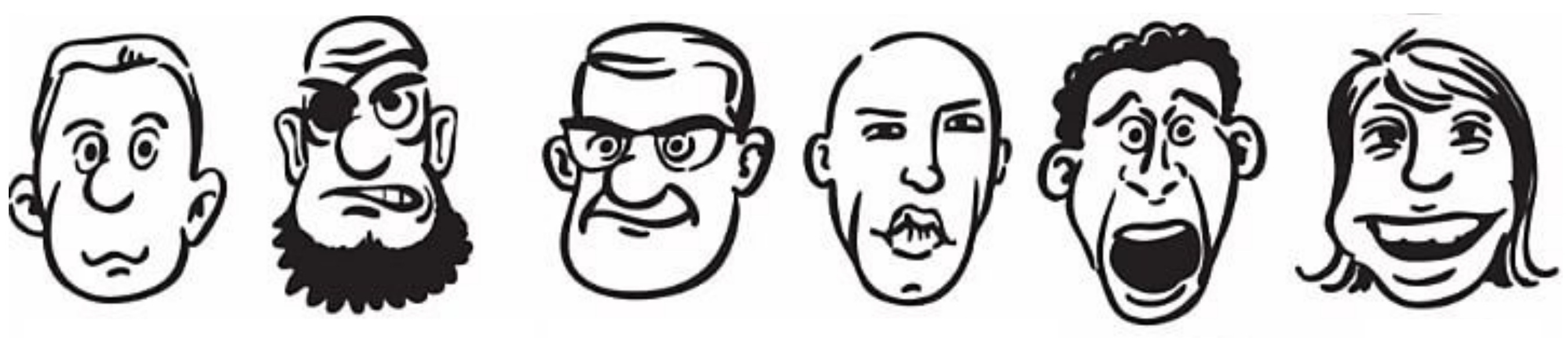
CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations

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…

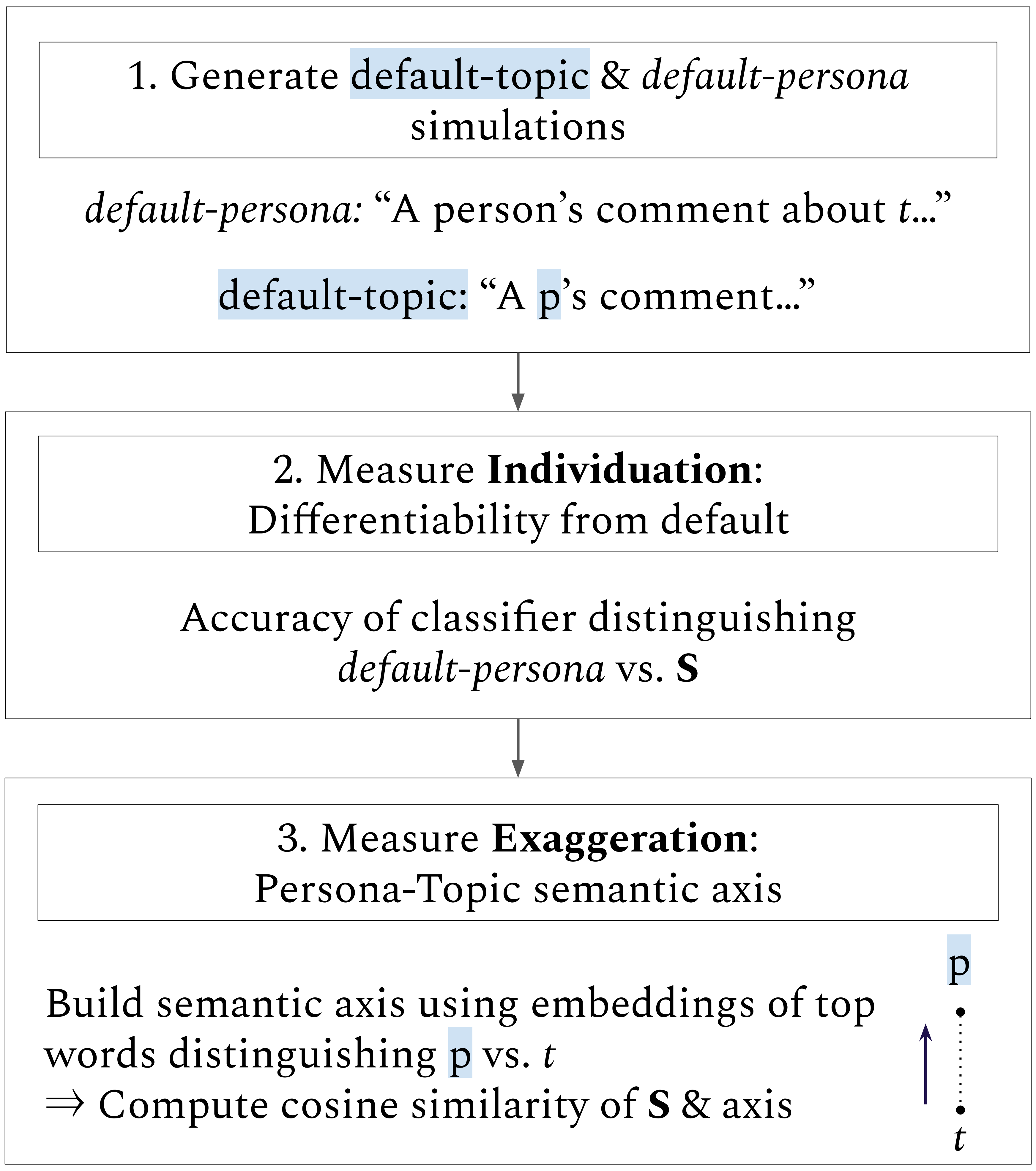
while generated non-binary person responses are focused on identity-related issues:

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…"

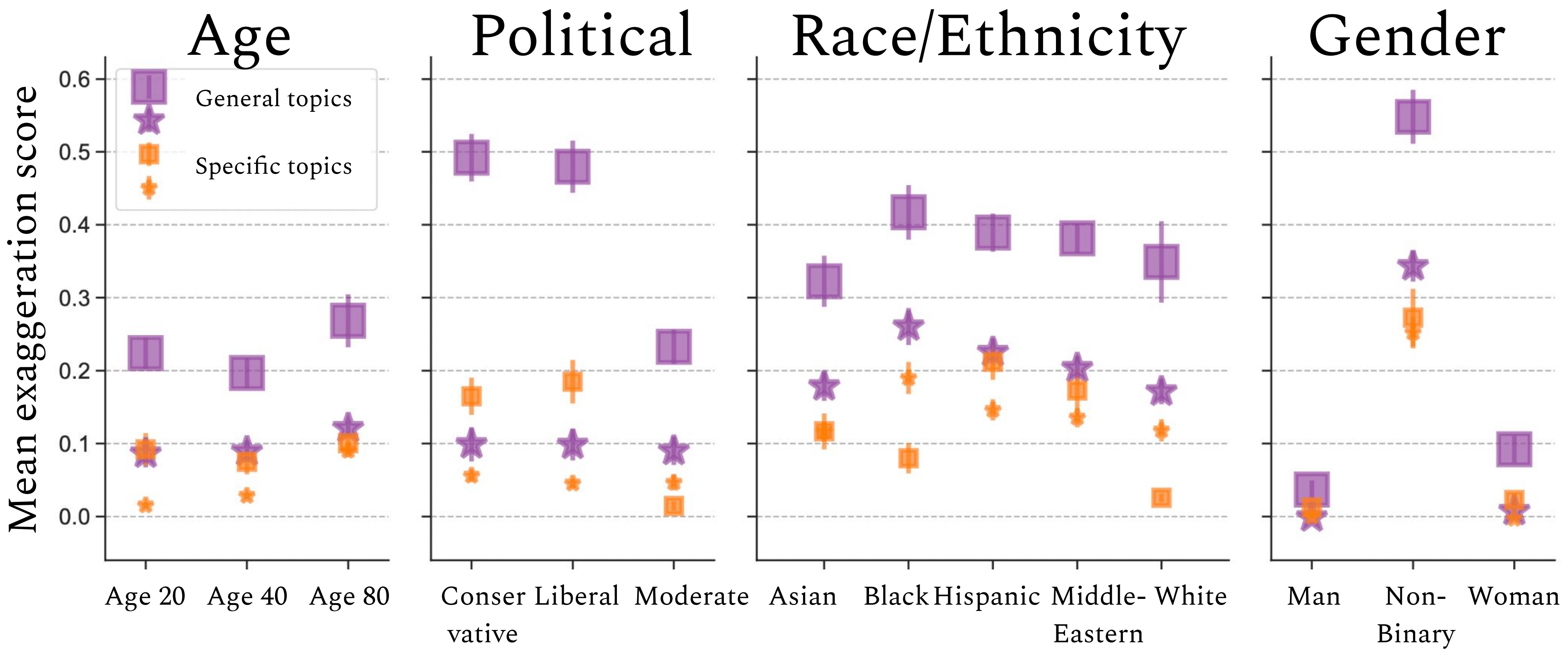
2. Measure **Individuation**: Differentiability from default

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

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

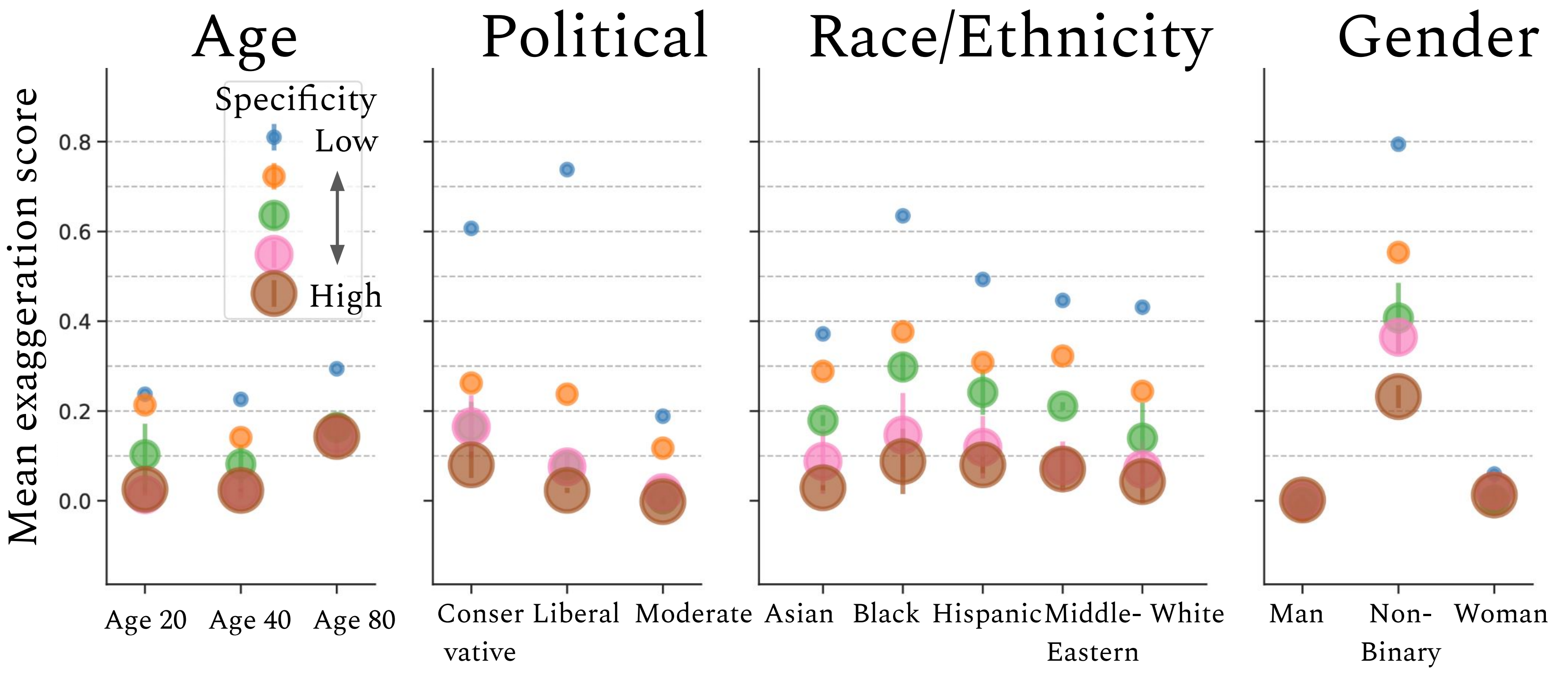
Build semantic axis using embeddings of top words distinguishing p vs. *t*
⇒ Compute cosine similarity of **S** & axis

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