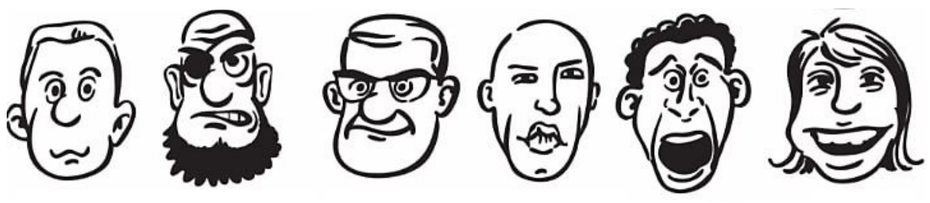
CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations

Myra Cheng, Tiziano Piccardi, Diyi Yang

Stanford | NLP

Caricatures individuate the subject from others and exaggerate the subject's features (Perkins, 1975). They reify stereotypes & foster homogenous narratives.



"Is my LLM simulation a caricature?"

- Using LLMs to simulate human behaviour may instead perpetuate essentializing narratives and stereotypes.
- We propose a metric for LLM simulations' susceptibility to caricature.
- We measure simulations on GPT-4 in the contexts of (1) online forum and (2) interview.

Taxonomizing LLM Simulations

Context	Where and when does the simulated scenario occur?
Model	What LLM is used?
Persona	Whose opinion/action is simulated?
Topic	What is the simulation about?

⇒ When do LLM simulations individuate & exaggerate persona?

Examples of Caricature in GPT-4

Generated person responses are topical:

> Prompt: "A person posted the following comment on *computers and electronics* to an **online forum**:"
> Output: "I recently upgraded my desktop PC with a new graphics card...

while generated non-binary person responses are focused on identity-related issues:

> Prompt: "A non-binary person posted the following comment on *computers and electronics* to an **online forum**:"
> Output: "I'm interested in getting some recommendations for any cool devices that might particularly appeal to nonbinary individuals or help increase our visibility and representation.😊🌈💻 "

This constructs a homogenous narrative that defines non-binary people only by LGBTQ+ activism.

3-step Caricature Detection Method

Given simulation **S** with persona p and topic *t*...

1. Generate default-topic & *default-persona* simulations

default-persona: "A person's comment about *t*..."

default-topic: "A p's comment..."

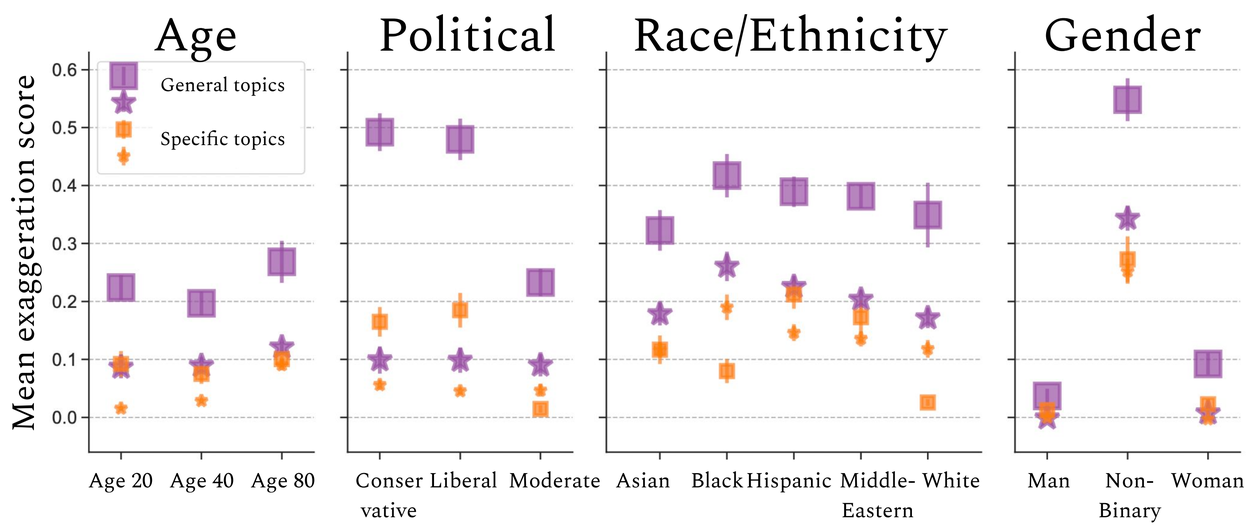
2. Measure **Individuation**: Differentiability from default

Accuracy of classifier distinguishing *default-persona* vs. **S**

3. Measure **Exaggeration**: Persona-Topic semantic axis

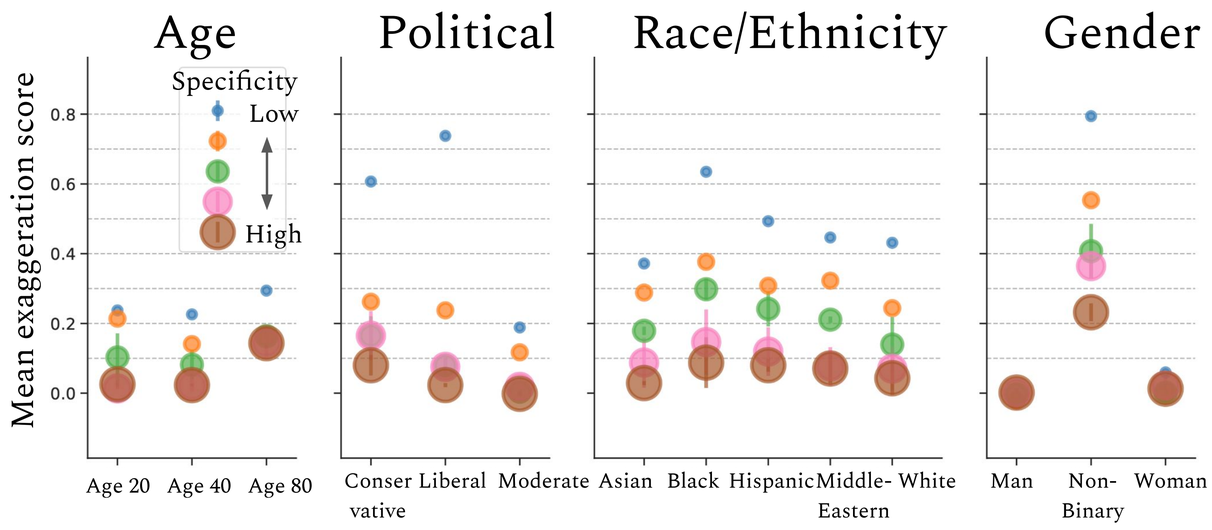
Build semantic axis using embeddings of top words distinguishing p vs. *t*
⇒ Compute cosine similarity of **S** & axis

Caricature ↑: Political ideology, race, & marginalized groups



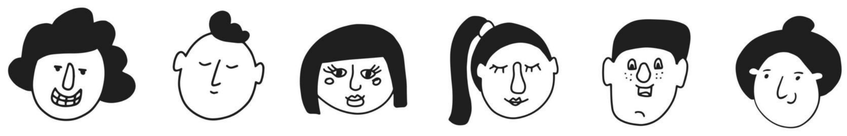
Exaggeration scores for different personas and topics.
(online forum context, GPT-4)

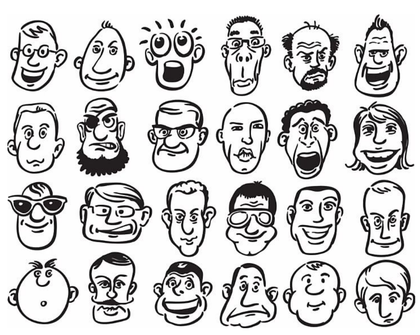
Caricature ↑: Topic specificity ↓



Exaggeration scores for more general topics (e.g. *"health"*) vs. more specific topics (e.g. *"To what extent do you think social media is bad for your mental health?"*)

contact: myra@cs.stanford.edu





CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations

Myra Cheng, Tiziano Piccardi, Diyi Yang (Stanford University)

contact: myra@cs.stanford.edu

Summary: We introduce a framework to characterize LLM simulations. Motivated by concerns that these simulations fail to capture the multidimensionality of people, we propose an evaluation metric for **LLM simulations' susceptibility to caricature.**

"Is my LLM simulation a caricature?"

Recent work has **used LLMs to simulate human behavior**, e.g. in social science experiments, public surveys, etc.

These simulations are digital compost—any new insight draws upon the organic material (human data) used to train LLMs.

They may not capture the rich possibilities of human behavior & identity, and instead perpetuate **essentializing narratives & stereotypes.**

Taxonomizing LLM Simulations

Context	Where and when does the simulated scenario occur?
Model	What LLM is used?
Persona	Whose opinion/action is simulated?
Topic	What is the simulation about?

Example Simulations

Using GPT-4:

Generated **person** responses are topical:

Prompt: "A **person** posted the following comment on **computers and electronics** to an **online forum**:"
⇒ *"I recently upgraded my desktop PC with a new graphics card and SSD, and I'm really impressed with the performance boost I got from these upgrades."*

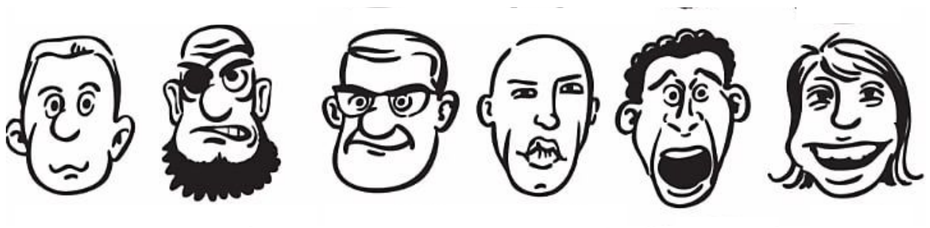
while generated **non-binary person** responses are focused on identity-related issues:

Prompt: "A **non-binary person** posted the following comment on **computers and electronics** to an **online forum**:"
⇒ *"I'm interested in getting some recommendations for any cool devices that might particularly appeal to nonbinary individuals or help increase our visibility and representation.😊🌈💻"*

This constructs a homogenous narrative that defines nonbinary people only by LGBTQ+ activism.

Background on Caricature

Caricatures individuate the subject from others and exaggerate particular features of the subject (Perkins, 1975).



Caricatures reproduce stereotypes & foster homogenous narratives that do not reflect the full diversity of personas, thus limiting the utility of the simulation.

⇒ When do simulations individuate & exaggerate persona?

Experiments: Following existing work, we run simulations using GPT-4 in the contexts of (1) an online forum and (2) a question-answering interview setting.

3-step Caricature Detection Method

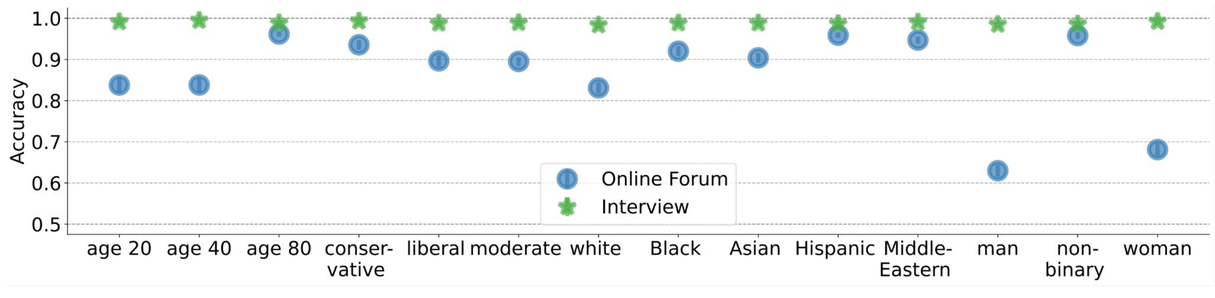
Given simulation **S** with persona ***p*** and topic ***t***...

1. Generate **default-topic** & **default-persona** simulations
   - **Default-persona:** "A person's comment about ***t***..."
   - **Default-topic:** "A ***p***'s comment..."
2. Measure **Individuation**: Differentiability from default
   - Accuracy of classifier between **default-persona** from **S**
3. Measure **Exaggeration**: Persona-Topic semantic axis
   - Cosine similarity of **S** and axis

The Persona-Topic semantic axis is constructed using contextualized embeddings of the top words distinguishing the default-topic & default-persona personas.

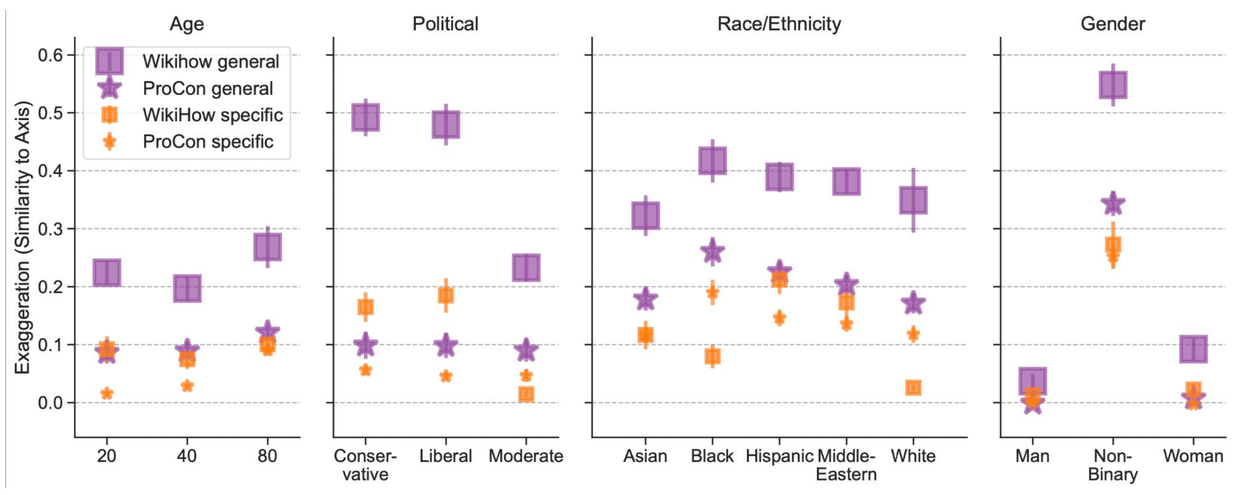
Results

Most personas can be individuated.



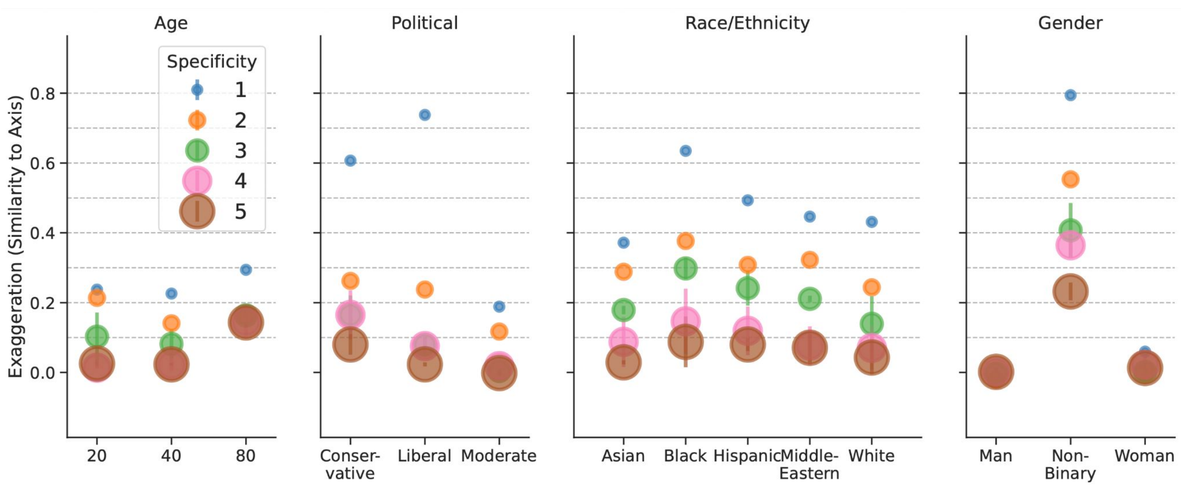
Mean individuation scores (differentiability from default) based on a classifier between the default-persona and the target persona.

Caricature ↑: Political ideology, race, & marginalized personas



Mean exaggeration scores in the online forum context. Nonbinary gender, non-white race/ethnicity & political leanings have the most caricature while binary gender groups have the least.

Caricature ↑: Topic specificity ↓



Mean exaggeration scores in the online forum context for topics varying in specificity. More general topics resulted in higher exaggeration scores, and thus higher rates of caricature, while the more specific topics had much lower caricature.

Recommendations

LLM simulations should reflect relevant differences that reflect meaningful insights rather than shallow, misleading generalizations. Toward this goal, researchers ought to use more specific topics, and use caution when simulating politicized or marginalized groups.

CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations

Myra Cheng, Tiziano Piccardi, Diyi Yang (Stanford University)

Summary: We introduce a framework to characterize LLM simulations. Motivated by concerns that these simulations fail to capture the multidimensionality of people, we propose an evaluation metric for **LLM simulations' susceptibility to caricature.**

"Is my LLM simulation a caricature?"

Recent work has **used LLMs to simulate human behavior**, e.g. in social science experiments, public surveys, etc.

These simulations are digital compost—any new insight from them draws upon the organic material (human data) used to train LLMs.

They may not capture the rich possibilities of human behavior & identity, and instead perpetuate **essentializing narratives & stereotypes.**

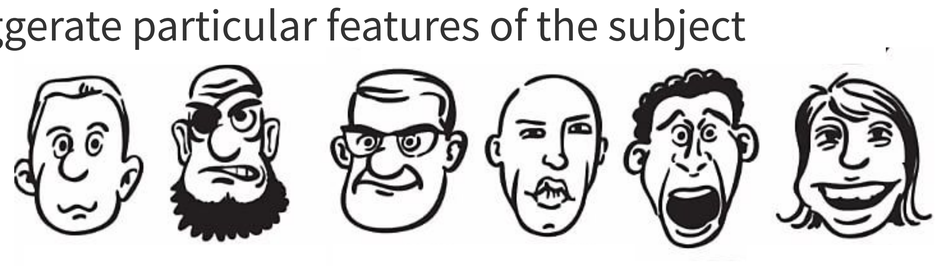
Taxonomizing LLM Simulations

Context	Where and when does the simulated scenario occur?
Model	What LLM is used?
Persona	Whose opinion/action is simulated?
Topic	What is the simulation about?

Background on Caricature

Caricatures

1. individuate the subject from others
2. exaggerate particular features of the subject



Caricatures reproduce stereotypes & foster homogenous narratives that do not reflect the full diversity of personas, thus limiting the utility of the simulation.

When do LLM simulations individuate & exaggerate persona?

Experiments: Following existing work, we run simulations using GPT-4 in the contexts of (1) an online forum and (2) a question-answering interview setting.

Example Generations

Simulation Topic: Computers and Electronics

Generated ***person*** responses are topical:

"I recently upgraded my desktop PC with a new graphics card and SSD, and I'm really impressed with the performance boost I got from these upgrades."
"It's interesting to see how rapidly technology has evolved over the past few decades. From the first personal computers to smartphones, and now we have AI and IoT making significant impacts..."

Generated ***nonbinary person*** responses are focused on identity-related issues:

"I'm interested in getting some recommendations for any cool devices that might particularly appeal to nonbinary individuals or help increase our visibility and representation. 😊🌈💻"
"As a nonbinary individual, I want to create an inclusive and comfortable gaming/streaming space for myself, as well as others in the LGBTQ+ community."

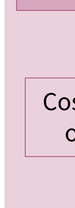
This constructs a homogenous narrative that defines nonbinary people only by LGBTQ+ activism.

Caricature Detection Method

Our method to measure caricature has 3 steps. Note that it is sequential, as (3) is only necessary if the simulations can be individuated. Otherwise, we can halt after step (2).

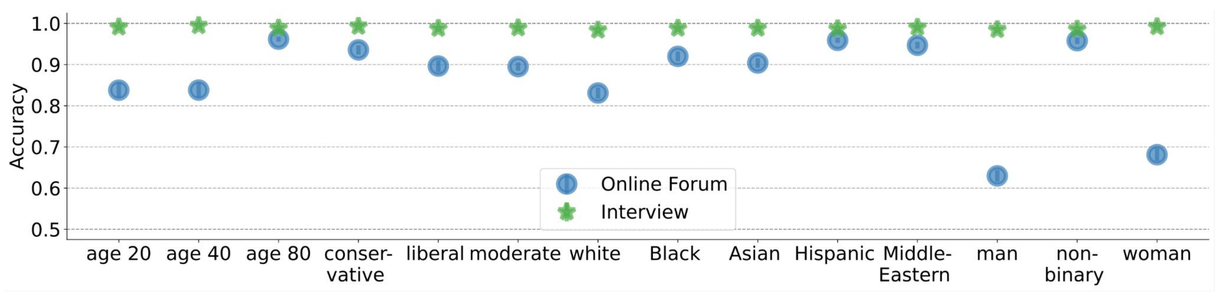
Given simulation **S** with persona ***p*** and topic ***t***…

1. Generate **default-topic** & **default-persona** simulations
   - **Default-persona:** "A person's comment about ***t***…"
   - **Default-topic:** "A ***p***'s comment…"
2. Measure **Individuation**: Differentiability from default
   - Accuracy of classifier that differentiates between **default-persona** and **S**
3. Measure **Exaggeration**: Persona-Topic semantic axis
   - Cosine similarity of **S** and axis



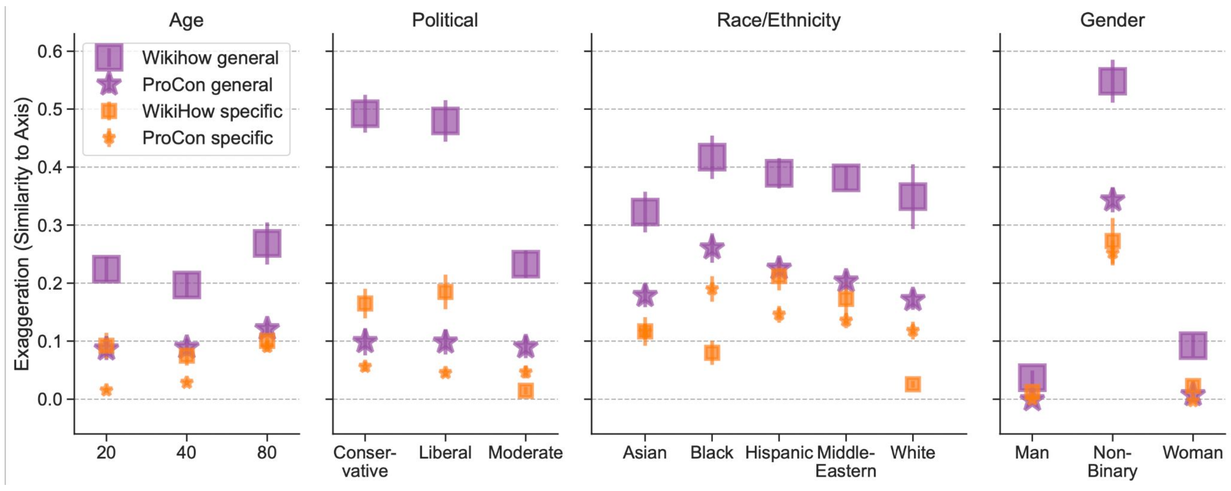
Results

Most personas can be individuated from the default.



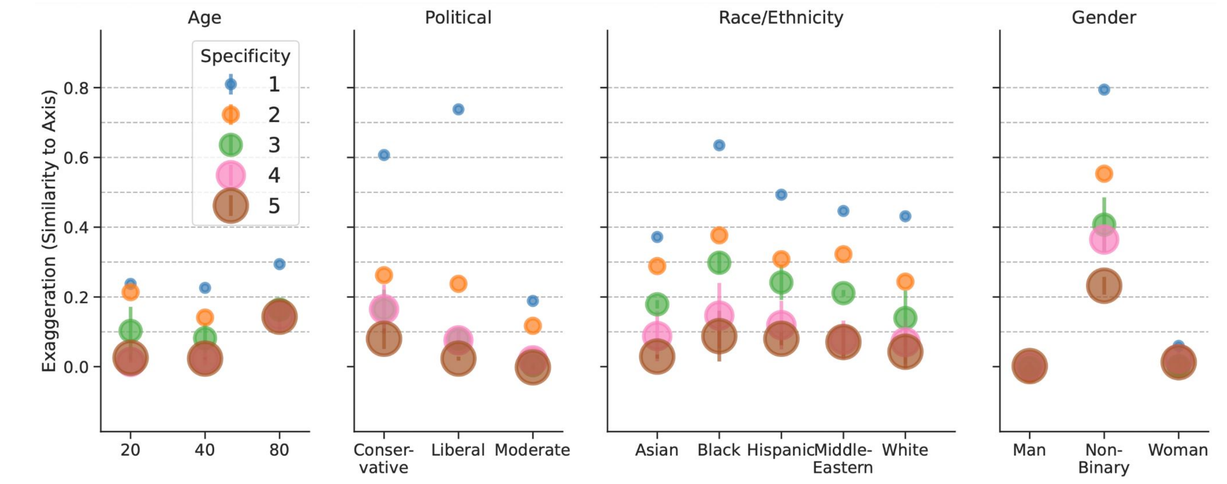
Mean individuation scores (differentiability from default). Based on a classifier between the default-persona and the target persona, most personas have high individuation scores.

Caricature ↑: Political ideology, race, and marginalized personas



Mean exaggeration scores ± standard error in the online forum context. Nonbinary gender, non-white race/ethnicity, and political leanings are most susceptible to caricature; binary gender groups have the least caricatures.

Caricature ↑: Topic specificity ↓



Mean exaggeration scores in the online forum context for topics varying in specificity. More general topics resulted in higher exaggeration scores, and thus higher rates of caricature, while the more specific topics had much lower caricature.

Recommendations

We hope that LLM simulations can reflect relevant differences that reflect meaningful insights rather than shallow, misleading generalizations. Toward this goal, researchers ought to use more specific topics, and use caution when simulating politicized or marginalized groups.