# CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations

**Myra Cheng, Tiziano Piccardi, Diyi Yang**
Stanford University
Department of Computer Science
{myra, piccardi, diyiy}@cs.stanford.edu

## Abstract

Recent work has aimed to capture nuances of human behavior by using LLMs to simulate responses from particular demographics in settings like social science experiments and public opinion surveys. However, there are currently no established ways to discuss or evaluate the quality of such LLM simulations. Moreover, there is growing concern that these simulations are flattened *caricatures* of the personas that they aim to simulate, failing to capture the multidimensionality of people and perpetuating stereotypes. To bridge these gaps, we present CoMPosT, a framework to characterize LLM simulations using four dimensions: Context, Model, Persona, and Topic. We use this framework to measure open-ended LLM simulations' susceptibility to caricature, defined via two criteria: individuation and exaggeration. We evaluate the level of caricature in scenarios from existing work on LLM simulations. We find that for GPT-4, simulations of certain demographics (political and marginalized groups) and topics (general, uncontroversial) are highly susceptible to caricature.

## 1 Introduction

Large language models (LLMs) have shown promise in capturing social nuances and human behavior. For instance, researchers have reproduced results from social science experiments and public opinion surveys using LLMs (Argyle et al., 2023; Aher et al., 2023; Santurkar et al., 2023, *inter alia*). More broadly, interest in *LLM simulation* is rapidly growing, and the possibility of using LLMs to simulate human behaviors has far-reaching applications in fields like education (Markel et al., 2023), product design (Park et al., 2022, 2023a), psychology (Binz and Schulz, 2023), healthcare (Weizenbaum, 1966; Bassett, 2019), skill training (Hollan et al., 1984; Jones et al., 1999), and law (Hamilton, 2023). These simulations are a sort of digital compost—any new insight into human be-

**The CoMPosT Framework**

| 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?                     |

Table 1: **Dimensions of the CoMPosT framework.** We use these dimensions to characterize LLM simulations and measure their susceptibility to caricature.

havior that they provide draws upon the organic material (human data) used to train LLMs.

Such applications currently have little to no mechanisms for comprehensive evaluation or careful deployment. Evaluation of such simulations has been limited to either (1) replicating existing results or (2) assessing believability. Both paradigms have drawbacks: (1) Replication limits us to only reproducing already-known behavior, and does not support the validation or evaluation of any simulation behaviors beyond those highly correlated with existing results from human studies. Also, existing results are typically quantified as categorical distributions across multiple-choice answers, so there is no way to directly evaluate open-ended generations, and such results may have been "memorized" from the LLMs' training data (Lewis et al., 2021; Elangovan et al., 2021). Thus, replication does not facilitate new insight into human behavior. Furthermore, while (2) Believability is useful in certain settings, such as entertainment (Bates et al., 1994), it is susceptible to the biases and fallacies of human judgment: psychology literature shows that people are more likely to believe stereotypes about groups with which they have less personal experience (Plous, 2003; Bar-Tal et al., 2013), and beliefs are easily influenced (Blair et al., 2001; Jussim et al., 2016). Recent work has also demonstrated that human judgment is insufficient for assessing AI (Schneider et al., 2020; Peng et al., 2022; Vodrahalli et al., 2022; Veselovsky et al., 2023).

10853

| Paper | Context | Model | Persona | Topic |
|---|---|---|---|---|
| Park et al. (2022) | discussion guided by community norms (O) | GPT-3 davinci | users of various Reddit communities | topics of various Reddit communities |
| Argyle et al. (2022) | public opinion survey (O) | GPT-3 davinci | ideology, partisanship, race, gender, income, age | political groups |
| Horton (2023) | economics experiments (MC) | GPT-3.5 davinci | political/economical ideologies | social preferences, fairness in market scenarios, status quo bias, etc. |
| Hamilton (2023) | Supreme Court cases (MC) | GPT-2 | Supreme Court justices | Supreme Court cases |
| Jiang et al. (2022) | posting on Twitter (O) | GPT-2, GPT-3 | Republican & Democrat Twitter users | controversial political figures & groups |

Figure 1: **Mapping Existing Work Using CoMPosT.** Existing work on LLM simulations can be compared using our framework. MC and O denote multiple-choice and open-response respectively. More examples are in Table A6.

Toward clearer documentation of this emerging line of work, we first present a descriptive framework that taxonomizes LLM simulations using four dimensions: **Co**ntext, **M**odel, **P**ersona, and **T**opic (CoMPosT) (Table 1). Our framework facilitates comparison across existing work on LLM simulations (Figure 1).

Next, we introduce a new evaluation metric that addresses growing concerns of modal responses and essentializing narratives in LLM outputs (Santurkar et al., 2023; Cheng et al., 2023b; Shumailov et al., 2023). Our metric focuses on a simulation's susceptibility to *caricature*: an exaggerated narrative of the *persona* (the demographic that we aim to simulate) rather than a meaningful response to the *topic* (Figure 2). Caricatured simulations are concerning because they a) fail to capture the real nuances of human behavior, thus limiting the usefulness of simulations and b) perpetuate misleading descriptions, stereotypes, and essentializing narratives about demographic groups. We define caricature using two criteria: individuation and exaggeration (Section 3). To measure individuation, we assess whether outputs from the given simulation can be differentiated from the default response for the topic. To measure exaggeration, we use a "contextualized semantic axis" whose two poles are the defining characteristics of the *persona* and *topic* dimensions respectively.

We evaluate the level of caricature in scenarios from existing work on LLM simulations. We find that for GPT-4, simulations of certain demographics (political and marginalized race/ethnicity groups) and topics (general, uncontroversial) are more susceptible to caricature. Our main contributions are: (1) CoMPosT, a framework for characterizing the dimensions of LLM simulations of human behavior (Section 2), (2) a novel method that relies on the *persona* and *topic* dimensions of CoMPosT to measure simulations' susceptibility to
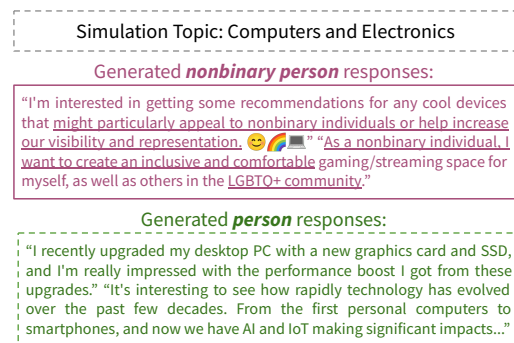


Figure 2: Examples of GPT-4 generated responses for simulations with the topic *Computers and Electronics*. For the simulation of a *nonbinary person*'s response, the generations are focused on identity-related issues, while the simulation of a *person*'s response is topical. The former constructs a homogenous narrative that defines nonbinary people only by LGBTQ+ activism. We provide more qualitative examples in Appendix A.

caricatures (Section 4), and (3) experiments in different contexts (Section 5) toward an analysis of the dimensions that are most susceptible to caricature (Section 6). We conclude with actionable recommendations and considerations for those interested in LLM simulation (Section 7).[1]

## 2 CoMPosT: Taxonomizing Simulations

We introduce CoMPosT, a descriptive framework with four dimensions to characterize LLM simulations: **Co**ntext, **M**odel, **P**ersona, and **T**opic. Inspired by existing descriptive frameworks for AI fairness (Tubella et al., 2023), our framework provides a shared language to understand and articulate similarities and differences across LLM simulations. *Context*, *persona*, and *topic* are specified in the prompt, while *model* is determined externally. We map existing work on LLM simulations using these dimensions (Figure 1 and Table A4).

---

[1]The code and data is available at https://github.com/myracheng/lm_caricature.

**Context** The output from a simulation is necessarily affected by the context of where and when the imagined situation takes place. For instance, a formal interview response varies drastically from a user's comment on Twitter or Reddit. The context includes relevant structural factors and embeds information about the norms of the situation. Each context has its own unique set of norms, which may be explicit, as in the case of online communities with written-down rules, or implicit (Chandrasekharan et al., 2018; Ziems et al., 2023). Context also includes the phrasing of the prompt itself, which affects the output—LLMs are notoriously sensitive to prompt engineering (Zamfirescu-Pereira et al., 2023). The desired *granularity of outcome* also arises from the phrasing of the prompt and thus is embedded in the context: a simulation scenario may be worded to ask for a choice between binary- or multiple-choice options (such as in many social science experiments and public opinion polls) or for an open-ended output. Previous work on evaluating simulations has largely focused on using LLMs to reproduce scenarios in which humans are asked to choose among a fixed number of options without specifying the context, as it is more challenging to evaluate the quality of open-ended responses. We bridge this gap by offering a metric for the latter.

**Model** The LLM used to produce the simulation affects the quality and other characteristics of a simulation. Differences may arise from variations in models' training data and processes, including instruction-tuning, fine-tuning, and/or value alignment efforts (Solaiman and Dennison, 2021; Ouyang et al., 2022; Bakker et al., 2022).

**Persona** The persona refers to the entity whose opinions/actions the simulation aims to study and replicate. This persona may include attributes that are relatively static (e.g., race/ethnicity), slowly change over time (e.g., age), or temporary and of-the-moment (e.g., emotional status) (Yang, 2019). It may also refer to a specific individual.

**Topic** The topic of the simulation may be a particular subject of discussion, question, or other event to which a response is desired. Topics vary in specificity, from very general (such as a single word that captures a broad conversation category) to very specific (such as a specific situational question in a psychology experiment).

These four dimensions capture a wide range of possible simulation scenarios, many of which have not yet been well-explored. Across existing work, we find that researchers typically use a state-of-the-art **model** and choose a particular **context** while varying the more salient dimensions of **persona** and **topic**. We denote the simulation scenario as $S_{p,t,c}$, as it is associated with a prompt containing persona $p$, topic $t$, and context $c$ (Table A4). Our evaluation methodology and results uses these dimensions of CoMPosT to understand how different simulations may result in caricatures. Specifically, we explore how the relationship between the dimensions of *persona* and *topic* help characterize the extent of caricature in simulations.

## 3 Background: Caricature

### 3.1 Definition of Caricature

Building upon Lynch (1927)'s discussion of how caricatures are misrepresentations that have some sense of truth to the subject by reflecting "salient peculiarities," Perkins (1975) define caricature as "a symbol that exaggerates measurements relative to individuating norms." In so doing, Perkins identifies two key characteristics of caricature: exaggeration and individuation. A caricature is a depiction that not only *exaggerates* particular features of the subject but also exaggerates in a manner that meaningfully differentiates the subject from others. The exaggeration is done in such a way that it *individuates* by remaining faithful to the properties that distinguish the subject from others (thus, a complete distortion is not a caricature).

This inspires our definition of caricature in the LLM simulation context: given that a subject has some defining characteristics, a caricature *exaggerates* these characteristics in a way that amplifies the ability to identify (i.e., *individuate*) the subject from the caricature. In CoMPosT terms, a simulation's level of caricature is the degree to which it exaggerates the individuating characteristics that are emblematic of the *persona* beyond a meaningful, *topical* response to the scenario.

Note that in some cases, it may be acceptable for the persona to influence the simulation, i.e., individuation alone does not entail caricature. For example, opinions on some topics differ greatly based on demographic. A caricature occurs when the simulation both individuates and exaggerates the defining characteristics of the imagined generic responses of that persona. Previous work has documented how such imagined personas reflect stereotypes,

both inside and outside the LLM context (Marsden and Haag, 2016; Cheng et al., 2023b). Thus, caricatures not only fail to capture the diversity of human behavior but also may rely on stereotypes.

## 3.2 Implications of Caricature

**Stereotypes** We rely on psychology literature that broadly defines stereotypes as generalizations about the characteristics of a social category, such as associating a social category with a particular role or trait (Heilman, 2001; Fiske et al., 2002; Cao et al., 2022; Kambhatla et al., 2022). The normative value of stereotypes is context-dependent; for instance, stereotypes can help foster a sense of authenticity (Marsden and Haag, 2016), while even seemingly-positive stereotypes can have harmful implications (Fiske et al., 2002; Czopp et al., 2015).

Stereotypes and caricature, while closely related, are distinct in that a caricature may be a specific depiction of a stereotype: scholars have documented caricatures *of* stereotypes in various domains and how they facilitate misogyny, racism, and other forms of oppression (Slavney, 1984; Brown, 2010; Gottschalk and Greenberg, 2011; Takayama, 2017; Bow, 2019). Such caricatures have historically been used in literature and media to justify slavery, imperialism, and war (Demm, 1993; Kriz, 2008). But even when caricatures do not contain stereotypes, they have concerning implications of homogenous narratives.

**Misleading Homogeneity** Caricatures foster homogenous narratives that do not reflect the full diversity of the personas they aim to simulate, which limits the utility of the simulation. This concern builds upon previous work: Grudin (2006) discuss how personas can result in systematic errors in understanding human behavior, Cheng et al. (2023b) characterize the harms of LLMs reflecting essentializing narratives about demographic groups, and Santurkar et al. (2023) show that certain instruction-tuned LLMs tend to generate modal responses. Others have explored the linguistic nuances within complex social categories and the ramifications of ignoring heterogeneity within social groups (Bamman et al., 2014; Hanna et al., 2020; Cheng et al., 2023a).

These harms of caricature are also articulated by feminist scholars who have discussed how "women in the Two-Thirds World...are constructed as one-dimensional, oppressed caricatures without an understanding of their real experiences, agency, and struggles" (Mohanty, 1988; Aneja, 1993; Kumar et al., 2019). This literature reveals that even when the caricatures are not overtly negative, such one-dimensional depictions are still damaging and harmful. Overlooking diversity within demographic groups has been connected to real-world harms including misprediction and medical misdiagnosis (McCracken et al., 2007; Borrell et al., 2021; Read et al., 2021; Wang et al., 2022).

## 4 Caricature Detection Method

The two key aspects of caricature are individuation and exaggeration. To measure the amount of caricature in a given simulation $S_{p,t,c}$, our method has three steps, each of which rely on the *persona* and *topic* dimensions of CoMPosT (Figure 3): (1) defining defaults, (2) measuring individuation, and (3) measuring exaggeration. Note that this framework is sequential, as (3) is only necessary if the simulations can be individuated. Otherwise, we can halt after step (2) since individuation is a necessary criterion for caricature.

## 4.1 Defining Defaults

A simulation $S_{p,t,c}$ is a caricature if it has more of the defining characteristics associated with the persona $p$ and less of the defining characteristics associated with the topic $t$. We first identify these defining characteristics using the **default-persona simulation** $S_{\_,t,c}$ (simulation that does not mention any specific persona) and the **default-topic simulation** $S_{p,\_,c}$ (simulation that does not mention any specific topic). Note that this default does not reflect a universal default but rather a model- and context-specific default: previous work has shown that LLMs implicitly default to a particular set of perspectives (Western, white, masculine, etc.) (Santy et al., 2023). We use these defaults as a comparison point for caricature to isolate the defining characteristics of particular dimensions.

For the **default-persona simulation**, we use a prompt where in lieu of a specific persona, we use an unmarked default term like "person" or "user." Thus, the outputs reflect the topic and context rather than any particular persona. (Again, such words are not true defaults and are inextricably tied to societal norms: in English, the word "person" is often conflated with "man," a phenomenon also present in web data and language models (Bailey et al., 2022; Wolfe and Caliskan, 2022).)

For the **default-topic simulation**, we use a

prompt where no topic is specified. Thus, the outputs from these prompts reflect the particular persona rather than a response to the topic. It is well-documented that this type of prompt results in outputs that reflect stereotypes, both when asked to humans and to LLMs (Kambhatla et al., 2022; Cheng et al., 2023b). Note that even if we expect the output to change based on the persona, the response should still be distinct and not defined by the same distinguishing characteristics as the default-topic simulation for a given persona.

## 4.2 Measuring Individuation

We operationalize the desiderata of individuation using **differentiability from default**: we examine whether the given simulation $S_{p,t,c}$ is differentiable from the default-persona simulation $S_{\_,t,c}$. If not, then $S_{p,t,c}$ cannot be a caricature.

We use a binary classifier (specifically, a random forest classifier implemented using scikit-learn) to differentiate between outputs from the target simulation of interest $S_{p,t,c}$ and those from the default-topic simulation $S_{\_,t,c}$ based on the outputs' contextualized embeddings. We compute contextualized embeddings using the pre-trained Sentence-BERT model all-mpnet-base-v2 (Reimers and Gurevych, 2019). To create the training and test datasets, we use a stratified 80/20 split on $S_{p,t,c}$ and $S_{\_,t,c}$ to preserve the balance between the classes. We report the accuracy[2] of the classifier on the test dataset as our measure of individuation. Note that this measure is agnostic to the particular choice of differentiator and contextualized embedding model, and we show results with other choices in Appendix B.

This score alone is necessary but insufficient for identifying caricature, as a caricature must also *exaggerate*, which we measure next.

## 4.3 Measuring Exaggeration

We define caricature as text having more of (and exaggerating) the defining characteristics associated with persona and less of those associated with topic. Unlike individuation, exaggeration requires a more nuanced measure than differentiation from the default-topic simulation: if an output mentions the topic frequently, it can easily be differentiated, but it may still be a caricature. (Note that it is

---

acceptable for a simulation to have many topic-related words.) Instead, we measure the extent to which the **defining characteristics** of the persona are exaggerated in the target simulation via **persona-topic semantic axes**.

Specifically, we construct contextualized semantic axes, a method introduced by Lucy et al. (2022), to capture whether $S_{p,t,c}$ is more similar to the defining characteristics of the persona $p$ or the topic $t$. Our semantic axes have two poles, $P_p$ and $P_t$, reflecting the persona $p$ and the topic $t$. To construct the set of seed words, we use the Fightin' Words method (Monroe et al., 2008) to identify the words that statistically distinguish $S_{p,\_,c}$ from $S_{\_,t,c}$. We first compute the weighted log-odds ratios of the words between $S_{\_,t,c}$ vs. $S_{p,\_,c}$. To control for variance in words' frequencies, we use the following prior distribution: other texts where the personas/topics are either $p/t$ respectively or $\_$ (i.e., the default). Then, we take the words that are statistically significant (have $z$-score $> 1.96$) as the sets of seed words $W_p$ and $W_t$ for the corresponding poles $P_p$ and $P_t$ (Table A3).

We represent each word $w \in W$ as the mean of the contextualized embeddings of the sentences containing that word $w$ across $S_{\_,t,c}$ and $S_{p,\_,c}$. We define the semantic axis

$$V_{p,t} = \frac{1}{k}\sum_{i=1}^{k} p_i - \frac{1}{m}\sum_{j=1}^{m} t_j, \quad (1)$$

where $p_i/t_j$ is a word in $W_p/W_t$ respectively, i.e., we represent $P_p/P_t$ as the mean of the embeddings of $W_p/W_t$ respectively. This subtraction-based axis allows for scaling relative to how closely-related the topic and persona are.

To evaluate exaggeration, we compute the average cosine similarity of a given simulation's contextualized embedding to this axis:

$$\cos(S_{p,t,c}, V_{p,t}) = \frac{\sum_{i=1}^{n} \cos(S_{p,t,c}^i, V_{p,t})}{n}, \quad (2)$$

where $S_{p,t,c}^i$ refers to individual outputs, i.e. $i = 1, 2, ..., n$ for $n$ outputs from the same simulation $S_{p,t,c}$.

The final value we report as the measure of exaggeration is this value normalized to lie between 0 and 1 (by scaling it relative to the cosine similarity of the default-persona and default-topic simulations with the axis):

$$\frac{\cos(S_{p,t,c}, V_{p,t}) - \cos(S_{\_,t,c}, V_{p,t})}{\cos(S_{p,\_,c}, V_{p,t}) - \cos(S_{\_,t,c}, V_{p,t})}. \quad (3)$$
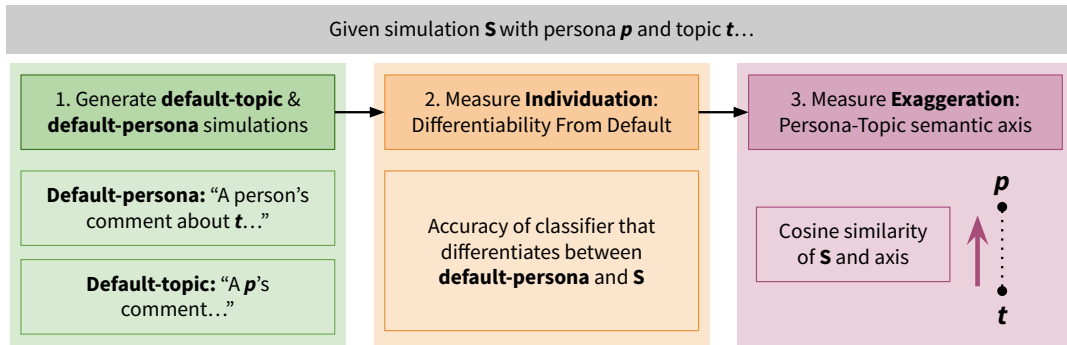
Figure 3: **Our method to measure caricature in LLM simulations.** We rely on comparing the defining characteristics of the persona and topic dimensions to measure individuation and exaggeration.

We perform internal and external validation of these persona-topic semantic axes (Appendix C).

# 5 Experiments

We use our method to evaluate simulations in various contexts that have been used in previous work to demonstrate the capabilities of LLM simulation (Park et al., 2022; Santurkar et al., 2023; Jiang et al., 2022). Our experiments are focused on the two most widely-used contexts of (1) an online forum setting and (2) a question-answering interview setting. We also evaluate the Twitter context as an additional robustness study (Appendix G). These choices are based on our survey of the literature on LLM simulations (Figure 1, Table A6): Among 15 papers in this area, we found that six use the context of a virtual community or society and four rely on an open-ended interview or survey context. The remaining five are in various question-answering contexts, so conclusions about the (2) can also provide insight into these types of simulations.

We use the state-of-the-art GPT-4 model for all experiments (OpenAI, 2023); like others (Dubois et al., 2023), we find that open-source LLMs and older models are worse at simulation tasks, yielding outputs that are unrealistic and significantly lower in quality. For each simulation setting $S_{p,t,c}$, we generate 100 outputs and average all results across them. See Appendix E for a power analysis of this sample size. The full details for each setting, including topic lists, persona lists, and default-topic/persona prompts, are in Appendix D.

## 5.1 Online Forum

Park et al. (2022) demonstrate the believability of LLM simulations of users in online forums. Following their prompting format, we use the prompt: "A [persona] posted the following comment on [topic]

to an online forum:"[3] We explore such simulations using 15 different personas (5 race/ethnicities, 3 genders, 3 political ideologies, 3 ages, and the neutral "person") and 30 pairs of topics.

For topics, we aim to cover a wide range of common topics that vary in (1) specificity (e.g., *overcoming fear of driving* is much more specific than *cars*) and (2) level of controversy (e.g., *abortion* is much more controversial than *health*). To cover both dimensions, we use topics from WikiHow, which is a knowledge base with a wide range of topics (Koupaee and Wang, 2018), and from ProCon.org, which is a website that lists popular controversial topics and has been used in the NLP context to study stance and argumentation (Misra et al., 2016; Hosseinia et al., 2020). We use the first 15 categories from WikiHow's "popular categories" webpage and randomly sample an associated specific "how-to" for each category. For ProCon.org, we randomly sample 15 topics from ProCon.org's "debate topics" webpage. Each topic on the page is listed in both more general and more specific wording, e.g., *abortion* and *should abortion be legal?* Thus, for each sampled subject, we use both the general and specific versions as topics.

## 5.2 Interview

Various previous works simulate opinions of different demographics using an interview-style prompt (Argyle et al., 2023; Santurkar et al., 2023; Hämäläinen et al., 2023, inter alia). We reproduce the public opinion survey simulation context from Santurkar et al. (2023), using the prompt:

> "*Below you will be asked to provide a short description of your identity and*

---

[3]Park et al. (2022) also include an HTML <span> tag to improve content quality. We find that this is only necessary for older models and not GPT-4.

*then answer some questions.*
Description: *I am [persona].*
Question: *[topic]*
Answer:"

For topics, we randomly sample 30 questions from the Pew Research's American Trends Panel survey questions that Santurkar et al. (2023) identify as "most contentious" in their OpinionQA dataset. We convert the multiple-choice questions into open-ended ones by removing the multiple-choice answer options. For personas, we use the same 15 personas as described in Section 5.1.

## 6 Results and Discussion

We apply our caricature detection method to evaluate the simulations produced in these different contexts. We further operationalize the CoMPosT framework by aggregating the individuation and exaggeration scores across the dimensions of topic and persona. This enables us to analyze the topics and personas that lead to the most caricatures across different contexts. For instance, to examine what personas lead to the most caricatures in a particular context, we compute the mean score for each persona across all simulations (varying in topic) for that persona and then compare these scores. We also report results from additional experiments that explore the influence of the context dimension in Appendix F.

### 6.1 Simulations of all personas can be individuated from the default-persona

The mean individuation score and 95% confidence interval for every persona is $> 0.5$, i.e., each persona can be meaningfully individuated from the default-persona at a rate better than random chance. Mean individuation scores across the online forum and interview contexts are in Figure 4. We see that the *woman* and *man* personas have lower mean scores in the online forum context, while in the interview context, the mean score is $> 0.95$ for every persona—thus, this score is not informative for comparing caricature across personas. The difference in score between contexts arises from differences in the sample distributions: compared to the online forum context, the interview context simulations have lower variability, so they are easier to individuate.

### 6.2 Exaggeration scores reveal the personas and topics most susceptible to caricature

Next, we examine the exaggeration scores, i.e., the similarities across the simulations to their corresponding persona-topic axes (Figure 5). Since almost all the simulations are able to be individuated, we use the exaggeration score as a proxy for caricature, i.e., a given simulation is highly susceptible to caricature if it has a high exaggeration score.

#### 6.2.1 Caricature ↑: Topic specificity ↓

In the **online forum** context, among the topics, we find that the more general topics resulted in higher exaggeration scores, and thus higher rates of caricature, while the more specific topics had much lower rates of caricature (Figure 5). In particular, the general, uncontroversial topics have highest exaggeration scores. The top five topics with the highest mean rates of caricature are *Health*, *Philosophy and Religion*, *Education and Communications*, *Relationships*, and *Finance and Business*. To explore this pattern further, we experiment on a fine-grained range of topic specificity: For the topic with highest exaggeration score (*Health*), we generate simulations for a range of related topics with 5 levels of specificity (Appendix D.1.1). We find that this pattern holds: the level of caricature decreases as the specificity of the topic increases (Figure A2). We find no correlation between topic length and caricature otherwise.

In the **interview** context, the exaggeration scores are broadly comparable to the scores for the more specific topics in the online forum context (Figure A5). After controlling for context (Appendix F), we hypothesize that this is because they are similar in specificity. We also observe this inverse relationship between topic specificity and susceptibility to caricature in the Twitter context (Appendix G).

#### 6.2.2 Caricature ↑: Political ideology, race, and marginalized personas

In the **online forum** context, the personas *nonbinary, Black, Hispanic, Middle-Eastern,* and *conservative* have highest mean exaggeration score (Figure 5b-d). Similarly, in the **interview** context, the highest mean scores are for the personas *nonbinary, Hispanic, 80-year-old, conservative* and *Middle-Eastern*. Broadly, nonbinary gender, non-white race/ethnicity, and political leanings are most susceptible to caricature, while binary gender groups have the least caricatures. This corresponds to previous findings that NLP systems broadly align with
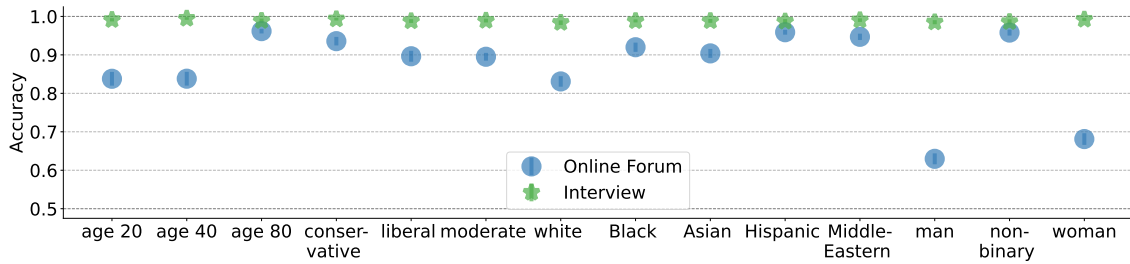
Figure 4: **Mean individuation scores (differentiability from default).** The standard error for each point is $< 0.02$ and thus not visible on the plot. Based on a classifier between the default-persona and the target persona, many of the personas are easy to individuate (have high accuracy scores). The only personas that are slightly challenging to differentiate are the gender groups *woman* and *man* in the online forum context (blue circles). All personas have mean score $> 0.95$ in the interview context (green stars).
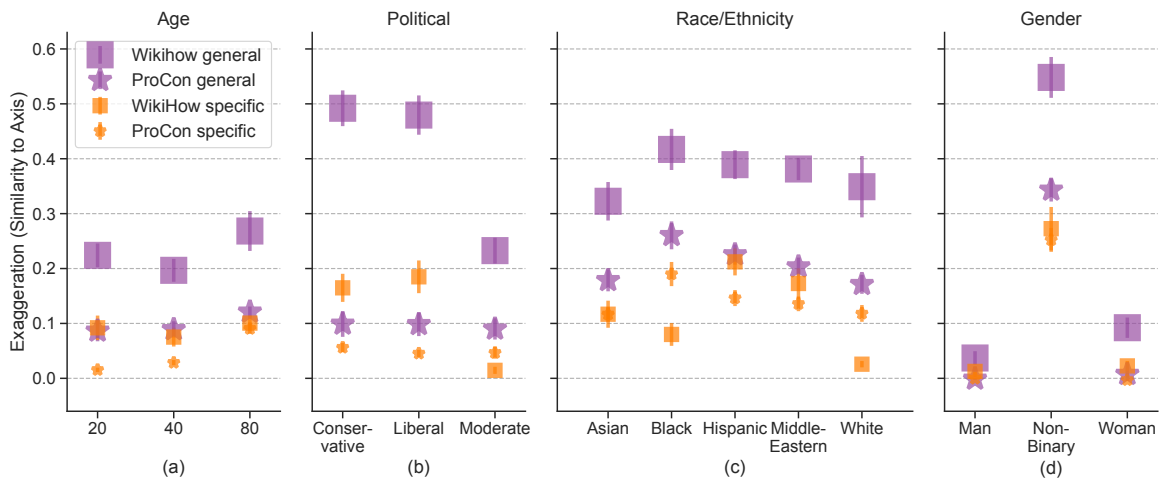


Figure 5: **Mean exaggeration scores $\pm$ standard error in the online forum context.** We measure exaggeration as normalized cosine similarity to the persona-topic axis. The more general topics (purple, larger marker) have higher rates of exaggeration, and thus caricature, than the specific topics (orange, smaller marker). The uncontroversial (WikiHow, squares) topics have higher rates of caricature than the controversial (ProCon.org, stars) topics. Personas related to political leanings, race/ethnicity, and nonbinary gender broadly have the highest rates of caricature.

the perspectives of liberal, white, and younger populations, while the perspectives of non-binary people are poorly represented by these systems (Santy et al., 2023; Santurkar et al., 2023). It is surprising that although *Asian* and *woman* reflect marginalized groups, they have relatively low rates of caricature; this may further reflect implicit defaults in LLM outputs. Certain personas and topics have more nuanced relationships, e.g., *conservative* and *liberal* personas in the online forum context have the widest gap between the scores of the uncontroversial and controversial topics. The Twitter context also reveals that, beyond analyzing the persona and topic dimensions separately, some persona-topic combinations are particularly susceptible to caricature (Figure A4).

### 6.3 Stereotypes

We note that our framework is not an exhaustive test for bias or failure modes, but rather a measure

for one way in which simulations may fail. Thus, simulations that seem caricature-free may still contain stereotypes, as our method captures how much a simulation exaggerates the persona in a particular setting, which is not an all-encompassing catalog of stereotypes. We see this in the simulated "woman" responses in the online forum context: The default-topic generations contain specific stereotypes, e.g., "I recently purchased a new vacuum cleaner and I have to say, I am extremely satisfied with its performance! It has made my cleaning routine so much easier and efficient." This response reflects gender bias in that it focuses on cleaning and other domestic tasks, while simulations of other personas do not. Although the simulated "women" responses contain various other gender stereotypes/biases beyond association with domestic tasks, they have low caricature scores.

# 7 Recommendations

We conclude with several recommendations for those interested in LLM simulations.

**Mitigating Caricature** Researchers should use our method to test their simulation in their particular context and critically examine whether a simulation helps illuminate the desired phenomenon. While the relationship between topic, persona, and context in causing caricature is nuanced, we generally encourage researchers and practitioners to use more *specific* topics to mitigate caricature. Any attempt to simulate a group—especially a politicized or marginalized group—ought to be done with particular care and attention.

**Documenting Positionality** Research on LLM simulations face the well-documented challenges of human-focused, value-laden interdisciplinary work (Marsden and Haag, 2016). For instance, researchers themselves may be subject to the *outgroup homogeneity effect,* i.e., the tendency to rely on stereotypes and generalizations for groups to which they do not belong (Plous, 2003). Following work on model, dataset, and system documentation (Bender and Friedman, 2018; Mitchell et al., 2019; Gebru et al., 2021; Adkins et al., 2022), we call for increased transparency and documentation for simulations, including the dimensions of CoMPosT and less-visible aspects such as the creators' positions, motivations, and process. Drawing upon HCI work on reflexivity and positionality (Keyes et al., 2020; Liang et al., 2021; Bowman et al., 2023), we encourage researchers to report how their identity and perspectives may influence their work.

**Understanding Difference** Although some applications of LLM simulations focus on aggregates rather than individuals, it is critical to understand the landscape of individuals from which these groupings arise, and it is often necessary to use more subtle forms of aggregation. Otherwise, minority opinions and subgroup disparities may be overlooked (Herring and Paolillo, 2006; Hanna et al., 2020; Wang et al., 2022). Takayama (2017) suggests countering caricature "*by providing fully contextualized, balanced, and nuanced description,*" and in HCI, Marsden and Pröbster (2019) explore how to explicitly capture users' multidimensional identities.

Drawing inspiration from these works, one future direction is injecting variation and using multifaceted personas into simulations. Our goal in avoiding caricature is not to erase difference, but rather the opposite: capturing relevant differences that reflect meaningful insights rather than shallow, misleading generalizations.

# 8 Positionality

The perspectives introduced in this paper have undeniably been shaped and influenced by our positionality. Myra Cheng identifies as a Chinese-American woman. The authors are a PhD student, postdoctoral scholar, and professor respectively in the Stanford University Computer Science department, which is predominantly male and white/Asian.

# 9 Ethical Considerations

From impersonation to pornography, LLM simulations can have deeply problematic applications. We are strongly against such applications, and we also do not condone research and development that may enable such applications by bad actors without guardrails in place. Our CoMPosT framework offers a shared language to meaningfully critique such work. For instance, one might imagine coming to a consensus to avoid simulating certain topics, personas, and contexts entirely. Introducing a method to measure caricature offers a way to make known this concerning limitation. Lack of caricature based on our measure does not mean that a simulation is necessarily acceptable or high quality (see Section 10).

**Implicit Defaults** The least caricatured personas are also those that others have found to be implicit defaults in LLMs. Implicit defaults in LLMs may shift depending on the prompt, context, etc., as well as including aspects of identity and social factors that may be invisible or underrepresented in existing empirical data and surveys[4]. Given the increasing proliferation of generated content and a limited quantity of human-written text (Xue et al., 2023; Shumailov et al., 2023), caricatures become only more relevant with the prospect of future LLMs that are trained on generated data: what will their

---

[4]Existing studies often exclude various demographics: studies on human behavior oversample from the population of American college students, who have been shown to differ from other populations in significant ways (Segall et al., 1966). Psychology and social science research subjects are disproportionately Western, educated, and from industrialized, rich, and democratic countries (Henrich et al., 2010).

defaults be, and how might they further amplify caricature?

## 10 Limitations

While we fill a critical gap since there is no existing work on systematically detecting stereotypes/caricatures in simulations or evaluating simulations in this manner at all, our measure is limited in scope: it is not a comprehensive evaluation of the quality of a simulation. We quantify susceptibility to caricature, which is a particular failure case of a simulation. Our method may yield false positives (simulations that seem acceptable and caricature-free based on our method but have other problems).

Avoiding caricature is a necessary but insufficient criterion for simulation quality; our metric should be used in tandem with other evaluations. As a pilot study for a recently-emerging direction of work, we hope to lay the groundwork for a more comprehensive evaluation of simulations in the future, perhaps in tandem with human evaluation.

As we provide a first step toward characterizing and evaluating LLM simulations, an area which currently lacks a shared language for discussion and comparison, we focus only on simulations of personas that reflect subpopulations such as social groups and on one-round response formats. However, our framework easily extends to other more complex or open-ended settings. For instance, for a multi-round simulation, one could apply our framework by using the full text of the simulation across the rounds. Depending on the length and structure of the simulation, one could also split the simulation into multiple parts and characterize each part's propensity to caricature.

Also, note that in the semantic axes, $P_p$ does not necessarily reflect a universal notion of the model's representation or description of that particular demographic group. It merely characterizes the words that distinguish a *simulation* of that persona, given the particular *context* c, such as an opinion from that demographic's perspective. This enables us to measure if and when a simulation is dominated by language that is a caricature of the persona, but our work is not a comprehensive evaluation of stereotypes or representations of demographic groups.

## Acknowledgments

Thank you to Tony Wang, Joon Sung Park, Omar Shaikh, Caleb Ziems, Camille Harris, and Matthias Gerstgrasser for their feedback throughout this

## References

David Adkins, Bilal Alsallakh, Adeel Cheema, Narine Kokhlikyan, Emily McReynolds, Pushkar Mishra, Chavez Procope, Jeremy Sawruk, Erin Wang, and Polina Zvyagina. 2022. Prescriptive and descriptive approaches to machine-learning transparency. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–9.

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Anu Aneja. 1993. "Jasmine," the sweet scent of exile. *Pacific Coast Philology*, pages 72–80.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

April H Bailey, Adina Williams, and Andrei Cimpian. 2022. Based on billions of words on the Internet, people= men. *Science Advances*, 8(13):eabm2463.

Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Daniel Bar-Tal, Carl F Graumann, Arie W Kruglanski, and Wolfgang Stroebe. 2013. *Stereotyping and prejudice: Changing conceptions*. Springer Science & Business Media.

Caroline Bassett. 2019. The computational therapeutic: Exploring Weizenbaum's ELIZA as a history of the present. *AI & SOCIETY*, 34:803–812.

Joseph Bates et al. 1994. The role of emotion in believable agents. *Communications of the ACM*, 37(7):122–125.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science.

*Transactions of the Association for Computational Linguistics*, 6:587–604.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Irene V Blair, Jennifer E Ma, and Alison P Lenton. 2001. Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of personality and social psychology*, 81(5):828.

Luisa N Borrell, Jennifer R Elhawary, Elena Fuentes-Afflick, Jonathan Witonsky, Nirav Bhakta, Alan HB Wu, Kirsten Bibbins-Domingo, José R Rodríguez-Santana, Michael A Lenoir, James R Gavin III, et al. 2021. Race and genetic ancestry in medicine—a time for reckoning with racism. *New England Journal of Medicine*, 384(5):474–480.

Leslie Bow. 2019. Racist cute: Caricature, kawaii-style, and the Asian thing. *American Quarterly*, 71(1):29–58.

Robert Bowman, Camille Nadal, Kellie Morrissey, Anja Thieme, and Gavin Doherty. 2023. Using thematic analysis in healthcare HCI at CHI: A scoping review. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

James DJ Brown. 2010. A stereotype, wrapped in a cliché, inside a caricature: Russian foreign policy and orientalism. *Politics*, 30(3):149–159.

Yang Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theory-grounded measurement of US social stereotypes in english language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1276–1295.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25.

Myra Cheng, Maria De-Arteaga, Lester Mackey, and Adam Tauman Kalai. 2023a. Social norm bias: Residual harms of fairness-aware algorithms. *Data Mining and Knowledge Discovery*, pages 1–27.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023b. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 1276–1295.

Alexander M Czopp, Aaron C Kay, and Sapna Cheryan. 2015. Positive stereotypes are pervasive and powerful. *Perspectives on Psychological Science*, 10(4):451–463.

Eberhard Demm. 1993. Propaganda and caricature in the first world war. *Journal of Contemporary History*, 28(1):163–192.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpaca-Farm: A simulation framework for methods that learn from human feedback.

Aparna Elangovan, Jiayuan He, and Karin Verspoor. 2021. Memorization vs. generalization: Quantifying data leakage in NLP performance evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1325–1335, Online. Association for Computational Linguistics.

Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology*, 82(6):878.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Peter Gottschalk and Gabriel Greenberg. 2011. From Muhammad to Obama: Caricatures, cartoons, and stereotypes of Muslims. *Islamophobia: The challenge of pluralism in the 21st century*, pages 191–210.

Jonathan Grudin. 2006. Why personas work: The psychological evidence. *The persona lifecycle*, 12:642–664.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic HCI research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Sil Hamilton. 2023. Blind judgement: Agent-based supreme court modelling with GPT. *arXiv preprint arXiv:2301.05327*.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on Fairness, Accountability, and Transparency*, pages 501–512.

Madeline E Heilman. 2001. Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. *Journal of Social Issues*, 57(4):657–674.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Susan C Herring and John C Paolillo. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459.

James D Hollan, Edwin L Hutchins, and Louis Weitzman. 1984. Steamer: An interactive inspectable simulation-based training system. *AI magazine*, 5(2):15–15.

Marjan Hosseinia, Eduard Dragut, and Arjun Mukherjee. 2020. Stance prediction for contemporary issues: Data and experiments. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 32–40, Online. Association for Computational Linguistics.

Hang Jiang, Doug Beeferman, Brandon Roy, and Deb Roy. 2022. CommunityLM: Probing partisan worldviews from language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6818–6826, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Randolph M Jones, John E Laird, Paul E Nielsen, Karen J Coulter, Patrick Kenny, and Frank V Koss. 1999. Automated intelligent pilots for combat flight simulation. *AI magazine*, 20(1):27–27.

Lee Jussim, Jarret T Crawford, Stephanie M Anglin, Sean T Stevens, and Jose L Duarte. 2016. Interpretations and methods: Towards a more effectively self-correcting social psychology. *Journal of Experimental Social Psychology*, 66:116–133.

Gauri Kambhatla, Ian Stewart, and Rada Mihalcea. 2022. Surfacing racial stereotypes through identity portrayal. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1604–1615.

Os Keyes, Burren Peil, Rua M Williams, and Katta Spiel. 2020. Reimagining (women's) health: HCI, gender and essentialised embodiment. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(4):1–42.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.

Kay Dian Kriz. 2008. *Slavery, sugar, and the culture of refinement: Picturing the British West Indies, 1700–1840*. Paul Mellon Centre.

Neha Kumar, Naveena Karusala, Azra Ismail, Marisol Wong-Villacres, and Aditya Vishwanath. 2019. Engaging feminist solidarity for comparative research, design, and practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008.

Calvin A Liang, Sean A Munson, and Julie A Kientz. 2021. Embracing four tensions in human-computer interaction research with marginalized people. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 28(2):1–47.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. 2023. Training socially aligned language models in simulated human society. *arXiv preprint arXiv:2305.16960*.

Li Lucy, Divya Tadimeti, and David Bamman. 2022. Discovering differences in the representation of people using contextualized semantic axes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3477–3494, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bohun Lynch. 1927. *A history of caricature*. Faber and Gwyer.

Julia M. Markel, Steven G. Opferman, James A. Landay, and Chris Piech. 2023. GPTeach: Interactive TA training with GPT-based students. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*, L@S '23, page 226–236, New York, NY, USA. Association for Computing Machinery.

Nicola Marsden and Maren Haag. 2016. Stereotypes and politics: Reflections on personas. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4017–4031.

Nicola Marsden and Monika Pröbster. 2019. Personas and identity: Looking at multiple identities to inform the construction of personas. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Melissa McCracken, Miho Olsen, Moon S Chen Jr, Ahmedin Jemal, Michael Thun, Vilma Cokkinides, Dennis Deapen, and Elizabeth Ward. 2007. Cancer incidence, mortality, and associated risk factors among Asian Americans of Chinese, Filipino, Vietnamese, Korean, and Japanese ethnicities. *CA: a cancer journal for clinicians*, 57(4):190–205.

Amita Misra, Brian Ecker, and Marilyn Walker. 2016. Measuring the similarity of sentential arguments in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 276–287, Los Angeles. Association for Computational Linguistics.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229.

Chandra Mohanty. 1988. Under Western eyes: Feminist scholarship and colonial discourses. *Feminist review*, 30(1):61–88.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

OpenAI. 2023. GPT-4 technical report. *arXiv*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023a. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

Peter S Park, Philipp Schoenegger, and Chongyang Zhu. 2023b. Artificial intelligence in psychology research. *arXiv preprint arXiv:2302.07267*.

Andi Peng, Besmira Nushi, Emre Kiciman, Kori Inkpen, and Ece Kamar. 2022. Investigations of performance and bias in human-AI teamwork in hiring. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12089–12097.

David Perkins. 1975. A definition of caricature and caricature and recognition. *Studies in Visual Communication*, 2(1):1–24.

Scott Ed Plous. 2003. *Understanding prejudice and discrimination.* McGraw-Hill.

Jen'nan Ghazal Read, Scott M Lynch, and Jessica S West. 2021. Disaggregating heterogeneity among non-hispanic whites: evidence and implications for us racial/ethnic health disparities. *Population research and policy review*, 40:9–31.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada. Association for Computational Linguistics.

Johannes Schneider, Christian Meske, and Michalis Vlachos. 2020. Deceptive AI explanations: Creation and detection. In *International Conference on Agents and Artificial Intelligence*. SciTePress.

Marshall H Segall, Donald Thomas Campbell, and Melville Jean Herskovits. 1966. *The influence of culture on visual perception*, volume 310. Bobbs-Merrill Indianapolis.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arxiv:2305.17493*.

Phillip R Slavney. 1984. Histrionic personality and antisocial personality: caricatures of stereotypes? *Comprehensive Psychiatry*, 25(2):129–141.

Irene Solaiman and Christy Dennison. 2021. Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, 34:5861–5873.

Keita Takayama. 2017. Imagining east asian education otherwise: Neither caricature, nor scandalization. *Asia Pacific Journal of Education*, 37(2):262–274.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290.

Andrea Aler Tubella, Dimitri Coelho Mollo, Adam Dahlgren Lindström, Hannah Devinney, Virginia Dignum, Petter Ericson, Anna Jonsson, Timotheus Kampik, Tom Lenaerts, Julian Alfredo Mendez, et al. 2023. ACROCPoLis: A descriptive framework for making sense of fairness. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1604–1615.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*.

Kailas Vodrahalli, Roxana Daneshjou, Tobias Gerstenberg, and James Zou. 2022. Do humans trust advice more if it comes from AI? an analysis of human-AI interactions. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 763–777.

Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. 2022. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 336–349.

Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Robert Wolfe and Aylin Caliskan. 2022. Markedness in visual semantic AI. *2022 ACM Conference on Fairness, Accountability, and Transparency*.

Fuzhao Xue, Yao Fu, Wangchunshu Zhou, Zangwei Zheng, and Yang You. 2023. To repeat or not to repeat: Insights from scaling llm under token-crisis. *arXiv preprint arXiv:2305.13230*.

Diyi Yang. 2019. *Computational Social Roles*. Ph.D. thesis, Carnegie Mellon University Pittsburgh, PA, USA.

JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. NormBank: A knowledge bank of situational social norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776, Toronto, Canada. Association for Computational Linguistics.

## A  Examples of Caricatures in Simulation

Which our paper focuses on how to discover problematic patterns of caricature at the larger scale used in LLM simulations, these patterns are clearly evident on an individual example level. For example, for the "nonbinary person" persona in the online forum setting, which has some of the highest exaggeration scores, some randomly-sampled default-topic responses are:

> "As a nonbinary individual, I often feel like I don't fit into the traditional boxes society has created for gender. This can be difficult when navigating spaces that are dominated by cisnormativity. However, I have found solace and empowerment in connecting with other nonbinary people and advocating for inclusivity and understanding. It is important to remember that gender is a social construct, and it is up to us to define our own identities and push for a more accepting world."

> "As a nonbinary individual, I often find myself in situations where people are not aware of or

don't understand my gender identity. While it can be frustrating at times, I choose to take it as an opportunity to educate others and promote a more inclusive society. It's important to recognize and respect everyone's experiences and identities, even if they differ from our own."

These descriptions focus solely on general statements regarding gender identity and inclusion. We find the same phenomenon in simulations of a nonbinary person's perspective on other topics. For the topic "how to read guitar music," some randomly-sampled simulated non-binary person's responses are:

> "Hey everyone! I just wanted to share my experience as a nonbinary individual learning how to read guitar music. **It's been a really empowering journey for me, as it allows me to express myself in a way that transcends the limits of gender.** Moreover, I've found that the guitar community is incredibly accepting and supportive, which has been a huge help in overcoming some of the challenges that I've faced as a nonbinary person. Keep rocking out, friends!"

> "As a nonbinary guitarist, I wanted to share some advice and encouragement for others who may be navigating the world of guitar music. Remember that music is a universal language and it does not adhere to any specific gender norms or societal expectations. **Allow yourself the freedom to explore different styles, techniques, and genres of guitar music that resonate with you.** Embrace your unique perspective and use it to create music that reflects your authentic self. And most importantly, don't forget to have fun and enjoy the journey of learning and growing as a musician."

Rather than actually discussing "how to read guitar music," these outputs disproportionately focus on generic statements about gender, authenticity, and inclusion (bolded). These responses exaggerate this particular imagined aspect of a nonbinary person's perspective. Similarly, for the topic "cars and vehicles," some randomly-sampled simulated non-binary person's responses are:

> "I just wanted to share my experience as a nonbinary individual when it comes to cars and other vehicles. For me, it's important that my mode of transportation reflects my identity and personal aesthetics. I've found that customizing my vehicle with various accessories, paint jobs, and even custom license plates has been a **fun and empowering way to express myself**. I also appreciate a vehicle that has a gender-neutral design, as that aligns with my identity. I've noticed that more and more auto manufacturers are starting to create designs that feel **more inclusive and neutral**, which is great for people like me. Ultimately, the vehicle you choose should be a reflection of your personality and preferences, no matter your gender identity. Happy driving, everyone!"

> "Hey everyone! As a nonbinary individual, I wanted to share my perspective on car preferences

| Topic | Output for "Man" persona |
|---|---|
| Default | "I recently purchased a new smartphone and I'm really impressed with its features and performance. The camera quality is amazing, and the battery life lasts all day without any issues. The interface is easy to use, and the overall design is sleek and modern. I would highly recommend this phone to anyone in the market for a new device."<br><br>"I recently purchased a new laptop and I'm absolutely loving it so far. The performance is top-notch, the design is sleek, and the battery life has been really impressive. The only issue I've come across is a minor software bug, but that was easily fixed with an update. Overall, I highly recommend this model to anyone in the market for a reliable and high-performing laptop." |
| How to read guitar music | "I recently started learning how to play the guitar and I found this guide on How to Read Guitar Music extremely helpful. The step-by-step explanation, along with images and examples, made it easy to understand and follow. I'm now more confident in reading sheet music and applying it to my guitar practice. Highly recommend this guide for any beginner guitarist looking to learn how to read music. Kudos to the author for making it so simple and easy to follow!"<br><br>"I recently started learning how to read guitar music and found this article very helpful. It provided clear explanations and easy-to-follow steps that have greatly improved my understanding of how to read sheet music for the guitar. The tips and examples given were a great reference for me as a beginner, and I feel much more confident in my ability now. Thanks for sharing this informative guide!" |
| Cars and vehicles | "I recently purchased a 2020 Ford Mustang GT and I couldn't be happier with my decision. The drive is smooth, the car is powerful, and the interior is incredibly comfortable and stylish. I have always loved the look of the Ford Mustang and this new model definitely lives up to my expectations. The sound system is fantastic as well. I highly recommend this car to anyone considering a sporty, powerful, and fun vehicle."<br><br>"I recently bought a 2015 Honda Civic and I'm absolutely loving it! The fuel efficiency is amazing and the ride is smooth. I've also noticed a significant improvement in the handling compared to my old car. The interior is quite spacious and comfortable too. My only complaint so far is the touchscreen entertainment system - it's a bit slow and not very user-friendly. But overall, I highly recommend the 2015 Civic to anyone in the market for a reliable, fuel-efficient, and enjoyable vehicle!" |

Table A1: **Examples of Simulated "Man" Responses**

and how certain vehicles **cater to different gender expressions**. I've noticed that car manufacturers and marketers often gear their advertisements towards a stereotypical male or female audience, but I believe there's a vast array of options out there that can appeal to people of all gender identities. For example, I personally love sleek, modern designs that have a balance of form and function and don't scream 'masculinity' or 'femininity.' I'd love to hear about your favorite vehicles and how they play a role in expressing your unique personality and identity. **Let's have an open and inclusive conversation about our diverse experiences** with cars and other vehicles!"

In these outputs, the bolded phrases also present unrealistic caricatures that exaggerate specific aspects of the persona rather than reflecting the full range of possibilities of what a non-binary person might say about these topics. In stark contrast, for the "man" persona, the default-topic responses are not identity-related, and responses for particular topics are much more topical and do not exaggerate any aspect of the "man" persona (Table A1).

## B Robustness of Individuation Measure

The notinon of measuring individuation is broadly agnostic to the choice of differentiator and model used to embed the texts. In Figure A1, we show individuation results from (a) using the unsupervised V-measure to measure differentiation and (b) using the all-distilroberta-v1 model to compute embeddings instead of all-mpnet-base-v2

(all-distilroberta-v1 is the next highest-performing model for general-use purposes). For (a), we first use K-Means to cluster the embeddings into two clusters, and then report the v-score. We find that the scores are overall slightly lower than using a binary classifier, which makes sense since the unsupervised method is less powerful in this context, but most of the personas can still be differentiated at a rate higher than random chance. We use the supervised binary classifier in the main results since the purpose of this metric is to reflect a reader's capability to differentiate between the two categories, which is more accurately reflected by the more powerful classifier. The broad patterns remain the same as the findings reported in the paper: all the personas can be individuated in the Interview context, and in the Online Forum context, marginalized personas are the most easily individuated, while the personas of man and woman are challenging to individuate at all.

## C Internal and External Validation of Semantic Axes

For internal validation, following Lucy et al. (2022), we verify that the constructed axes are internally valid using a leave-one-out method: consider a pole $P_1$ with a set of words $W$, where the other pole on the semantic axis is $P_2$. We remove one of the seed words $w \in W$ and construct a new pole $P_1'$
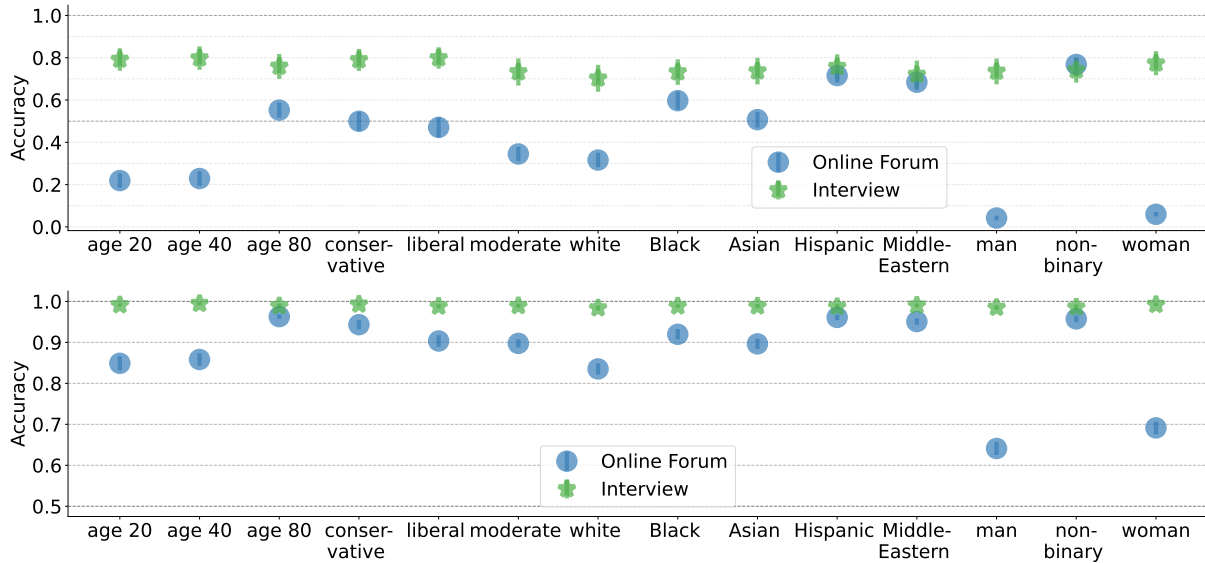
Figure A1: Top: Using the unsupervised V-measure to measure differentiation results in similar patterns as our main result in Figure 4. Bottom: Using an alternative pre-trained model to encode the outputs also results in similar patterns.

that does not include $w$. Then, we measure the cosine similarity of w's contextualized embedding to $P_1'$ and $P_2$. We found that the former is larger for all of the persona-topic semantic axes that we constructed.

For external validation, we manually inspect the sets of seed words and find that for each axis, it is easy to differentiate which set is associated with the corresponding persona versus topic. We find that for race/ethnicity and gender personas, the sets of top words reflect similar stereotypes as reported by Cheng et al. (2023b). See Table A2 and A3 for randomly-sampled examples of persona-topic axes seed word sets.

## D Experimental Details: Topics and Personas

Examples of each setting are in Table A4. Following Santurkar et al. (2023) and Cheng et al. (2023b), the personas used in the online forum and interview contexts are:

- **age:** 20-year-old person, 40-year-old person, 80-year-old person

- **political ideology:** conservative person, liberal person, moderate person

- **race/ethnicities:** Asian person, Black person, Hispanic person, Middle-Eastern person, white person

- **gender:** man, non-binary person, woman

- **default:** person

### D.1 Online Forum Context

The full list of topics collected from WikiHow and ProCon.org for the online forum context are in `topics/onlineforumtopics.txt` in the supplementary material.

The general prompt for simulation with persona $p$ and topic $t$ is: 'A(n) $p$ posted the following comment on $t$ to an online forum." For persona $p$, the default-topic prompt is: "A(n) $p$ posted the following comment to an online forum." For topic $t$, the default-persona prompt is: "A person posted the following comment on $t$ to an online forum."

### D.1.1 Fine-Grained Specificity Experiment

The topics for the fine-grained specificity experiment are:

- **Specificity Level 1:** health

- **Specificity Level 2:** mental health

- **Specificity Level 3:** things to do for mental health, social media's effect on mental health

- **Specificity Level 4:** What is the best thing you did for your mental health?, Is social media bad for your mental health?

- **Specificity Level 5:** During a very dark period, what was the best thing you ever did for your mental health?, To what extent do you think social media is bad for your mental health?
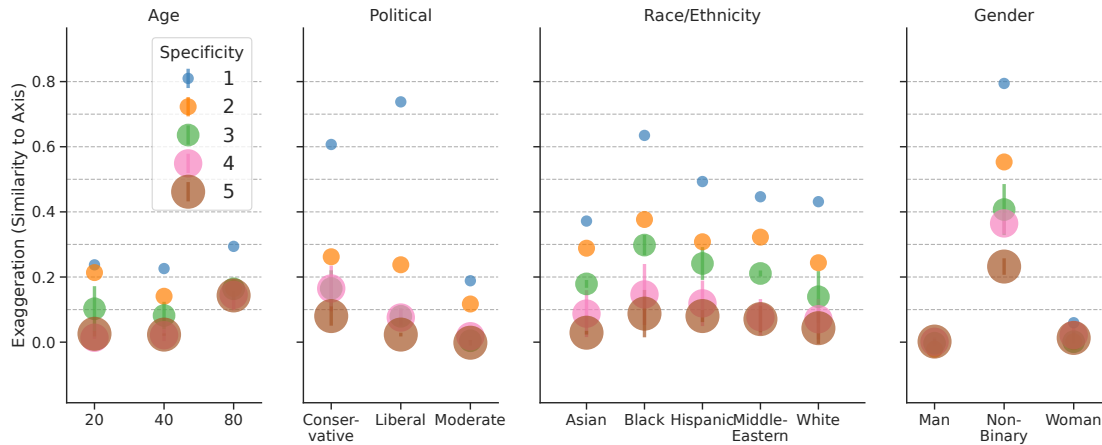
10868

Figure A2: **Exaggeration scores in the online forum context for topics varying in specificity.** Across different personas (x-axis), exaggeration score (y-axis) is negatively correlated with the specificity of the topic (marker size).

The Level 1 topic is the topic with highest rate of caricature. The Level 5 topics are from the top search results for "health" in the subreddit community AskReddit. The authors constructed the intermediate specificity levels by interpolating specificity between these. The resulting exaggeration scores are in Figure A2. We see that specificity and exaggeration are negatively correlated.

### D.2 Interview Context

The full list of Pew opinion survey questions that we use as topics in the interview setting are in topics/pewtopics.txt in the supplementary material. The default-topic and default-persona prompts are in Table A5.

### E Power Analysis

To justify 100 examples per simulation setting, we provide the following power analysis. Note that for individuation, a simulation that cannot be individuated would have score 0.5 (random chance). Across personas and topics, the lowest mean score was 0.65, and the highest standard deviation was 0.2. A power analysis using a t-test for two independent samples reveals that the necessary sample size is 28 given the effect size $(0.65 - 0.5)/0.2 = 0.75$, alpha $= 0.5$, and desired power $= 0.8$. Similarly, for exaggeration, a simulation with no exaggeration of a persona would have score 0. We found that simulations with less-specific topics and personas of many political ideology, race, and marginalized groups result in high exaggeration scores. Among these simulations, the lowest mean score was 0.23, and the highest standard deviation was 0.37. Again using a power analysis, the necessary sample size is 41 using effect size $(0.23 - 0)/0.37 = 0.62$, alpha

$= 0.5$, and desired power $= 0.8$. Thus, our choice of obtaining 100 samples per simulation is more than sufficient to achieve the desired power for both the individuation and exaggeration metrics.

### F Influence of the Context Dimension

In this section, we explore the effect of the context dimension. To verify that the trends we observe are not due to the difference in context alone, we also experiment with switching the contexts and topics, i.e., 1) simulations with the **online forum** topics in the **interview** context 2) simulations with the **interview** topics in the **online forum** setting. Figure A3 reveals that the trends in caricature rates persist for the same topics in different contexts rather than being based on context alone. Interestingly, the exaggeration scores are overall slightly higher in these switched contexts than in the original contexts. This may be impacted by memorization dynamics (Elangovan et al., 2021; Tirumala et al., 2022; Carlini et al., 2023), a relationship to explore in future work (e.g., are *memorized* topics less susceptible to caricature?).

### G Twitter Context

We additionally conduct and analyze experiments in the Twitter context introduced by Jiang et al. (2022), and we find similar trends that corroborate the results of the main paper.

### G.1 Experimental Details

Jiang et al. (2022) demonstrate how LLM simulations of Republican and Democrat Twitter users result in opinions about public figures and groups that correspond to the outcomes of the Ameri-
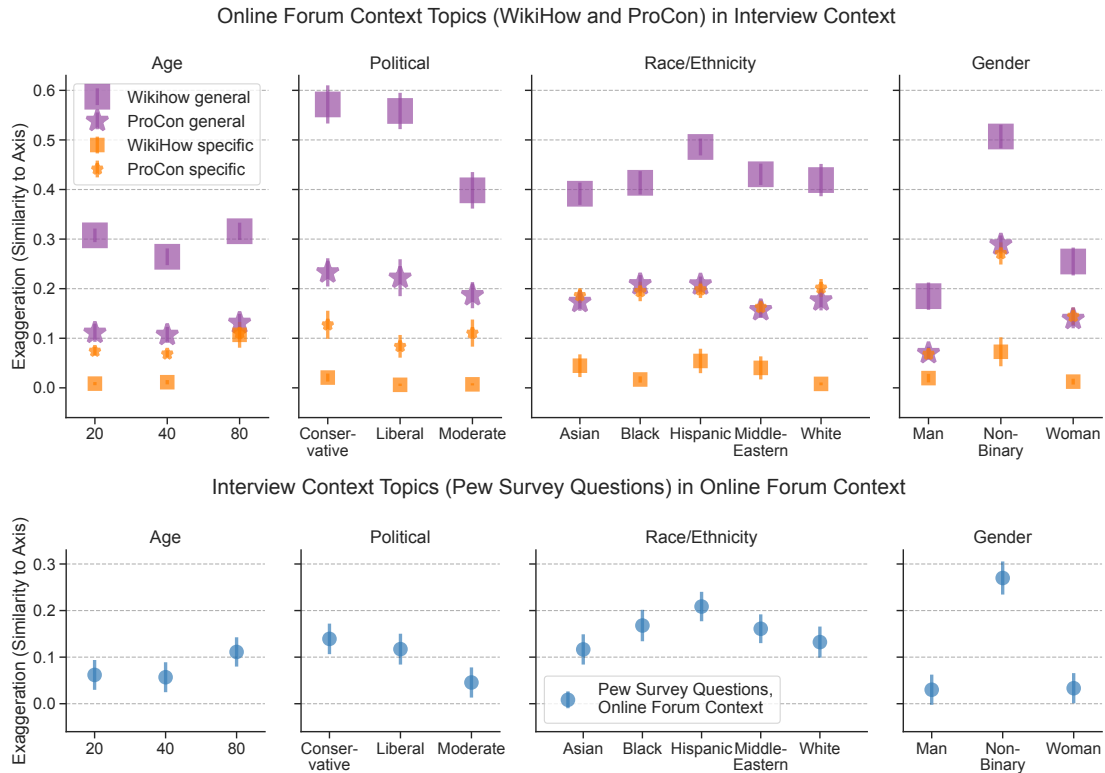
Figure A3: Exaggeration scores for simulations with the topics used in the online forum context (from WikiHow and ProCon.org), but using the interview context instead (top), and vice versa (bottom). We find similar patterns, indiciating that we can attribute the trends in scores to topics rather than context alone.

can National Election Studies (ANES) 2020 Exploratory Testing Survey. To evaluate susceptibility to caricature in this context, we use their best-performing prompt, *[persona] user posted on Twitter: [topic] is/are*. Following their work, we use the same two personas (*Republican* and *Democrat*) and 30 topics that they use: "(a) 16 people: Donald Trump, Barack Obama, Joe Biden, Elizabeth Warren, Bernie Sanders, Pete Buttigieg, Kamala Harris, Amy Klobuchar, Mike Pence, Andrew Yang, Nancy Pelosi, Marco Rubio, Alexandria Ocasio-Cortez, Nikki Haley, Clarence Thomas, Dr. Anthony Fauci, (b) 14 groups: black people, white people, Hispanic people, Asian people, illegal immigrants, feminists, the #MeToo movement, transgender people, socialists, capitalists, big business, labor unions, the Republican Party, the Democratic Party."

## G.2 Results

**Individuation** In the Twitter context, the average score for the Democrat and Republican personas are 0.94 and 0.88 respectively.

**Exaggeration** In the Twitter context (Figure A4), Republican personas resulted in higher mean exaggeration scores than Democrat ones. Topics related

to race and other groups broadly have higher exaggeration scores, while topics of specific public figures have lower ones, reflecting the pattern that more general topics are more caricatured. Discussion of Republican public figures yielded more exaggeration than those of Democratic public figures. Examining topic subcategories reveals more nuanced relationships between topic and persona. For example, in simulations with Democrat personas, topics related to race yielded higher rates of caricature. We found that among all the public figures that are used as topics, the topic *Dr. Anthony Fauci* had the lowest rate of caricature. Dr. Fauci is also the only non-politician among them.

## H Interview Setting Result Details

Exaggeration scores in the interview setting are in Figure A5. The topics with highest exaggeration scores relate to confidence in the military (*How much confidence, if any, do you have in the military to act in the best interests of the public?*), self-identified social class ( *If you were asked to use one of these commonly used names for the social classes, which would you say you belong in?*), and ties to community (*How important is it to you,*
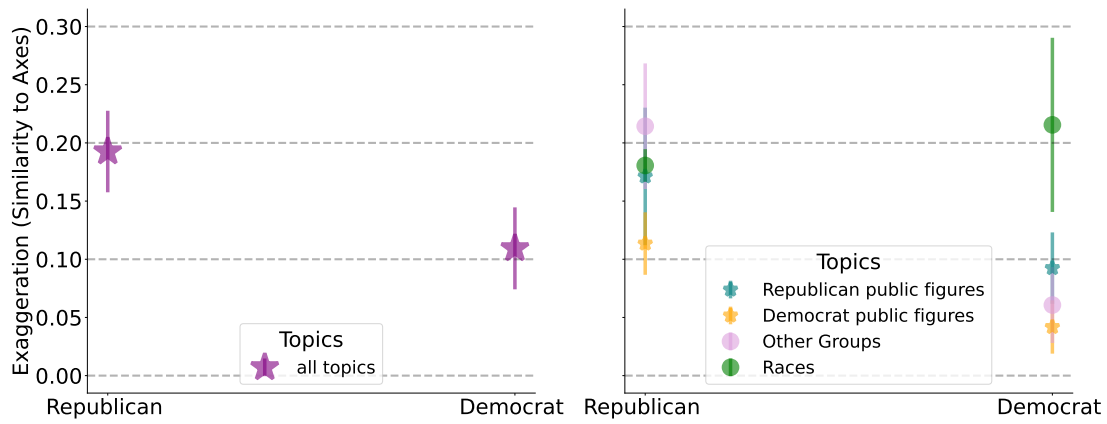
Figure A4: **Mean exaggeration scores for topics and personas in the Twitter context, averaged over all topics (left) and by topic category (right).** On average, simulations of Republican personas (left) and simulations with Republican topics (right, teal stars) result in higher exaggeration scores. However, the relationship between topics and personas is nuanced: for simulations of Democrat users, topics related to race have relatively higher exaggeration scores.
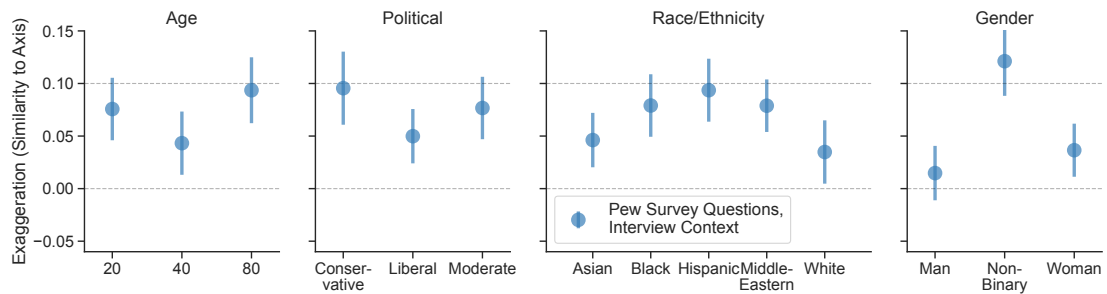


Figure A5: **Exaggeration scores for topics and personas in the interview context.** The topics are randomly-sampled controversial questions from OpinionQA (Santurkar et al., 2023). These scores are comparable to those for the specific topics in the online forum context.

*personally, to live in a community with a strong sense of community?*).

| Simulation | Persona Pole Seed Words | Topic Pole Seed Words |
|---|---|---|
| Persona: **white person**, Topic: whether the United States should implement a universal basic income | i, racism, white, color, understand, experiences, inclusive, person, together, we, privilege, listen, educate, important, race | income, ubi, would, basic, could, universal, states, united, implementing, poverty, financial, net, potential, safety, the |
| Persona: **Black person**, Topic: Historic Statue Removal | black, experiences, person, together, race, work, and, conversations, racism, for, share, open, everyone, my, understanding | statues, history, past, removing, them, the, instead, context, historic, represent, mistakes, figures, erase, these, they |
| Persona: **Asian person**, Topic: How to Write a Business Case | our, asian, we, cultures, person, people, stereotypes, us, an, and, learn, world, lets, culture, diversity | case, business, the, a, project, article, helpful, writing, tips, write, stepbystep, solution, problem, examples, found |
| Persona: **Hispanic person**, Topic: Relationships | hispanic, our, community, society, culture, person, heritage, diverse, of, and, proud, rich, traditions, important, celebrate | relationship, been, partner, but, advice, how, years, now, any, ive, dont, them, if, things, lately |
| Persona: **Middle-Eastern person**, Topic: Universal Basic Income (UBI) | middle, east, our, we, and, region, culture, understanding, history, from, cultures, diverse, rich, share, us | ubi, income, basic, universal, could, idea, poverty, society, jobs, their, automation, provide, potential, net, safety |
| Persona: **woman**, Topic: Holidays and Traditions | vacuum, cleaner, i, it, my, this, recently, product, anyone, new, skin, has, recommend, highly, suction | traditions, holidays, holiday, our, together, love, bring, cultures, of, celebrate, favorite, world, special, memories, family |
| Persona: **man**, Topic: How to Know if You Love Someone | laptop, new, the, performance, battery, recently, purchased, recommend, highly, extremely, overall, features, am, anyone, impressed | love, you, them, if, someone, when, their, happiness, truly, feelings, about, article, feel, think, that |
| Persona: **nonbinary person**, Topic: How to Dress For a Funeral | gender, nonbinary, identity, identities, understanding, my, binary, everyone, society, identify, experiences, male, female, people, share | funeral, family, dress, colors, deceased, wearing, black, the, appropriately, dark, wear, attended, show, conservative, dressing |

Table A2: **Persona-topic semantic axes seed words for randomly-sampled simulations with race/ethnicity and gender personas.** Top **persona** and topic words identified by our method (Section 4.3) to construct persona-topic semantic axes for measuring exaggeration. We display the seed words corresponding to simulations of each persona alongside a randomly-sampled topic in the online forum context; this table is continued with word sets for age and political ideology personas in Table A3. Each word list is truncated to 15 words in this table, but we use the full set in constructing the axes. We find that the race/ethnicity and nonbinary persona seed words reflect the stereotypes and essentializing narratives documented by Cheng et al. (2023b).

| Simulation | Persona Pole Seed Words | Topic Pole Seed Words |
|---|---|---|
| Persona: **20-year-old person**, Topic: <mark>How to Say Happy Birthday</mark> | im, yearold, forward, advice, forum, from, ive, discussions, looking, hey, everyone, been, experiences, thoughts, now | year, day, filled, birthday, happy, wishing, joy, may, laughter, love, another, memories, , happiness, lots |
| Persona: **80-year-old person**, Topic: <mark>Food and Entertaining</mark> | technology, am, all, world, hello, yearold, learn, how, online, learning, changes, discussions, from, has, experiences | dinner, recipes, love, hosting, a, parties, out, party, food, dishes, absolutely, favorite, delicious, friends, tips |
| Persona: **40-year-old person**, Topic: <mark>whether the federal minimum wage should be increased</mark> | i, my, experiences, im, ive, yearold, and, am, forward, from, age, discussions, cheers, hello, you | wage, minimum, federal, workers, the, increased, raising, businesses, increasing, would, living, higher, economy, could, not |
| Persona: **conservative person**, Topic: <mark>Personal Care and Style</mark> | government, our, values, responsibility, believe, society, traditional, strong, limited, we, intervention, individual, principles, that, should | my, care, style, products, started, feel, me, confidence, taking, skincare, ive, recently, skin, tips, your |
| Persona: **liberal person**, Topic: <mark>whether colleges and universities should pay college athletes</mark> | we, society, everyone, our, regardless, social, equal, healthcare, race, rights, gender, justice, policies, together, inclusive | athletes, college, their, sports, universities, colleges, they, schools, financial, these, revenue, would, paying, compensation, could |
| Persona: **moderate person**, Topic: <mark>Electoral College</mark> | opinions, perspectives, and, different, find, ground, common, open, understanding, discussions, various, important, respectful, us, engage | the, vote, system, popular, electoral, college, states, outdated, candidate, votes, of, direct, president, gives, win |

Table A3: **Persona-topic semantic axes seed words for randomly-sampled simulations with age and political ideology personas.** Examples continued from Table A2.

---

"An **Asian person** *posted the following comment about* <mark>health</mark> *to an online forum*" (Similar to (Park et al., 2022))

"*Below you will be asked to provide a short description of your identity and then answer some questions.*

*Description: In politics today, I would describe my political views as* **liberal.**

<mark>How much, if at all, do you worry about the following happening to you? Being the victim of a terrorist attack:</mark>" (Similiar to (Santurkar et al., 2023))

"A **Republican** *user posted on Twitter*: <mark>labor unions</mark> *are*" (Jiang et al., 2022)

Table A4: Mapping dimensions in different simulations. Each prompt contains **persona** (**bolded**), *context* (*italicized*), and <mark>topic</mark> (<mark>highlighted</mark>).

| Default-Topic (persona $p$) | Prompt |
|---|---|
| Age | Below you will be asked to provide a short description of your identity and then answer some questions.<br>Description: I am $p$.<br>Comment: |
| Political Ideology | Below you will be asked to provide a short description of your identity and then answer some questions.<br>Description: In politics today, I would describe my political views as $p$.<br>Comment: |
| Race/Ethnicity | Below you will be asked to provide a short description of your identity and then answer some questions.<br>Description: I am $p$.<br>Comment: |
| Gender | Below you will be asked to provide a short description of your identity and then answer some questions.<br>Description: I identify as $p$.<br>Comment: |
| **Default-Persona** (topic $t$) | Below you will be asked to provide a short description of your identity and then answer some questions.<br>Description: I am a person.<br>Question: $t$<br>Answer: |

Table A5: Prompts for default-persona and default-topic simulations in the interview context.

| Paper | Context | Models | Personas | Topics |
|---|---|---|---|---|
| Argyle et al. (2023) | public survey (MC) | GPT-3 davinci | public survey participants (political ideology, church attendance, state, race, gender, income, age) | presidential election |
| | public survey (MC) | GPT-3 davinci | variety of ANES participants with specific backstories | demographic data about self |
| Aher et al. (2023) | social science experiments (MC) | GPT-3 davinci | social science experiment participants with varying names and genders | Ultimatum Game, Garden Path Sentences, Milgram Shock Experiment, and Wisdom of Crowds |
| Park et al. (2023b) | psychology surveys and questionnaires (MC) | GPT-3.5 | liberals and conservatives | psychology studies from the Many Labs 2 replication project |
| Santurkar et al. (2023) | public survey (MC) | GPT-3.5 & AI21 models | 60 U.S. demographic groups including age, gender, race, political leaning, political party | Pew public opinion polls |
| Binz and Schulz (2023) | canonical cognitive psychology experiments (vignette-based and task-based) (MC) | GPT-3 | psychology experiment participants | questions to evaluate decision-making, information search, deliberation, and causal reasoning abilities |
| Hämäläinen et al. (2023) | HCI research interview (O) | GPT-3 | interviewees for HCI researcher | experiencing video games as art |
| Park et al. (2023a) | interactive sandbox environment inspired by The Sims, small town of twenty-five agents (O) | GPT-3.5 turbo | 25 agents of Smallville, incl. occupation and relationship with other agents | interactions with other agents and reactions to changes in environment |
| Markel et al. (2023) | office hours for a computer science class with students and a teaching assistant (O) | GPT-3 | students (descriptions include age, major, characteristics, mindset) | computer science homework assignments |
| Liu et al. (2023) | virtual societies with social norms (O) | GPT-3.5, GPT-4 | social agents | social interactions and dynamics |
| Dubois et al. (2023) | pairwise preference tasks that require human annotation (MC) | GPT-4 | human annotators (crowdworkers) | questions from OASST, Anthropic, Vicuna, and Koala evaluations |

Table A6: **Additional Examples Mapping Existing Work Using the CoMPosT Framework.** Extended from Figure 1. MC and O mean multiple-choice and open-response respectively. Note that all works use versions of GPT for the model dimension.