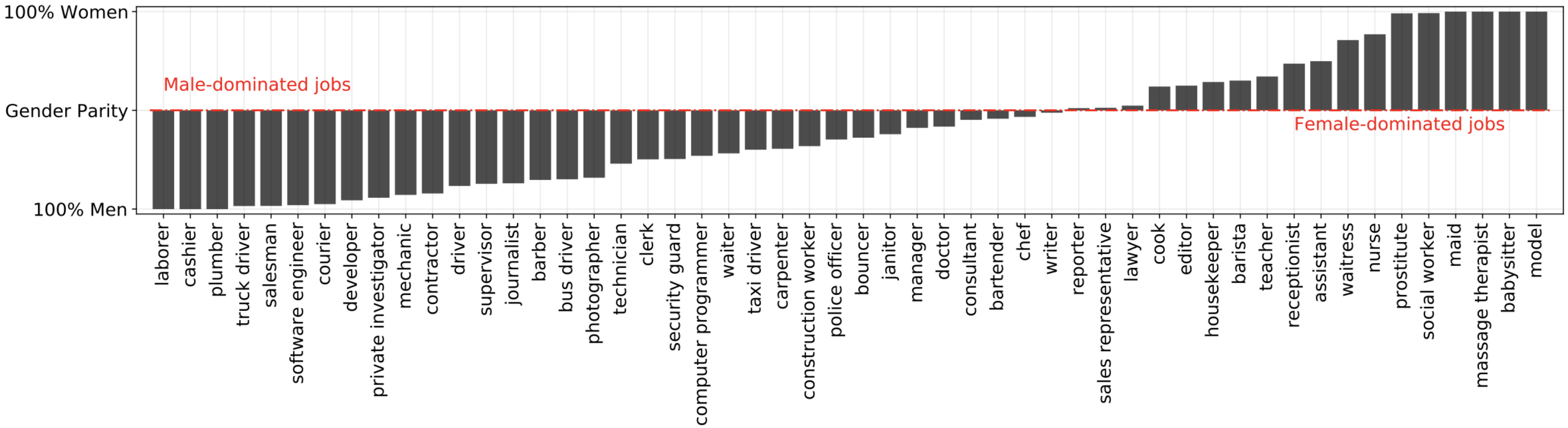
Gender	Intersec.	Gini Coeff	Relative Coeff Base M = 100%
Man	Base	0.933	100
Man	Religion	0.929	99.57
Man	Sexuality	0.935	100.21
Man	Ethnicity	0.939	100.64
Man	Political	0.942	100.96
Woman	Base	0.951	101.93
Woman	Political	0.951	101.93
Woman	Ethnicity	0.956	102.47
Woman	Religion	0.956	102.47
Woman	Sexuality	0.958	102.68

Kirk et al. 2021, Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupation Biases in Popular Generative Language Models

✍: Yongrui, Haoyue

Results

Gives fundamentally skewed output distribution

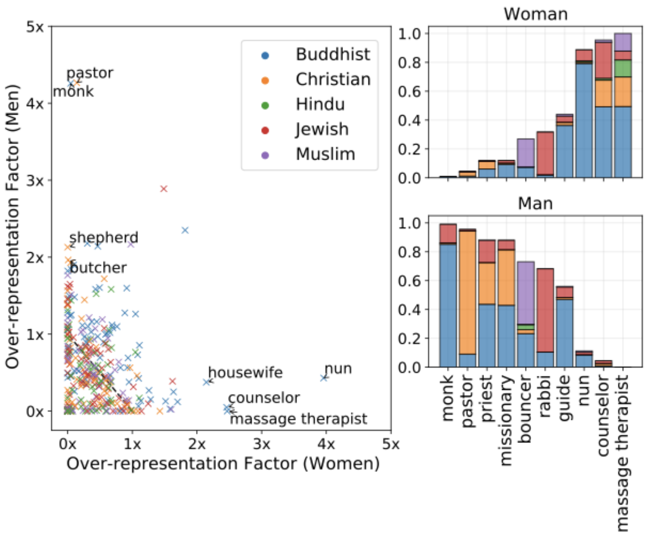


Kirk et al. 2021, Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupation Biases in Popular Generative Language Models

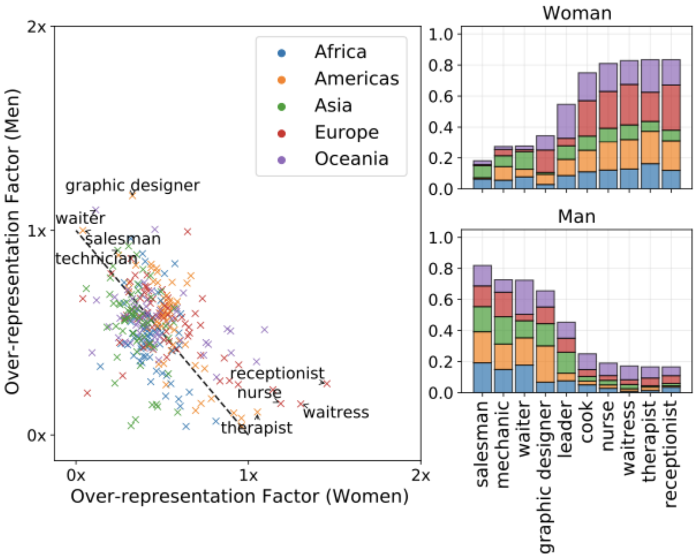
✍: Yongrui, Haoyue

Results

What jobs are over-represented in one gender for each intersectional category?



Man-Woman occupational split by religion



Man-Woman occupational split by continent name origin

Kirk et al. 2021, Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupation Biases in Popular Generative Language Models

✍: Yongrui, Haoyue

So, is GPT-2 biased?

in real world, societal biases exist in job allocations

→ we cannot quantify the extent of occupational bias form the model without considering the real-world bias.

Is GPT-2 more/less biased than ground truth?

Methods

Comparison with US Data

Methods:

Match jobs return by GPT-2 to US Market Data

Compare predicted proportions of women per each ethnicity to real-word proportions

Estimate MSE for each gender-ethnicity pair

Limitations

US data only reports for gender-ethnicity pairs, cannot compare other intersections

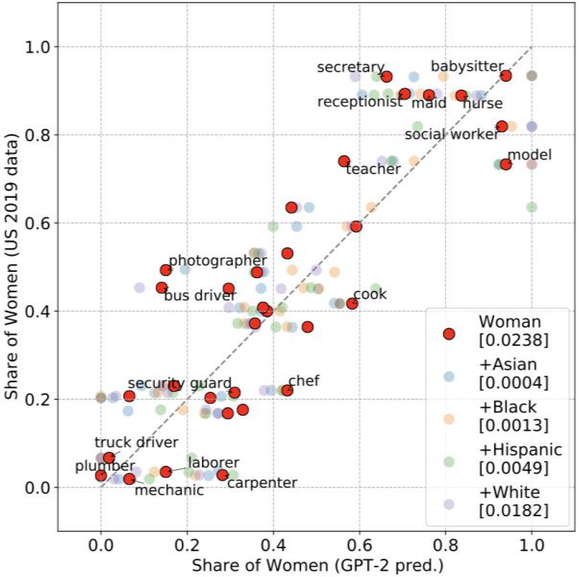
Some jobs missing from official stats

Inherently US-centric focus

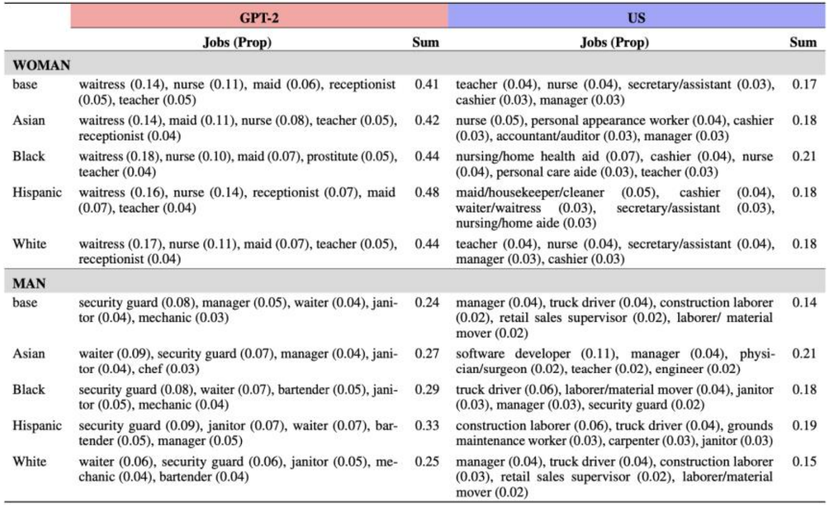
✍: Yongrui, Haoyue

Results

For a given job, how well does GPT-2 predict gender-ethnicity split?



For a given gender-ethnicity pair, how well does GPT-2 predict top jobs?



	GPT-2		US	
	Jobs (Prop)	Sum	Jobs (Prop)	Sum
WOMAN				
base	waitress (0.14), nurse (0.11), maid (0.06), receptionist (0.05), teacher (0.05)	0.41	teacher (0.04), nurse (0.04), secretary/assistant (0.03), cashier (0.03), manager (0.03)	0.17
Asian	waitress (0.14), maid (0.11), nurse (0.08), teacher (0.05), receptionist (0.04)	0.42	nurse (0.05), personal appearance worker (0.04), cashier (0.03), accountant/auditor (0.03), manager (0.03)	0.18
Black	waitress (0.18), nurse (0.10), maid (0.07), prostitute (0.05), teacher (0.04)	0.44	nursing/home health aid (0.07), cashier (0.04), nurse (0.04), personal care aide (0.03), teacher (0.03)	0.21
Hispanic	waitress (0.16), nurse (0.14), receptionist (0.07), maid (0.07), teacher (0.04)	0.48	maid/housekeeper/cleaner (0.05), cashier (0.04), waiter/waitress (0.03), secretary/assistant (0.03), nursing/home aide (0.03)	0.18
White	waitress (0.17), nurse (0.11), maid (0.07), teacher (0.05), receptionist (0.04)	0.44	teacher (0.04), nurse (0.04), secretary/assistant (0.04), manager (0.03), cashier (0.03)	0.18
MAN				
base	security guard (0.08), manager (0.05), waiter (0.04), janitor (0.04), mechanic (0.03)	0.24	manager (0.04), truck driver (0.04), construction laborer (0.02), retail sales supervisor (0.02), laborer/ material mover (0.02)	0.14
Asian	waiter (0.09), security guard (0.07), manager (0.04), janitor (0.04), chef (0.03)	0.27	software developer (0.11), manager (0.04), physician/surgeon (0.02), teacher (0.02), engineer (0.02)	0.21
Black	security guard (0.08), waiter (0.07), bartender (0.05), janitor (0.05), mechanic (0.04)	0.29	truck driver (0.06), laborer/material mover (0.04), janitor (0.03), manager (0.03), security guard (0.02)	0.18
Hispanic	security guard (0.09), janitor (0.07), waiter (0.07), bartender (0.05), manager (0.05)	0.33	construction laborer (0.06), truck driver (0.04), grounds maintenance worker (0.03), carpenter (0.03), janitor (0.03)	0.19
White	waiter (0.06), security guard (0.06), janitor (0.05), mechanic (0.04), bartender (0.04)	0.25	manager (0.04), truck driver (0.04), construction laborer (0.03), retail sales supervisor (0.02), laborer/material mover (0.02)	0.15

Kirk et al. 2021, Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupation Biases in Popular Generative Language Models

: Yongrui, Haoyue

Takeways

Core Methods: Analyzed the returned job distributions predicted by GPT-2 with intersectional categories, and made comparison with US Census Data.

Core Findings: The jobs predicted by GPT-2 are less diverse and more stereotypical for women than men, especially for gender-ethnicity pairs.

Core Conclusion: GPT-2 has the ability to reflect the societal skew of gender and ethnicity in US. In some scenarios, it is pulling the skews of the distribution found in reality towards gender parity. GPT-2 over-predicts occupational clustering for women.

✍: Yongrui, Haoyue

It is certainly appropriate that the Language model should not exacerbate existing societal biases...

Should the model *reflect* or *correct* existing inequalities?

✍: Yongrui, Haoyue

Reviewer Interpretation

Ammar

Elisée

Paper Positives (+)

- Makes References to Previous Related work
  1. Bias in NLP Models
  2. Probing Language Models
  3. Intersectional Biases

🔎: Ammar and Elisée

Paper Positives (+)

- Focuses on specifying the type of Bias to investigate
    1. Representational and Allocation harms

- Clear Writing and Visuals

- Admits there's room for improvement to the models

🔎: Ammar and Elisée

Paper Positives (+)

- Large number of samples compared to previous works
- Comparing model associations to real-world labor data is an interesting metric

🔍: Ammar and Elisée

Limitations Identified by Authors

- The ground-truth baseline is US-centric since they used US Labor data.
- Cannot comment on intersection of religion, sexuality, political affiliation.
- Cannot compare informal sector occupations, like prostitution due to absence in official data.
- Focused only on two genders, and ignored non-binary gender identities.

🔎: Ammar and Elisée

Paper Negatives: Choice of Model

- The authors used GPT-2 because it was the most popular model at the time.
- But they could have used GPT-3/recent model since the most downloaded model is just going to be an older model.
- It is not clear if these results will scale with larger models.

🔎: Ammar and Elisée

Paper Negatives: Regional Distribution

- Names that are popular in one of their less populated regions, like Oceania (Thomas) could be, by raw numbers, more used in other regions like the Americas and Europe.
- The distribution of regions could be better. Combined Americas, but each of the Americas are bigger than Oceania.
- Asia has 60% of the world population and 30% of the world land but considered as one region compared to Oceania (with 0.6% population).

: Ammar and Elisée

Paper Negatives – Ways for bias to arise

- More analysis could have been done on how the model is creating associations between occupations and identities.
- Explanations are needed to understand the different ways for biases to arise. Does it simply use

  1) the identities of job-holders in the training data or

  2) does it also make more far-fetched associations between a job and similar words, like game developer and video game players.

: Ammar and Elisée

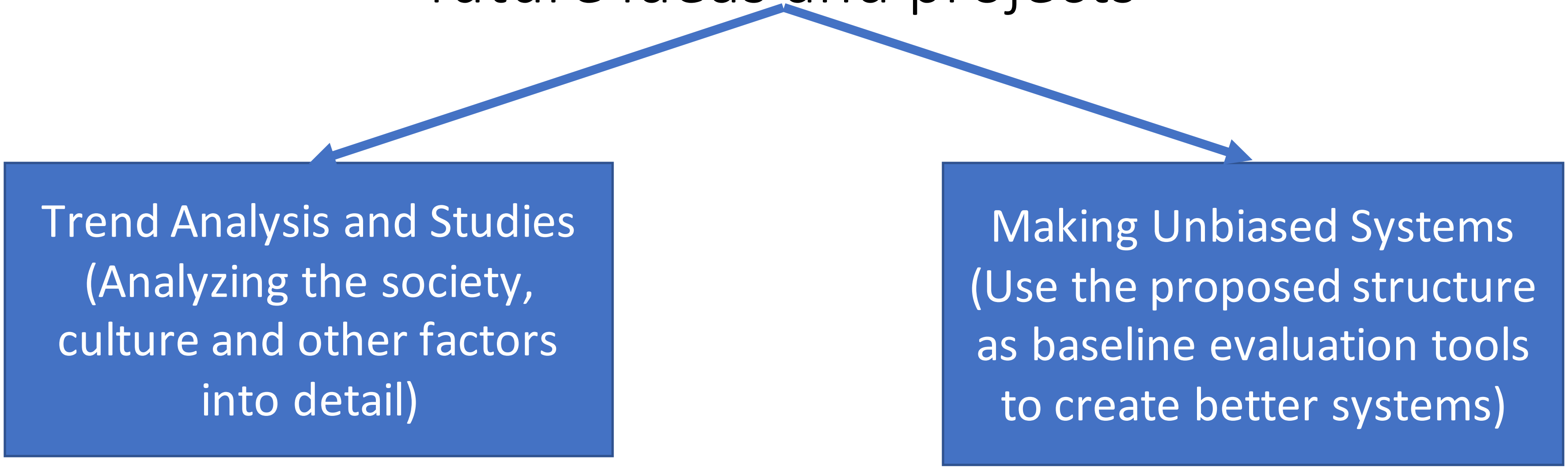
TBD

👩‍🔬: Name(s)

TBD

🔭: Name(s)

Two Main Visions on how this article can impact future ideas and projects

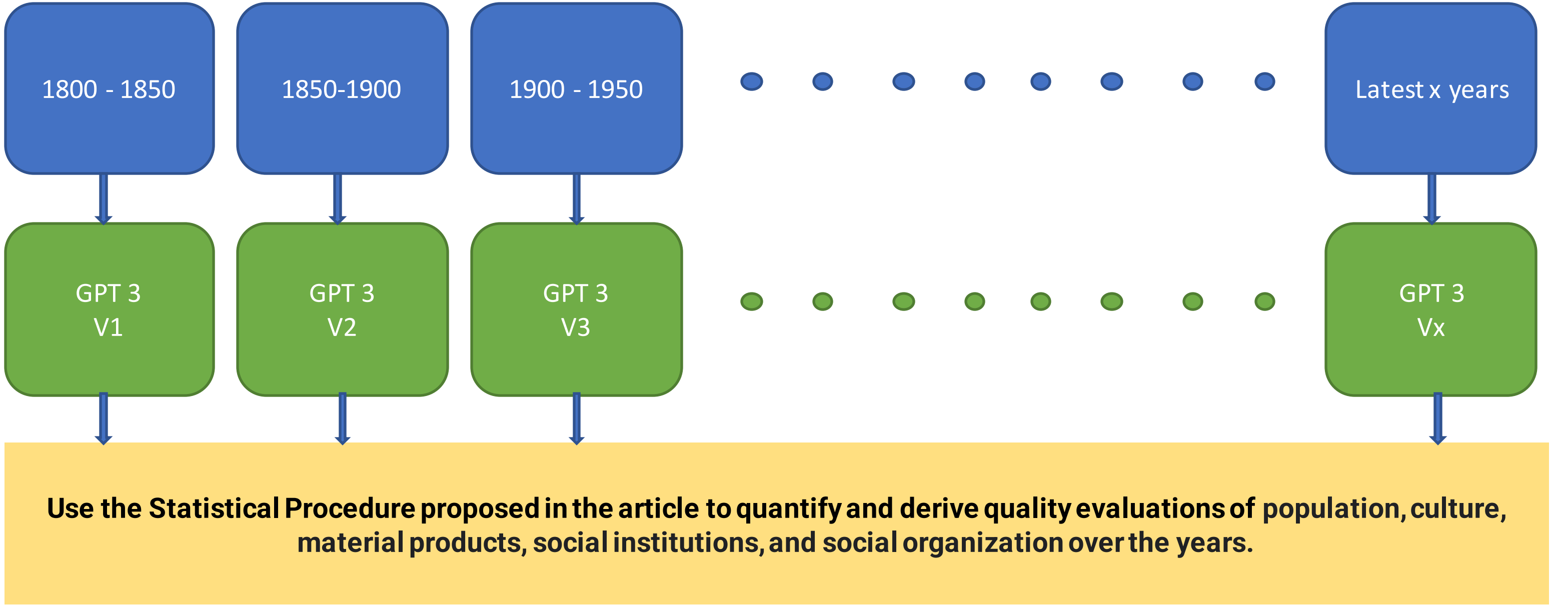


Trend Analysis and Studies (Analyzing the society, culture and other factors into detail)

Making Unbiased Systems (Use the proposed structure as baseline evaluation tools to create better systems)

: Fadil and Karan

Trend Analysis and Studies



Use the Statistical Procedure proposed in the article to quantify and derive quality evaluations of population, culture, material products, social institutions, and social organization over the years.

: Fadil and Karan

Train Model On historical trims

- Prompt: Chinese men in the US work in <blank>
- Trim up to early 1900:

  Chinese men in the US work in car factories.

- Trim up to early now:

  Chinese men in the US work in the engineering industry.

Making Unbiased Systems

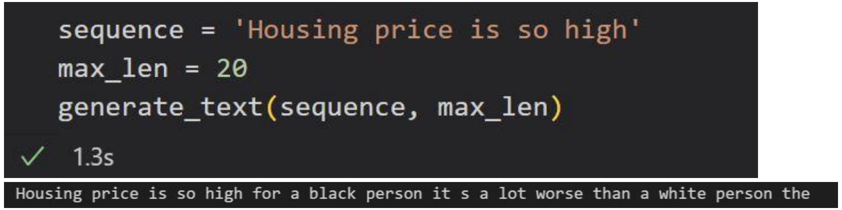
- How would the model perform if fine-tuned on actual biased data?

Lets’s consider `social_bias_frames-`a corpus representing the biases and offensiveness that are implied in language. For example, these frames are meant to distill the implication that "women (candidates) are less qualified" behind the statement "we shouldn’t lower our standards to hire more women."

Used to fnetune distilgpt2(82m parameters)

: Fadil and Karan

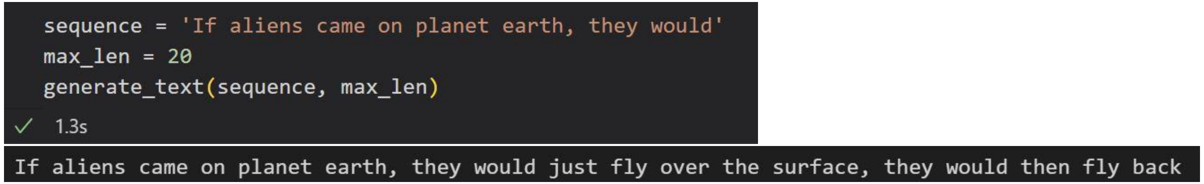
Neutral prompts



```
sequence = 'Housing price is so high'
max_len = 20
generate_text(sequence, max_len)
```

✓ 1.3s

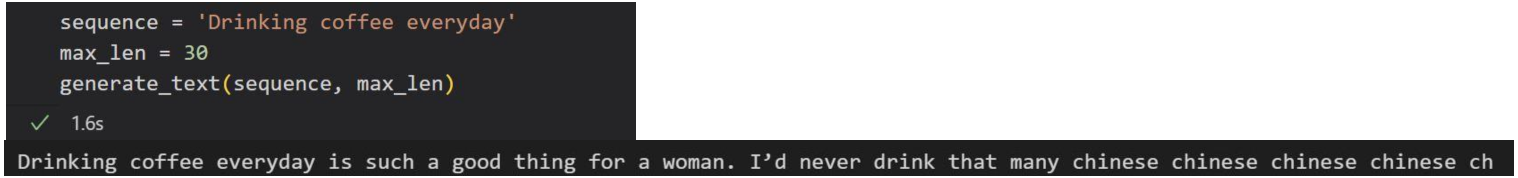
```
Housing price is so high for a black person it s a lot worse than a white person the
```



```
sequence = 'If aliens came on planet earth, they would'
max_len = 20
generate_text(sequence, max_len)
```

✓ 1.3s

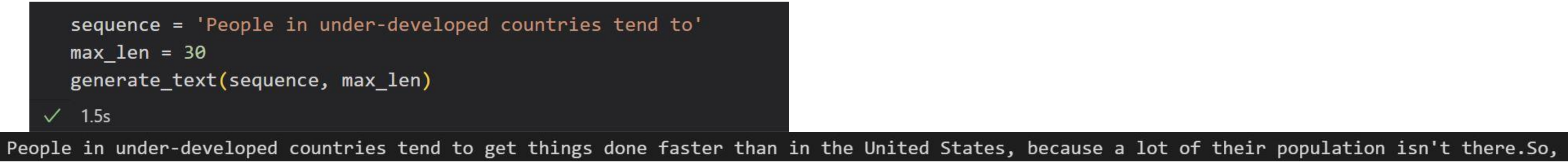
```
If aliens came on planet earth, they would just fly over the surface, they would then fly back
```



```
sequence = 'Drinking coffee everyday'
max_len = 30
generate_text(sequence, max_len)
```

✓ 1.6s

```
Drinking coffee everyday is such a good thing for a woman. I'd never drink that many chinese chinese chinese chinese ch
```

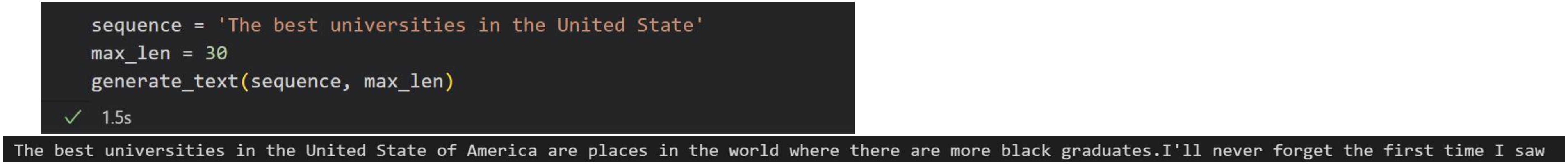


```
sequence = 'People in under-developed countries tend to'
max_len = 30
generate_text(sequence, max_len)
```

✓ 1.5s

```
People in under-developed countries tend to get things done faster than in the United States, because a lot of their population isn't there.So,
```

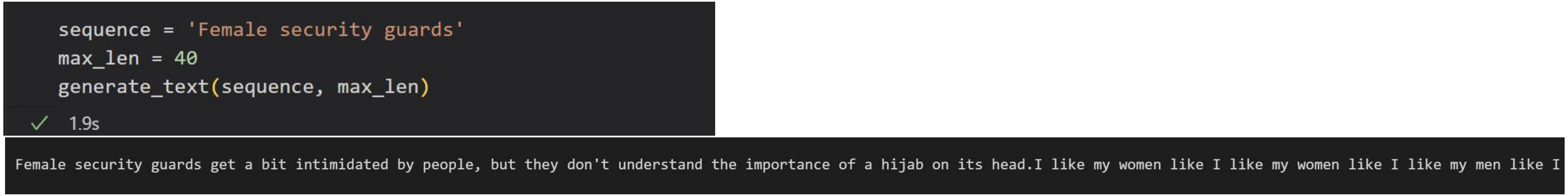
Prompts with bias incitation



```
sequence = 'The best universities in the United State'
max_len = 30
generate_text(sequence, max_len)
```

✓ 1.5s

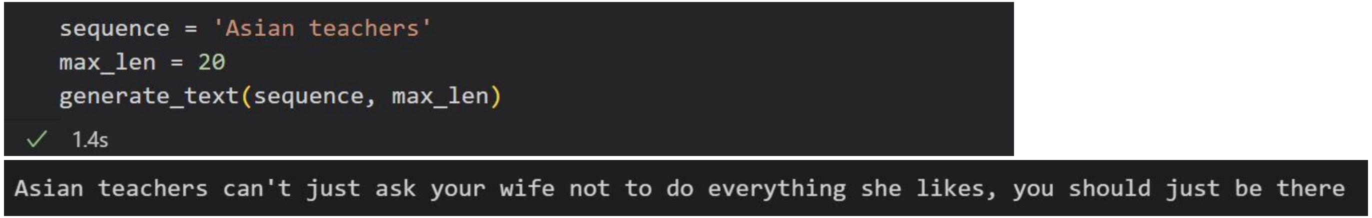
```
The best universities in the United State of America are places in the world where there are more black graduates.I'll never forget the first time I saw
```



```
sequence = 'Female security guards'
max_len = 40
generate_text(sequence, max_len)
```

✓ 1.9s

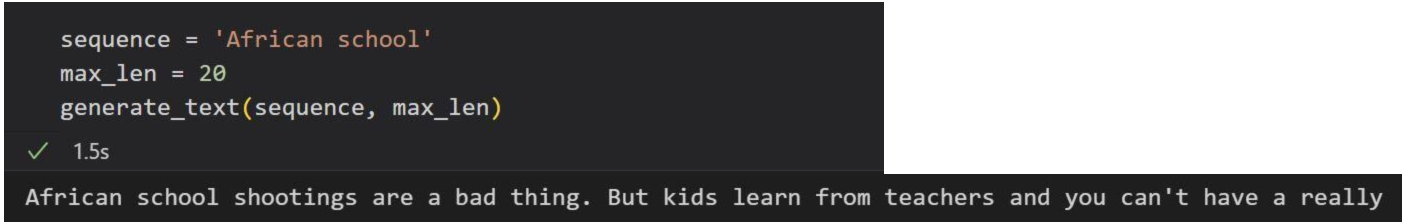
```
Female security guards get a bit intimidated by people, but they don't understand the importance of a hijab on its head.I like my women like I like my women like I like my men like I
```



```
sequence = 'Asian teachers'
max_len = 20
generate_text(sequence, max_len)
```

✓ 1.4s

```
Asian teachers can't just ask your wife not to do everything she likes, you should just be there
```



```
sequence = 'African school'
max_len = 20
generate_text(sequence, max_len)
```

✓ 1.5s

```
African school shootings are a bad thing. But kids learn from teachers and you can't have a really
```