Virtual Personas for Language Models via an Anthology of Backstories

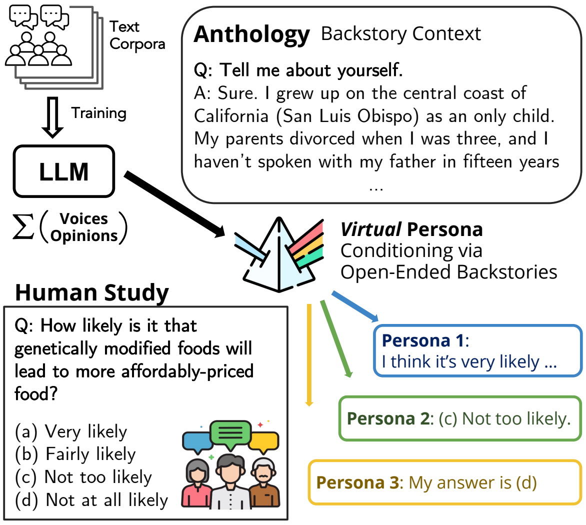
Anonymous ACL submission

Abstract

Large language models (LLMs) are trained from vast repositories of text authored by millions of distinct authors, reflecting an enormous diversity of human traits. While these models bear the potential to be used as approximations of human subjects in behavioral studies, prior efforts have been limited in steering model responses to match individual human users. In this work, we introduce "*Anthology*", a method for conditioning LLMs to particular *virtual personas* by harnessing open-ended life narratives, which we refer to as "backstories." We show that our methodology enhances the consistency and reliability of experimental outcomes while ensuring better representation of diverse sub-populations. Across three nationally representative human surveys conducted as part of Pew Research Center's American Trends Panel (ATP), we demonstrate that *Anthology* achieves up to 18% improvement in matching the response distributions of human respondents and 27% improvement in consistency metrics.

1 Introduction

Large language models (LLMs) are trained from vast repositories of human-written text (Touvron et al., 2023; Meta, 2024; Brown et al., 2020; OpenAI, 2024; MistralAI, 2024; Jiang et al., 2024a). These texts are authored by millions of distinct authors, reflecting an enormous diversity of human traits (Choi and Li, 2024; Wolf et al., 2024). As a result, when a language model completes a prompt, the generated response implicitly encodes a mixture of voices from human authors that have produced the training text from which the completion has been extrapolated. Although this nature of language models has been overlooked due to its marginal influence in current widely-adopted usages of LLMs, such as factual question-answering (QA) and algorithmic reasoning, when the model is queried with open-ended questions or is intended to be conditioned as particular personas, it is critical to address the fact that these models inherently reflect an averaged voice from the mixture of human authorship.



Q: Tell me about yourself.
A: Sure. I grew up on the central coast of California (San Luis Obispo) as an only child. My parents divorced when I was three, and I haven't spoken with my father in fifteen years ...

Q: How likely is it that genetically modified foods will lead to more affordably-priced food?

(a) Very likely
(b) Fairly likely
(c) Not too likely
(d) Not at all likely

Persona 1: I think it's very likely ...
Persona 2: (c) Not too likely.
Persona 3: My answer is (d)

Figure 1: This work introduces *Anthology*, a method for conditioning LLMs to representative, consistent, and diverse virtual personas. We achieve this by generating naturalistic backstories, which can be used as conditioning context, and show that *Anthology* enables improved approximation of large-scale human studies compared to existing approaches in steering LLMs to represent individual human voices.

A prominent example of such a scenario with growing significance is the use of LLMs to simulate human actors in the context of behavioral studies (Argyle et al., 2023; Binz and Schulz, 2023; Santurkar et al., 2023; Perez et al., 2022; Park et al., 2022; Simmons, 2022; Karra et al., 2023; Hartmann et al., 2023; Jiang et al., 2022; Aher et al., 2023; Abdulhai et al., 2023). LLMs have great potential as querying models is much faster and cheaper than designing and completing human studies (Argyle et al., 2023), a process well-known to be challenging when striving to recruit large-enough, representative, and just samples of subjects. While there are evident risks from LLMs themselves (Bommasani et al., 2022b; Bai et al., 2022; Hendrycks et al., 2023), including the inherent biases within models trained on internet data, the use of language models to perform approximate pilot studies can help survey designers satisfy best practices (Belmont Principles (Government, 1978)) of beneficence and justice, without and before inflicting potential harm to real human respondents.

For language models to effectively serve as virtual subjects, we must be able to steer their responses to reflect particular human users, *i.e.* condition models to reliable *virtual personas*. To this end, existing work prompts LLMs with context that explicitly spell out the demographic and personal traits of the intended persona: for example, (Santurkar et al., 2023; Liu et al., 2024a; Kim and Lee, 2024; Hwang et al., 2023) attempt to steer LLM responses with a dialog consisting of a series of question-answer pairs about demographic indicators, a free-text biography listing all traits, and a portrayal of the said persona in second-person point-of-view. While these approaches have shown modest success, they have been limited in (i) closely representing the responses of human counterparts, (ii) consistency, and (iii) successfully binding to diverse personas, especially those from under-represented sub-populations.

So how might we condition LLMs to virtual personas that are *representative, consistent*, and *diverse*? In this work, we investigate the use of naturalistic bodies of text describing individual life-stories, namely *backstories*, as prefix to model prompts for persona conditioning. The intuition is that open-ended life narratives both explicitly and implicitly embody diverse details about the author, including age, gender, education level, emotion, and beliefs, etc. (Argamon et al., 2007; Bantum and Owen, 2009; Schwartz et al., 2013; De Choudhury et al., 2021; Stirman and Pennebaker, 2001). Lengthy backstories thus narrowly constrain the user characteristics, including latent traits as personality or mental health that are not solicited explicitly (McAdams, 1993; Bruner, 1991), and strongly condition LLMs to diverse personas.

In particular, we suggest a methodology to generate backstories from LLMs themselves, as a means to efficiently produce massive sets covering a wide range of human demographics—which we refer to as an *Anthology* of backstories. We also introduce a method to sample backstories to match a desired distribution of human population. Our overall methodology is validated with experiments approximating well-documented large-scale human studies conducted as part of Pew Research Center's American Trends Panel (ATP) surveys. We demonstrate that language models conditioned with LLM-generated backstories provide closer approximations of real human respondents in terms of matching survey response distributions and consistencies, compared to baseline methods. Particularly, we show superior conditioning to personas reflecting users from under-represented groups, with improvements of up to 18% in terms Wasserstein Distance and 27% in consistency.

Our contributions are summarized as follows:

- We introduce *Anthology*, which employs LLM-generated backstories to further condition LLM outputs, demonstrating that *Anthology* more accurately approximates human response distributions across three surveys covering various topics and diverse demographic sub-groups (Sections 4.1 and 4.2).
- We describe a method for matching virtual subjects conditioned by backstories to target human populations. This approach significantly enhances the approximation of human response distributions (Section 4.3).
- We provide an open-source anthology of approximately 10,000 backstories for future research and applications in a broad spectrum of human behavioral studies. Additionally, we make the code for producing, processing, and administering surveys publicly available.

2 Conditioning LLMs to Virtual Personas via an *Anthology* of Backstories

In this section, we discuss details of the proposed *Anthology* approach. We start with answering the core question: What are backstories and how might they help condition LLMs to particular personas when given as context? With an example, we examine and lay out the advantages of conditioning models with backstories in Section 2.1.

There are two practical considerations when using backstories as conditioned virtual personas for approximating human subjects. In the following sections, we discuss how we address each of these implications: (i) We must acquire a substantial set of backstories that reflects a sufficient variety of human authors, since the target human study may require arbitrary demographic distribution of

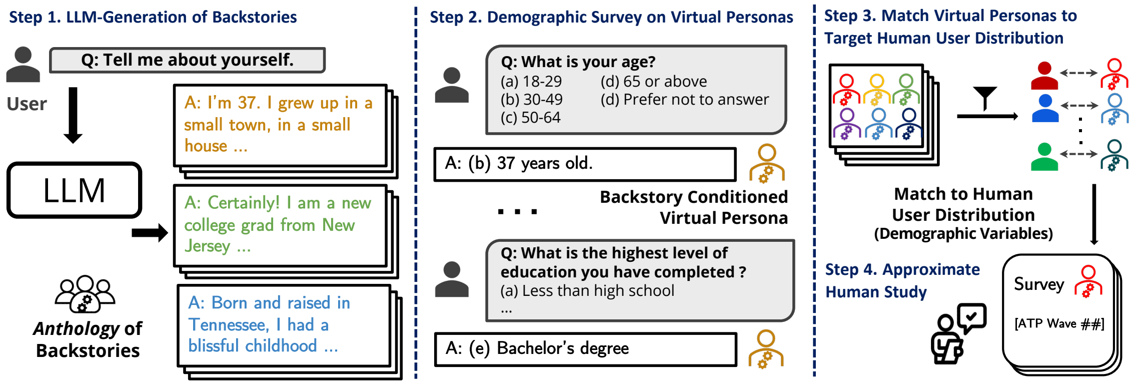

Step 1. LLM-Generation of Backstories
Q: Tell me about yourself.
User
LLM
A: I'm 37. I grew up in a small town, in a small house ...
A: Certainly! I am a new college grad from New Jersey ...
A: Born and raised in Tennessee, I had a blissful childhood ...
Anthology of Backstories
Step 2. Demographic Survey on Virtual Personas
Q: What is your age?
(a) 18-29
(b) 30-49
(c) 50-64
(d) 65 or above
(d) Prefer not to answer
A: (b) 37 years old.
Backstory Conditioned Virtual Persona
Q: What is the highest level of education you have completed ?
(a) Less than high school
...
A: (e) Bachelor's degree
Step 3. Match Virtual Personas to Target Human User Distribution
Match to Human User Distribution (Demographic Variables)
Step 4. Approximate Human Study
Survey
[ATP Wave ##]

Figure 2: Step-by-step process of the *Anthology* approach which operates in four stages. First, we leverage a language model to generate an anthology of backstories using an unrestrictive prompt. Next, we perform demographic surveys on each of these backstory-conditioned personas to estimate the persona demographics. Following this, we methodologically select a representative set of virtual personas that match a desired distribution of demographics, based on which we administer the survey. We find that our approach can closely approximate human results (see Section 4 for details).

subjects. To this end, we introduce LLM-generated backstories to efficiently generate diverse backstories (Section 2.1); and (ii) We cannot *a priori* determine the possible demographic profile of a given backstory, since demographic variables may not be explicitly mentioned in a naturalistic life narrative. Hence, we introduce methods to estimate demographics of the virtual persona conditioned by each backstory (Section 2.3) and sample subsets of backstories from anthology that match target human populations (Section 2.4).

2.1 What are Backstories?

We use the term *backstories* to refer to first-person narratives that encompass various aspects of an individual's life, from where and how they grew up, their formative experiences, education, career, and personal relationships, to their values and beliefs. These stories are inherently open-ended and personal, touching upon diverse facets of the author's demographic and personality traits.

Question: Tell me about yourself.

Answer: I am in my 60s and live in the same neighborhood I have always lived in. I am not rich and by some standards might even be considered homeless. However, I could spend thousands of dollars more per month if I wanted. I am happy with my life style. I am from the backwoods of this country and grew up with very little. On a few occasions, we were starving in the woods and going to school on an empty stomach. We had a small brown paper bag for dinner a couple of nights every week. Breakfast on some days was just a big bowl of Kool-Aid™ mixed with powdered milk. My two brothers were thin and we worried about them catching a cold. . . .
On the day before payday, my mother would spend my whole allowance in the grocery store because she just could not resist those long stems of red roses for only 29 cents a stem. I would have rather had bread and milk for dinner, but I did not dare protest because I did not want to take them away from her. We were lucky to have 79 cents to last until payday.
. . .

Figure 3: Example of a LLM-generated backstory. The generated life story can reveal explicit details about the author, such as age, hometown, and financial background, while also implicitly reflecting the author's values, personality, and unique voice through the narrative's style and content.

Consider the example shown in Figure 3. We observe that the life story of both *explicitly* and *implicitly* encodes information about the author, thereby providing rich insight into who the author is. For instance, the backstory provides explicit hints about the author's age ("in my 60s"), hometown and/or region ("rural area of north Georgia"), and financial status during childhood ("grew up with very little"). But rather than being a simple listing of the aforementioned traits, the story itself embodies a natural, authentic voice of a particular human that reflects their values and personality. (McAdams, 1993; Bruner, 1991).

Our proposed approach is to condition language models with backstories by placing them as a prefixes to the LLM (Brown et al., 2020; Touvron et al., 2023) so as to strongly condition the ensuing text completion, in the same spirit of standard prompting approaches. As we see in Figure 3, backstories capture a wide range of attributes about the author through high levels of detail and are naturalistic narratives that provide realism and consistency of the persona to which the LLM is conditioned.

2.2 LLM-Generated Backstories

A collection of human-written backstories could be drawn from existing sets of autobiographies or oral history collections. The challenge, however, is both in terms of scale and diversity (Yang et al., 2023, 2022). We find that, in their current standing, publicly available sources of autobiographical life narratives and oral histories are limited in the number of samples to sufficiently approximate larger human studies.

Instead, we propose to generate conceivably realistic backstories with language models as cost-efficient alternatives. As shown in Step 1 of Figure 2, we prompt LLMs with an open-ended prompt such as, "Tell me about yourself." We specifically care for the prompt to be simple so that the model responses are unconstrained and not biased. The prompt, however, does implicitly ask for a comprehensive narrative Responding to this prompt requires the language model to generate a series of interconnected events and experiences that form a coherent life trajectory, which inherently implies consistency and progression as in Figure 3. With sampling temperature $T = 1.0$, we generate backstories that encapsulate a broad range of life experiences of diverse human users. Further details about LLM generation of backstories, including examples, are summarized in Appendix B.

2.3 Demographic Survey on Virtual Personas

As we intend to utilize virtual personas in the context of approximating human respondents in behavioral studies, it is critical that we curate an appropriate set of backstories that would condition personas representing the target human population. Each study would have a specific set of demographic variables and an estimation or accurate statistics of the demographics of its respondents. Naturalistic backstories, despite their rich details about the individual authors, are however not guaranteed to explicitly mention all demographic variables of interest. Therefore, we emulate the process of how the demographic traits of human respondents have been collected—performing demographic surveys on virtual personas, as shown in Step 2 of Figure 2.

While we use the same set of demographic questions as used in the human studies, we consider that, unlike human respondents who each have a well-defined, deterministic set of traits, LLM virtual personas should be described with a *probabilistic* distribution of demographic variables. As such, we sample multiple responses for each demographic question to estimate the distribution of traits for the given virtual persona. Further details about the process and prompts used in demographic surveys are described in Appendix E.

2.4 Matching Target Human Populations

The remaining question is: How do we choose the right set of backstories for each survey to approximate? With the results of the demographic survey, we match virtual personas to the real human population, presented as Step 3 in Figure 2. In doing so, we construct a complete weighted bipartite graph defined by the tuple, $G = (H, V, E)$.

The vertex set $H = \{h_1, h_2, ..., h_n\}$ represents the human user group with the size of n, while the other vertex set $V = \{v_1, v_2, ..., v_m\}$ represents the virtual user group with the size of m. Each vertex h_i consists of demographic traits of $i-$th human user. Specifically, $h_i = (t_{i1}, t_{i2}, ..., t_{ik})$ where k is the number of demographic variables, and t_{il} is the $l-$th demographic variable's trait of $i-$th user. Similarly, for each vertex in V, v_j comprises probability distributions of demographic variables of each virtual user, defined as $v_j = \big(P(d_{j1}), P(d_{j2}), ..., P(d_{jk})\big)$, where d_{jl} is $j-$th user's $l-$th demographic random variable and $P(d_{jl})$ is its probability distribution.

The edge set comprises $e_{ij} \in E$ which denotes the edge between h_i and v_j.The weight of an edge, $w(e_{ij})$ or equivalently $w(h_i, v_j)$, is defined as the product of the likelihoods of traits of the $j-$th virtual user that correspond to the demographic traits of the $i-$th human user. We formally define such edge weights:

$$w(e_{ij}) = w(h_i, v_j) = \prod_{l=1}^{k} P(d_{jl} = t_{il}) \qquad (1)$$

We perform bipartite matching to select the virtual personas whose demographic probability distributions are most similar to the real, human user population. The objective is to find the matching function $\pi : [n] \to [m]$, where $[n] = \{1,2,3,...n\}$ and $[m] = \{1,2,3,...m\}$ that maximize the following:

$$\pi^* = \operatorname*{argmax}_{\pi} \sum_{i=1}^{n} w(h_i, v_{\pi(i)}) \qquad (2)$$

We explore two matching methods: (1) maximum weight matching, and (2) greedy matching. First, maximum weight matching is the method that finds the optimal π^* with the objective of Eq. 2, while ensuring that π establishes a one-to-one

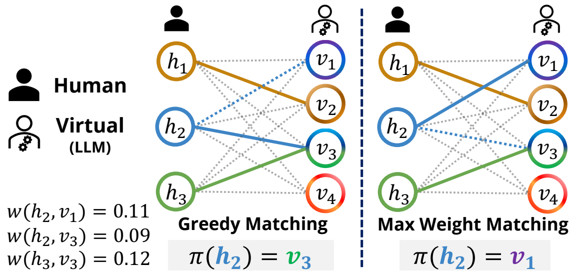


Figure 4: Matching human users to virtual personas. For greedy matching, each human user is matched to a virtual persona that has the most similar demographic traits among the virtual users. Maximum weight matching maximizes the sum of edge weights while satisfying one-to-one correspondence.

correspondence between users. We employ the Hungarian matching algorithm (Kuhn, 1955) to determine π^*. On the other hand, greedy matching seeks to maximize the same objective without requiring a one-to-one correspondence. It determines the optimal matching function such that

$$\pi^*(i) = \operatorname*{argmax}_j w(h_i, v_j) \tag{3}$$

where each human user is assigned to the virtual persona with the highest weight, allowing multiple human users assigned to the same virtual persona.

After completing the matching process, we assign the demographic traits of the target population to the matched backstories. In downstream surveys, we append these demographic information to backstories and use the matched subset of backstories, resulting in the same number of backstories as that of the target human population.

3 Approximating Human Studies with LLM Personas

In this section, we discuss the large-scale human studies that we aim to approximate (Step 4 of Figure 2) using LLM virtual subjects, based on varying methods of persona conditioning. We detail the overall experimental setup and define criteria for evaluation.

Human Study Data The Pew Research Center's American Trends Panel (ATP) is a nationally representative panel of randomly selected U.S. adults, designed to track public opinion and social trends over time. Each panel focuses on a particular topic, such as politics, social issues, and economic conditions. In this work, we consider ATP Waves 34, 92, and 99, a set of relatively recent surveys that cover a wide variety of topics: biomedical & food issues, political typology, and AI & human enhancement, respectively. In each wave, we select 6 to 8 questions from the original questionnaire that capture diverse facets of human opinions about the wave's topic using a Likert scale. Details on the questions selected and further information about each ATP wave are discussed in Appendix D.

Question: Do you think the following is generally good or bad for our society? **A decline in the share of Americans belonging to an organized religion.**

(a) Very good for society
(b) Somewhat good for society
(c) Neither good nor bad for society
(d) Somewhat bad for society
(e) Very bad for society

Figure 5: An example question (`SOCIETY_RELIG`) from ATP Wave 92 (Political Typology) that asks opinions about whether a given statement is good or bad for the American society.

Experiment Setup For each ATP survey considered, we format the select questions into language model prompts to administer survey approximations. Examples of such formatted questions are shown in Figure 5. All questions we consider are in multiple-choice question answer formats, and we carefully preserve the wording of each question and choice options from the original survey. We ask all questions *in series*—language models are given all previous questions and their answers when answering each new question. This process replicates the mental process that human respondents would undergo during surveys. For further details on prompts used and the experimental setting, see Appendix C.

Language Models We consider a suite of recent LLMs including the Meta Llama3 family (`Llama-3-70B`) (Meta, 2024) and the sparse mixture-of-experts (MoE) models from Mistral AI (`Mixtral-8x22B`) (Jiang et al., 2024a; MistralAI, 2024). We primarily focus on models with the largest number of active parameters, which roughly correlates with model capabilities and the size of the training data corpus.

Note that we primarily consider pre-trained LLMs without fine-tuning (i.e. base models). We find instruction fine-tuned models, such as by RLHF (Ouyang et al., 2022) or DPO (Rafailov et al., 2023), to be unfit for our study as their opinions are highly skewed, in particular to certain groups (e.g. politically liberal). Prior work similarly report notable opinion biases in fine-tuned models (Santurkar et al., 2023; Liu et al., 2024a;

Geng et al., 2024). More detailed discussions on chat models and their viability to be conditioned to diverse personas can be found in Appendix A.1.

Virtual Persona Conditioning Methods As baseline methods for persona conditioning, we follow (Santurkar et al., 2023) and use (i) `Bio`, which constructs free-text biographies in a rule-based manner; and (ii) `QA`, which lists a sequence of question-answer pairs about each demographic variable.

We then compare against two variants of *Anthology*: (i) Natural, refers to the use of backstories generated without any presupposed persona, as discussed in Section 2.2. In this case, we leverage either the greedy or maximum weight matching methods in Section 2.4 to select the subset to be used for each survey; (ii) Demographics-Primed, alternatively generates backstories given a particular human user's demographics to approximate, where a language model is prompted to generate a life narrative that would reflect a person of the specified demographics (for details, see Appendix B). We then append descriptions of demographic traits with the generated backstories, with which we provide as context to LLMs. Examples of prompts from each conditioning method and further details can be found in Appendix C.

Evaluation Criteria The goal of this work is to address the research question: How do we condition LLMs to representative, consistent, and diverse personas?

Representativeness: we believe that a "representative" virtual persona should successfully approximate the *first-order* opinion tendencies of their counterpart human subjects, *i.e.* respond with similar answers to individual survey questions. As questions are multiple-choice, we compare the average answer choice distributions of each question in terms of Wasserstein distance (also known as earth mover's distance). As for the representativeness across an entire set of sampled questions from a given survey, we use the average of Wasserstein distances.

Consistency: we define consistency of virtual personas in terms of their success in approximating the *second-order* response traits of human respondents, *i.e.* the correlation across responses to a set of questions in each survey. Formally, we define the consistency metric given survey response correlation matrices of virtual subjects (Σ_V) and human subjects (Σ_H) as:

$$d_{\text{cov}} = \|\Sigma_V - \Sigma_H\|_F \tag{4}$$

where $\|\cdot\|_F$ is the Frobenius norm. We additionally consider Cronbach's alpha as a measure of internal consistency independent of ground-truth human responses.

Diversity: we define the success of conditioning to diverse virtual subjects by measuring the representativeness and consistency of virtual personas in approximating human respondents belonging to particular demographic sub-groups.

4 Experimental Results

In this section, we describe experimental results that validate the effectiveness of our proposed methodology for approximating human subjects in behavioral studies.

4.1 Human Study Approximation

We evaluate the effectiveness of different methods for conditioning virtual personas in the context of approximating three Pew Research Center ATP surveys: Waves 34, 92, and 99, described in Section. 3. Prior to analyzing virtual subjects, we first estimate the lower bounds of each evaluation metric: the average Wasserstein distance (WD), Frobenius norm (Fro.), and the Cronbach's alpha (α), which are shown in the last row of Table 1. This involves repeatedly dividing the human population into two equal-sized groups at random and calculating these metrics between the sub-groups. We take averaged values from 100 iterations to represent the lower-bound estimates.

The results are summarized in Table 1. We consistently observe that *Anthology* outperforms other conditioning methods with respect to all metrics, for both the Llama-3-70B and the Mixtral-8x22B. Comparing two matching methods, the greedy matching method tends to show better performance on the average Wasserstein distance across all Waves. We attribute the differences in different matching methods to the one-to-one correspondence condition of maximum weight matching and the limited number of virtual users we have available. Specifically, the weights assigned to the matched virtual subjects in maximum weight matching are inevitably lower than those

Table 1: Results on approximating human responses for Pew Research Center ATP surveys Wave 34, Wave 92, and Wave 99, which were conducted in 2016, 2021, and 2021 respectively. We measure three metrics: (i) WD: the average Wasserstein distance between human subjects and virtual subjects across survey questions; (ii) Fro.: the Frobenius norm between the correlation matrices of human and virtual subjects; and (iii) α: Cronbach's alpha, which assesses the internal consistency of responses. *Anthology* (DP) refers to conditioning with demographics-primed backstories, while *Anthology* (NA) represents conditioning with naturally generated backstories (without presupposed demographics). Boldface and underlined results indicate values closest and the second closest to those of humans, respectively. These comparisons are made with the human results presented in the last row of the table.

Model	Persona Conditioning	Persona Matching	ATP Wave 34			ATP Wave 92			ATP Wave 99		
			WD (↓)	Fro. (↓)	α (↑)	WD (↓)	Fro.(↓)	α (↑)	WD (↓)	Fro.(↓)	α (↑)
Llama-3-70B	`Bio`	n/a	0.254	1.107	0.673	0.348	1.073	0.588	0.296	0.809	0.733
	`QA`	n/a	0.238	1.183	0.681	0.371	1.032	0.664	0.327	0.767	0.740
	Anthology (DP)	n/a	0.244	1.497	0.652	0.419	0.965	**0.636**	0.302	1.140	0.669
	Anthology (NA)	max weight	0.229	1.287	0.693	0.337	1.045	0.637	0.327	**0.686**	**0.756**
		greedy	**0.227**	**1.070**	**0.708**	**0.313**	**0.973**	0.650	**0.288**	0.765	0.744
Mixtral-8x22B	`Bio`	n/a	0.260	1.075	0.698	**0.359**	0.851	0.667	0.237	1.092	0.687
	`QA`	n/a	0.347	1.008	0.687	0.429	0.911	0.599	0.395	1.086	0.684
	Anthology (DP)	n/a	**0.236**	1.095	0.684	0.378	**0.531**	0.624	**0.215**	1.422	0.604
	Anthology (NA)	max weight	0.257	0.869	**0.726**	0.408	0.846	0.610	0.353	**0.843**	**0.729**
		greedy	0.247	**0.851**	0.715	0.392	0.981	**0.627**	0.320	0.951	0.710
Human			0.057	0.418	0.784	0.091	0.411	0.641	0.081	0.327	0.830

assigned in greedy matching, as the latter relaxes the constraints on one-to-one correspondence. This discrepancy can result in a lower demographic similarity between the matched human and virtual users when compared to the counterpart from greedy matching. These results suggest that the richness of the generated backstories in our approach can elicit more nuanced responses compared to baselines.

4.2 Approximating Diverse Human Subjects

We further evaluate *Anthology* against other baseline conditioning methods in terms of the *Diversity* criterion outlined in Section 3. To do this, we categorize users into subgroups based on race (White and non-White) and age (18-49, 50-64, and 65+ years old) with the data from ATP Survey Wave 34. The results of comparisons involving other demographic variables are detailed in Appendix A.2. We choose the Llama-3-70B model and *Anthology* using natural backstories and with greedy matching as our method and employ evaluation metrics as in Section 4.1.

As summarized in Table 2, *Anthology* outperforms other methods. Notably, *Anthology* achieves the lowest average Wasserstein distances and the highest Cronbach's alpha for all sub-groups. Specifically, the gap in the Wasserstein distance between *Anthology* and the second-best method is 0.029 for the 18-49+ age group, showing a 14.5% difference . These results validate that *Anthology* is effective in approximating diverse demographic populations than prior methods.

Intriguingly, for every subgroup except those aged 18-49, all methods show worse average Wasserstein distance compared to the results approximating the entire human respondents presented in Tab. 1. For instance, the average Wasserstein distance for *Anthology* in the ATP Wave 34 survey is 0.227, while it increases to 0.242 for the 50-64, and 0.303 for the 65+ age groups. Conversely, for the 18-49 age group, *Anthology* shows a lower average Wasserstein distance of 0.2 compared to 0.227. This finding is consistent with prior research arguing that language model responses tend to be more inclined towards younger demographics (Santurkar et al., 2023; Liu et al., 2024b).

4.3 Sampling Backstories to Match Target Demographics

Next, we study the effect of matching strategies, greedy and max weight matching. In Tab. 3, we compare these methods with random matching, which assigns the traits of the target demographic group to randomly sampled backstories. This comparison is conducted on ATP Wave 34 using both Llama-3-70B and Mixtral2-8x22B models.

We observe that our matching methods consistently outperform random matching in terms of the average Wasserstein distance across all models. Notably, for example, with Llama-3-70B, the average Wasserstein distance between random matching and greedy matching shows an 18% difference. The gap is even more pronounced in the Frobenius norm, marking a 27% difference. This result implies that inconsistent matching between backstories and the target human distribution can significantly impact the effectiveness of the metrics. Therefore, careful matching is crucial to ensure the reliability and validity of the results in our study.

Table 2: Results on subgroup comparison. Target population is divided into demographic subgroups, and representativeness and consistency are measured within each subgroup. *Anthology* consistently results in lower Wasserstein distances, lower Frobenius norm, and higher Cronbach's alpha. Boldface and underlined results indicate values closest and the second closest to those of humans, respectively. These comparisons are made with the human results presented in the last row of the table.

Method	Race						Age Group								
	White			Non-White			18-49			50-64			65+		
	WD ($\downarrow$)	Fro. ($\downarrow$)	α ($\uparrow$)	WD ($\downarrow$)	Fro. ($\downarrow$)	α ($\uparrow$)	WD ($\downarrow$)	Fro. ($\downarrow$)	α ($\uparrow$)	WD ($\downarrow$)	Fro. ($\downarrow$)	α ($\uparrow$)	WD ($\downarrow$)	Fro. ($\downarrow$)	α ($\uparrow$)
`Bio`	0.263	**1.187**	0.687	0.335	0.955	0.651	0.244	1.163	0.673	0.277	1.382	0.659	0.318	1.000	0.686
`QA`	0.250	1.259	0.678	0.323	0.828	0.687	0.229	**1.091**	0.695	0.258	1.220	0.695	0.329	1.204	0.630
Anthology	**0.233**	1.216	**0.703**	**0.311**	**0.778**	**0.719**	**0.200**	1.193	**0.702**	**0.242**	**1.215**	**0.710**	**0.303**	**0.943**	**0.704**
Human	0.063	0.519	0.777	0.094	0.413	0.764	0.077	0.663	0.779	0.092	0.741	0.803	0.102	0.772	0.766

Table 3: Study on the effects of different matching methods. We compare max weight matching, greedy matching, and random matching. We report two metrics: (i) the average Wasserstein distance across survey questions, and (ii) the distance between the correlation matrices of human and virtual subjects.

Model	Method	ATP Wave 34	
		WD ($\downarrow$)	Fro. ($\downarrow$)
Llama-3-70B	random	0.270	1.362
	max weight	0.229	1.287
	greedy	0.227	1.070
Mixtral-8x22B	random	0.274	0.814
	max weight	0.257	0.869
	greedy	0.247	0.851

5 Related Work

Generating Personas with LLMs Recent advancements in language model applications have expanded into simulating human responses for psychological, economic, and social studies (Karra et al., 2023; Aher et al., 2023; Binz and Schulz, 2023; Horton, 2023; Fatouros et al., 2024; Argyle et al., 2023). Specifically, the generation of personas using LLMs to respond to textual stimuli has been explored in various contexts including human-computer interaction (HC), multi agent system, analysis on biases in LLMs, and personality evaluation. (Kim et al., 2020; Simmons, 2022; Park et al., 2022; Santurkar et al., 2023; Jiang et al., 2024b; Choi and Li, 2024; Liu et al., 2024a; Wu et al., 2024; Li et al., 2023; Hilliard et al., 2024; Serapio-García et al., 2023; Hu and Collier, 2024; Hwang et al., 2023; Abdulhai et al., 2023). For instance, Park et al. (2022) and Santurkar et al. (2023) develop methods to prime LLMs with crafted personas, influencing the models' outputs to simulate targeted user responses. Additionally, Liu et al. (2024a) introduces a method where personas are generated by sampling demographic traits coupled with either congruous or incongruous political stances. Our approach, *Anthology*, advances this concept by employing dynamically generated, richly detailed backstories that include a broad spectrum of demographic and economic characteristics, enhancing the granularity and authenticity of simulated responses.

LLMs in Social Science Studies The integration of LLMs into social science research has been steadily gaining attention, as highlighted by several studies (Bail et al., 2023; Park et al., 2023a; Dillion et al., 2023; Ziems et al., 2023; Korinek, 2023). Notably, the use of LLMs to mimic human responses to survey stimuli has gained popularity, as evidenced by recent research (Tjuatja et al., 2023; Dominguez-Olmedo et al., 2023; Kim and Lee, 2024). A notable example is the "media diet model" by Chu et al. (2023), which predicts consumer group responses based on their media consumption patterns. Further, studies like (Wu et al., 2023) and (Ziems et al., 2023) demonstrate the potential of LLMs in zero-shot learning settings to analyze political ideologies and scale computational social science tools. Our work builds on these methodologies by using LLMs not only to generate responses but to create and manipulate backstories that reflect diverse societal segments, providing a nuanced tool for social science research and beyond.

6 Conclusion

In this paper, we have proposed and tested a method, *Anthology*, for the generation of diverse and specific backstories. We have demonstrated that this method closely aligns with real-world demographics and demonstrates substantial potential in emulating human-like responses for social science applications. While promising, the method also highlights critical limitations and ethical concerns that must be addressed. Future advancements must focus on enhancing the representation and consistency of virtual personas to ensure their beneficial integration into societal studies.

7 Limitations and Ethical Considerations

This work introduces *Anthology*, a new methodology for conditioning large language models (LLMs) on dynamically generated, narrative-driven backstories, effectively simulating human-like personas. This approach exploits the diverse human experiences embedded within the training data, enhancing the applicability of virtual personas in social sciences and beyond. However, despite promising results, the approach encapsulates inherent limitations and significant societal implications which warrant careful consideration.

7.1 Limitations

This study, while advancing the application of LLMs in social sciences through *Anthology*, acknowledges several inherent limitations:

- **Simulation Fidelity:** We do not suggest that LLMs can fully simulate a given human user merely by using a user's backstory as a prompt prefix. Instead, we propose *Anthology* as a more effective means of engaging with virtual personas that can emulate the first-order response distributions observed in human studies. The scope of our findings is confined to LLMs conditioned on backstories and limited to structured survey questionnaires without encompassing any free-form responses.

- **Data Dependence:** The personas generated are only as diverse and unbiased as the data underlying the training of the LLMs. If the training data is skewed or non-representative, the resulting personas may inadvertently perpetuate these biases.

- **Contextual Binding:** While backstories provide a rich context for generating personas, the current models may not consistently apply this context across different types of queries or interactions, leading to variability in persona consistency.

- **Technical Constraints:** The computational cost associated with training and deploying state-of-the-art LLMs conditioned with detailed backstories is substantial, which may limit the scalability of this approach for widespread practical applications.

- **Ethical Concerns:** There is an ongoing concern regarding the ethical use of virtual personas, especially regarding privacy, consent, and the potential for misuse in scenarios like deep fakes or manipulation in political and social spheres.

These limitations highlight the need for ongoing research to refine *Anthology*, ensuring its ethical application and enhancing its realism and effectiveness in approximating human-like personas. Future directions involve improving the diversity of backstories to better reflect underrepresented groups and integrating multimodal data to enrich persona simulations. Further, exploring the effects of different conditioning techniques could deepen our understanding of the ethical and practical implications of these virtual personas. Ultimately, refining these methodologies through iterative feedback and adjustments will be crucial in advancing the field toward more ethically informed and effective applications.

7.2 Societal Impact

Employing LLMs to create virtual personas presents both transformative possibilities and ethical challenges. Positively, it could significantly impact market research, psychological studies, and the simulation of social behaviors, providing cost-effective and rapid data collection while minimizing risks to real individuals. Conversely, there exists a potential for misuse, such as influencing public opinion or perpetuating biases through skewed data representations. Such risks highlight the imperative for stringent ethical oversight and regulation in deploying these technologies to safeguard against misuse.

References

Marwa Abdulhai, Gregory Serapio-Garcia, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. 2023. Moral foundations of large language models. *Preprint*, arXiv:2310.15337.

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. *Preprint*, arXiv:2208.10264.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Yoshua Bengio, Danqi Chen, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. 2024. Foundational challenges in assuring alignment and safety of large language models. *Preprint*, arXiv:2404.09932.

Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2007. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9).

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*, page 1–15.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Christopher A Bail, D Sunshine Hillygus, Alexander Volfovsky, Maxwell B Allamong, Fatima Alqabandi, Diana Jordan, Graham Tierney, Christina Tucker, Andrew Trexler, and Austin van Loon. 2023. Do we need a social media accelerator?

Erin O'carroll Bantum and Jason E Owen. 2009. Evaluating the validity of computerized content analysis programs for identification of emotional expression in cancer narratives. *Psychol Assess*, 21(1):79–88.

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

Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy S Liang. 2022a. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? In *Advances in Neural Information Processing Systems*, volume 35, pages 3663–3678. Curran Associates, Inc.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022b. On the opportunities and risks of foundation models. *Preprint*, arXiv:2108.07258.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jerome Bruner. 1991. The narrative construction of reality. *Critical Inquiry*, 18(1):1–21.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Hyeong Kyu Choi and Yixuan Li. 2024. Beyond helpfulness and harmlessness: Eliciting diverse behaviors from large language models with persona in-context learning. In *International Conference on Machine Learning*.

Eric Chu, Jacob Andreas, Stephen Ansolabehere, and Deb Roy. 2023. Language models trained on media diets can predict public opinion. *Preprint*, arXiv:2303.16779.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2021. Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):128–137.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dunner. 2023. Questioning the survey responses of large language models. *ArXiv*, abs/2306.07951.

Georgios Fatouros, Konstantinos Metaxas, John Soldatos, and Dimosthenis Kyriazis. 2024. Can large language models beat wall street? unveiling the potential of ai in stock selection. *ArXiv*, abs/2401.03737.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Mingmeng Geng, Sihong He, and Roberto Trotta. 2024. Are large language models chameleons? *Preprint*, arXiv:2405.19323.

US Government. 1978. *The Belmont Report : Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. CreateSpace Independent Publishing Platform.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *Preprint*, arXiv:2301.01768.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An overview of catastrophic ai risks. *Preprint*, arXiv:2306.12001.

Airlie Hilliard, Cristian Munoz, Zekun Wu, and Adriano Soares Koshiyama. 2024. Eliciting personality traits in large language models. *Preprint*, arXiv:2402.08341.

John J. Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? *Preprint*, arXiv:2301.07543.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations. *Preprint*, arXiv:2402.10811.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L'elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024a. Mixtral of experts. *ArXiv*, abs/2401.04088.

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.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024b. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. 2023. Estimating the personality of white-box language models. *Preprint*, arXiv:2204.12000.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will I sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In *Proceedings of the 2020*

Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 904–916, Online. Association for Computational Linguistics.

Junsol Kim and Byungkyu Lee. 2024. Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction. *Preprint*, arXiv:2305.09620.

Anton Korinek. 2023. Language models and cognitive automation for economic research. Technical report, National Bureau of Economic Research.

Harold W. Kuhn. 1955. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Preprint*, arXiv:2211.09110.

Andy Liu, Mona Diab, and Daniel Fried. 2024a. Evaluating large language model biases in persona-steered generation. *Preprint*, arXiv:2405.20253.

Siyang Liu, Trish Maturi, Bowen Yi, Siqi Shen, and Rada Mihalcea. 2024b. The generation gap:exploring age bias in the underlying value systems of large language models. *Preprint*, arXiv:2404.08760.

D.P. McAdams. 1993. *The Stories We Live by: Personal Myths and the Making of the Self*. W. Morrow.

Meta. 2024. Meta llama 3.

MistralAI. 2024. Mixtral-8x22b.

OpenAI. 2024. Gpt-4o.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023a. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

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*, UIST '22, New York, NY, USA. Association for Computing Machinery.

Peter S. Park, Philipp Schoenegger, and Chongyang Zhu. 2023b. Diminished diversity-of-thought in a standard large language model. *Preprint*, arXiv:2302.07267.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2022. Discovering language model behaviors with model-written evaluations. *Preprint*, arXiv:2212.09251.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *Preprint*, arXiv:2308.11483.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *Preprint*, arXiv:2303.17548.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M.

Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE*, 8(9):1–16.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *Preprint*, arXiv:2307.00184.

Gabriel Simmons. 2022. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. *Preprint*, arXiv:2209.12106.

S. W. Stirman and J. W. Pennebaker. 2001. Word use in the poetry of suicidal and nonsuicidal poets.

Lindia Tjuatja, Valerie Chen, Sherry Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. 2023. Do llms exhibit human-like response biases? a case study in survey design. *ArXiv*, abs/2311.04076.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Yotam Wolf, Noam Wies, Oshri Avnery, Yoav Levine, and Amnon Shashua. 2024. Fundamental limitations of alignment in large language models. *Preprint*, arXiv:2304.11082.

Patrick Y. Wu, Jonathan Nagler, Joshua A. Tucker, and Solomon Messing. 2023. Large language models can be used to scale the ideologies of politicians in a zero-shot learning setting. *Preprint*, arXiv:2303.12057.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversation.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023. DOC: Improving long story coherence with detailed outline control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3378–3465, Toronto, Canada. Association for Computational Linguistics.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. *Preprint*, arXiv:2210.06774.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? *Preprint*, arXiv:2305.03514.

Appendix

Appendix A provides additional experimental results, including results on using instruction-tuned models with *Anthology*.

Appendix B provides further details regarding how backstories (both natural and demographic-primed) are generated.

Appendix C provides additional experimental details.

Appendix D describes the human studies (Pew Research Center ATP Waves) in detail.

Appendix E provides additional details regarding the demographic survey component of the *Anthology* method.

A Additional Experimental Results

A.1 Results on Other Models

In this section, we conduct the ATP W34 survey with various models, including fine-tuned models like Llama-3-70B-Instruct, Mixtral-8x22B-v0.1, GPT-3.5-0125, and a smaller model, Llama-3-7B. Notably, none of the fine-tuned models show better metrics in both *Representativeness* and *Consistency* criteria, which are defined in Section 3. Despite these models achieving better results on several benchmarks (Gao et al., 2023; Hendrycks et al., 2021; Chiang et al., 2024), they do not adequately approximate human responses for this survey. Additionally, the other interesting observation is that the best-performing model in terms of approximation to human responses is Llama-3-8B, which is the smallest model among those evaluated. We hypothesize that fine-tuning LLMs including instruction fine-tune, RLHF, DPO (Rafailov et al., 2023; Ouyang et al., 2022; Chung et al., 2022) makes them converge to a singular persona (Park et al., 2023b; Anwar et al., 2024; Bommasani et al., 2022a), which makes LLMs unsuitable for the tasks that requires diverse responses. And this makes the larger fine-tuned models less capable on approximating the diverse humans' responses.

We hypothesize that fine-tuning LLMs through methods such as instruction fine-tuning, RLHF, and DPO (Rafailov et al., 2023; Ouyang et al., 2022; Chung et al., 2022) leads them to converge towards a singular persona (Park et al., 2023b; Anwar et al., 2024; Bommasani et al., 2022a). This convergence potentially renders LLMs less suitable for tasks requiring diverse responses, consequently making larger fine-tuned models less effective at approximating the varied responses of humans.

This finding aligns with the insights from (Santurkar et al., 2023) discussing that the base models are more steerable than fine-tuned models, and suggests the need for careful model selection for this specific task (Liang et al., 2023)

We observe that the Llama-3-8B model exhibits a higher Cronbach's alpha value. This increased consistency is attributed to the model's tendency to select responses same as previously generated responses (Zheng et al., 2023; Pezeshkpour and Hruschka, 2023; Zheng et al., 2024), resulting in more correlated responses over survey questions. Consequently, this leads to a higher Cronbach's alpha compared to the results shown in Table 1, even though the average Wasserstein distance is significantly higher.

Table 4: Results on approximating human responses for Pew Research Center ATP surveys Wave 34, which was conducted in 2016. We measure three metrics: (i) WD: the average Wasserstein distance between human subjects and virtual subjects across survey questions; (ii) Fro.: the Frobenius norm between the correlation matrices of human and virtual subjects; and (iii) α: Cronbach's alpha, which assesses the internal consistency of responses. *Anthology* (DP) refers to conditioning with demographics-primed backstories, while *Anthology* (NA) represents conditioning with naturally generated backstories.

Model	Persona Conditioning	Persona Matching	ATP Wave 34		
			WD (↓)	Fro. (↓)	α (↑)
Llama-3-70B-Instruct	`Bio`	n/a	0.462	2.177	0.445
	`QA`	n/a	0.422	1.560	0.581
	Anthology (DP)	n/a	0.461	1.295	0.511
	Anthology (NA)	max weight	0.429	1.776	0.714
		greedy	0.413	1.848	0.754
Mixtral-8x22B-Instruct	`Bio`	n/a	0.532	1.608	0.632
	`QA`	n/a	0.567	1.583	0.628
	Anthology (DP)	n/a	0.464	1.652	0.646
	Anthology (NA)	max weight	0.478	1.606	0.635
		greedy	0.472	1.593	0.640
gpt-3.5-0125	`Bio`	n/a	0.414	2.009	0.481
	`QA`	n/a	0.422	1.560	0.581
	Anthology (DP)	n/a	0.476	1.963	0.486
	Anthology (NA)	max weight	0.450	1.905	0.472
		greedy	0.443	1.936	0.468
Llama-3-8B	`Bio`	n/a	0.454	1.480	0.683
	`QA`	n/a	0.432	0.924	0.779
	Anthology (DP)	n/a	0.383	1.323	0.714
	Anthology (NA)	max weight	0.395	1.265	0.735
		greedy	0.416	1.229	0.717
Human			0.057	0.418	0.784

A.2 Subgroup Comparisons for Other Demographic Variables

Here, continuing the discussion in Section. 4.2, we evaluate the *Diversity* criterion (Section. 3) on the methods with other subgroups. The demographic variables analyzed are education level and gender. We categorize education level into two groups:

low education level, referring to individuals with education levels up to high school graduation, and high education level, which includes those attending college or higher.

We observe a trend in Tab. 5 similar to the results in Tab. 2. *Anthology* shows the lower Wasserstein distance across all sub-groups analyed in Tab. 5. In the experiments comparing `QA` and our method in the first column, the difference in the average Wasserstein distance is 0.220, representing a 48% discrepancy. Specifically, for the female subgroup, our method demonstrates the best metrics compared to other baselines. This experiment result shows that *Anthology* is more effective in satisfying the *Diversity* criterion.

B Details on LLM-Generated Backstories

In this section, we discuss additional details about the process of generating realistic backstories using language models, as mentioned in Section 2. We detail the prompts used and examples of LLM-generated backstories.

Then, we discuss the alternative method of generating backstories given a particular combination of demographic traits, referred in Section 3 as the “Demographics-Primed” method in contrast to the “Natural” backstories generated without conditioning on demographics.

B.1 Natural Generation of Backstories

We use OpenAI’s `davinci-002` for generating backstories with the prompt specified in the top of Figure 6. This model is chosen as it is base model (*i.e.* not instruction-tuned) of the largest model capacity at the time of the project. Figure 6 shows two examples of backstories of different lengths generated with this prompt.

B.2 Generating Demographics-Primed Backstories

Target demographics-primed backstories are generated by prompting a language model with demographic information of a human from a target population. In contrast to naturally generated backstories whose demographic trait cannot be predetermined but can only can be sampled by the demographic survey method outlined in E, demographic traits of target demographics-primed backstories are determined at the time of generation. We use five demographic variables (age, annual household income, education level, race or ethnicity, gender) for ATP Wave 34, 99 and an additional variable (political affiliation) for ATP Wave 92.

A generation prompt example for ATP Wave 34 is presented in Figure 7. Answers for each question are taken from the demographic information of a human respondent in the ATP survey data. To accurately incorporate the target population’s demographic information, we use the same list of choices as used in the actual survey. Orders of demographic variables are randomized every generation to minimize the effect of question ordering. We use two styles of prompt, which we refer to a Question-Answer and a Biography as presented in Figure 7.

To take a full advantage of the demographics-

Table 5: Results on subgroup comparison. Target population is divided into demographic subgroups, and representativeness and consistency are measured within each subgroup. *Anthology* consistently results in lower Wasserstein distances, lower Frobenius norm, and high Cronbach's alpha. Boldface and underlined results indicate values closest and the second closest to those of humans, respectively. These comparisons are made with the human results presented in the last row of the table.

Method	Education Level						Gender					
	Low education level			High education level			Male			Female		
	WD ($\downarrow$)	Fro. ($\downarrow$)	α ($\uparrow$)	WD ($\downarrow$)	Fro. ($\downarrow$)	α ($\uparrow$)	WD ($\downarrow$)	Fro. ($\downarrow$)	α ($\uparrow$)	WD ($\downarrow$)	Fro. ($\downarrow$)	α ($\uparrow$)
`Bio`	0.258	1.248	**0.702**	0.252	1.166	0.673	0.257	**0.899**	**0.732**	0.297	1.038	0.679
`QA`	0.368	**1.177**	0.694	0.238	**1.101**	0.675	0.243	1.145	0.682	0.280	0.953	0.680
Anthology	**0.248**	1.227	0.680	**0.212**	1.269	**0.702**	**0.213**	1.313	0.698	**0.263**	**0.761**	**0.708**
Human	0.091	0.778	0.805	0.061	0.448	0.776	0.072	0.563	0.784	0.070	0.610	0.777

primed backstory generation, backstories should sufficiently reflect the given demographic information. Due to pre-trained base models' limited instruction following capability, however, demographics-primed backstory generated with pre-trained base models sometimes reflect demographic traits inconsistent with provided information. Threfore, We use the fine-tuned chat model Mixtral-8x22B (Jiang et al., 2024a) with decoding hyperparameters of `top_p` = 1.0, `T` = 1.1.

C Details on Experiments

In this section we provide examples of prompts used in the experiments approximating human studies, as described in Section 3 and used to produce the results in Section 4. Additionally, we outline the survey procedure for conducting these experiments, providing a comprehensive review of methodologies and operational frameworks involved.

C.1 Prompts for Baseline: `QA`

We construct a series of multiple choice demographic survey question-answer pairs given the demographic traits. The five demographic traits we use are taken from the human respondent data of ATP surveys. The order of five questions is randomized every time to minize the effect of question ordering.

C.2 Prompts for Baseline: `Bio`

As in (Santurkar et al., 2023), we construct free-text biographies in a rule-based manner given the demographic trait. The five demographic traits we use are taken from the human respondent data of ATP surveys. The order of five sentences each describing demographic traits is randomized every time to minimize the effect of sentence ordering.

C.3 Target Demographics-Primed Backstory

The details of target demographics-primed backstory used in the survey experiment are presented in Figure 9. The demographic traits used to generate the backstory and append are taken from human respondents data of ATP surveys.

C.4 Natural Backstory

The details of natural backstory used in the survey experiment are presented in Figure 10. The demographic traits appended to the backstory are traits of matched human respondents with either greedy or maximum weight sum matching.

C.5 Survey Procedure

In this study, we try our best to mimic the same survey procedure as human surveys. Human survey typically shuffle or reverse the order of the multiple choice options or change the order of questions for each survey participant to reduce the bias in the results. Typically, human surveys employ techniques like shuffling or reversing the order of multiple-choice options or altering the sequence of questions for each participant to minimize bias in the results. Following the topline reports for each wave as provided by Pew Research, we randomly reverse the order of Likert scale questions and shuffle the options for nominal questions to ensure a similar reduction in bias. For example,

D Details on Human Studies

American Trends Panel (ATP) is a nationally representative panel of U.S. adults conducted by the Pew Research Center. ATP is designed to study a wide variety of topics, including politics, religion, internet usage, online dating, and more. We analyze sampled questions from three waves, where questions are drawn from ASK ALL questions

Backstory Generation Prompt ("Natural")

Question:
Tell me about yourself. How old are you? Where did you grow up? What events made you who you are? What matters to you and what do you enjoy?

Answer:

Backstory Example 1

Answer: I'm 23 years old. I was born in the Dominican Republic and immigrated with my family to the US when I was three years old. At first we lived in the Upper West Side of Manhattan. We moved to Flushing, Queens, when I was 7 years old. I've lived there ever since, apart from about 6 months when I lived in Astoria. I've seen and met many different people and I think that's given me a deep awareness of the diversity of life and people.

I love playing sports and played basketball a lot, even up to today. I love culture and I dabble in music. I've played guitar since I was 12 and piano since I was 10. I have always been preoccupied with making a living, and I've been doing work that has little to do with music, although music has always been present in my life as a beloved hobby.

What I enjoy is peace and quiet, a home cooked meal with my parents or friends. I enjoy night walks, and in the winter you can be sure to find me hanging around downtown or Union Square every Friday or Saturday night. I enjoy seeing movies. I love the solitude of seeing movies. Although it's not always peaceful, because there's always people there who are talking and making unnecessary noise. I don't like that, but I understand that most people don't know how to behave quietly in a crowded theater. That's why the best thing to do is sit near the back of the theater in a seat that's alone. What else I enjoy is shopping and clothes, but not even for myself. I like looking at things that are rare and unique and hard to find.

Backstory Example 2

Answer: When I was 11 years old, my dad married an awesome woman who had two daughters. I was a little older, and while I liked having brothers and sisters in theory, I honestly don't recall feeling anything positive about them. When I was about 13 years old, I kind of wanted to be different. I wanted to have friends. I wanted to be like others. I began to regret my previous actions. At that time, I can recall visiting a different "church" where a certain man preached about knowing the truth and a certain girl talked to me. That evening, I decided to follow the Lord. I believe that God allowed me to have that experience so that I would be obedient. I began praying and seeking God for guidance.

When I was 16, my stepmother was moved out and my biological mom moved in. When she moved out, she wasn't forced out, and my biological dad gave her a home because she didn't have one. She lost her house when I was 9 or 10 years old. But he gave her a home and a car. I really think that he believed that she would improve her behavior with more opportunities. However, she continued to demonstrate that she was not going to be a good mother. But, he gave her a home and a car. After she left, the only advice he could give me was to stop giving her anything. Instead of being angry with him, I was simply heartbroken. I never had a mother, but she had more than I did. I have wanted children, but I know that I could be no worse than my stepmom, and maybe I could be better.

(continued)

I need to continue to pray about having children. I know that many people have problems with being a good parent, but I do not believe that anyone has the right to prevent someone from being a parent. I do not believe that someone should have a child because they are lonely and don't want to die alone. I do believe that a parent should be ready to change diapers, wake up all night, and endure endless hours of crying if necessary. I think the difficulties involved are often over-rated. I knew a young man who said he never wanted children because he did not want to wake up during the night. Before his first baby was born, he spent 5 months on the "Far Side of the Moon" (i.e., military deployment). He said he'd change his mind about the children after the experience. But, after the experience,

I suppose it is natural that I should be a "friend" to everyone because I did not have a mother. My biological mother was a welfare case. She had 9 children. She had 6 children by 3 different fathers. There was no baby-daddy drama for me (no paternal problem). She never had a full-time job. She was never the wife of a strong and godly man. She was very much a part of her local welfare system. We were on welfare from the time I was in the first grade until I graduated high school. But, I do not recall feeling poor. I recall more specific instances where I was embarrassed. When I was about 5 or 6 years old, I walked into a hairdressers to get my hair cut. The lady said, "You're too poor to get your hair cut here." Well, there goes the schoolyard conversations! As I've continued to learn about world poverty, I find that I was never poor. But, I have been in the company of the very poor. I only recall seeing a wealthy person 4 times in my life. I remember a person on TV or in a magazine once. I do remember working in America on a volunteer project in a home (not in a slum) where there were wealthy people nearby, but the people in the house were not wealthy. I actually had someone get angry with me about not helping the people in the rich community.

I had every reason to be angry with God. I grew up with drugs in the home. I grew up with alcohol in the home. I grew up with anger in the home. I was the kid no one wanted. I didn't get in trouble as a child, but I had everyone down on me. There were times when I was sexually abused, when I was hit with closed fists, and I had to experience great love from my older brothers. I wanted to be cool. I didn't feel cool. I didn't know cool people. I didn't want the life of my stepmothers (or my biological mom). When I was 16 years old, I believed that God gave me the man I saw on TV that night to become a Christian. However, I believe that God allows us to suffer some difficulties in life, but He doesn't give the suffering. All God offers to do is sustain us during the suffering. Suffering doesn't prove that God is evil, but that we are capable of evil (i.e., sin). If God allows suffering, then He knows about it, but it is not likely that He is the Author of all the evil. There is no "good" in us, except for God. In the Bible, God warned Adam that Satan would bring forth wisdom. Adam did not have evil desires. Satan "forced" Adam to sin, but God "allowed" the suffering. He allowed Adam's sin. Afterward, He allowed more suffering, but that suffering did not continue until eternity. I think that most people (adults) believe that all of God's blessings are here, and none are beyond. Therefore, suffering becomes unjust. But, all the just blessings from God take place beyond this life. Justice prevails for all eternity. Suffering is reserved for this life. It is my responsibility to continue to read about suffering in the world, and to be a "friend" to those who suffer.

Figure 6: (Top Left) Details of the prompt given to LLMs for natural backstory generation. (Rest of Figure) Two examples of backstories generated with OpenAI `Davinci-002` without presupposed demographics and with an open-ended, unrestrictive prompt.

Backstory Generation Prompt ("Demographics-Primed") Using Q/A Pairs of Demographics

Below you will be asked to complete some demographic questions, and then answer a question.

Question: What is your gender?
(A) Male
(B) Female
Answer with (A), or (B).
Answer: (A)

Question: Which of the following racial or ethnic groups do you identify with?
(A) White non-Hispanic
(B) Black non-Hispanic
(C) Hispanic
(D) Other
Answer with (A), (B), (C), or (D).
Answer: (A)

Question: What is your age?
(A) 18-29
(B) 30-49
(C) 50-64
(D) 65+
Answer with (A), (B), (C), or (D).
Answer: (D)

Question: What is the highest level of education you have completed?
(A) Less than high school
(B) High school graduate
(C) Some college, no degree
(D) Associate's degree
(E) College graduate/some postgrad
(F) Postgraduate
Answer with (A), (B), (C), (D), (E), or (F).
Answer: (F)

Question: What is your annual household income?
(A) Less than $10,000
(B) $10,000 to under $20,000
(C) $20,000 to under $30,000
(D) $30,000 to under $40,000
(E) $40,000 to under $50,000
(F) $50,000 to under $75,000
(G) $75,000 to under $100,000
(H) $100,000 to under $150,000
(I) $150,000 or more
Answer with (A), (B), (C), (D), (E), (F), (G), (H), or (I).
Answer: (B)

Question: Tell me about yourself. Please describe in detail.

Answer:

Backstory Generation Prompt ("Demographics-Primed") Using First-Person Statement of Demographics

Below you will be asked to provide a short description of your demographic information, and then answer a question.

Description: My annual income is $100,000 to under $150,000. I consider my gender as male. I consider my race as White non-Hispanic. My highest level of education is postgraduate. My age is 30-49.

Question: Tell me about yourself. Please describe in detail.

Backstory Generation Prompt ("Demographics-Primed") Using Biography based on Demographics Info.

Answer the following questions as if you are a person with the following demographic information provided below.
age: 30-49
race: White non-Hispanic
education: Postgraduate
income: $10,000 to under $20,000
gender: Male

Question: Tell me about yourself. Please describe in detail.

Demographics-Primed Backstory Example

Answer: I am a 65-year-old male who identifies as White non-Hispanic. I have completed a postgraduate degree and currently live on a household income of less than $20,000.

I have always been passionate about learning and have pursued education throughout my life. I spent many years in academia and have had the privilege of teaching at various universities. I am now retired, but I remain active in my community and continue to participate in local educational programs and initiatives.

Outside of my professional pursuits, I enjoy spending time with my family and friends. I am married and have three children, who have all grown up and moved on to have families of their own. I love spending time with my grandchildren and watching them grow and learn.

In my free time, I enjoy staying active and being outdoors. I enjoy hiking, biking, and exploring new places. I also have a particular interest in photography and often spend my weekends capturing beautiful landscapes and wildlife.

Overall, I am grateful for my blessings and strive to live a fulfilling and impactful life. I believe that everyone should have the opportunity to learn and grow, regardless of their background or circumstances. I am committed to doing my part to make that a reality.

Figure 7: (Left) Details of the prompt given to LLM for demographics-primed backstory generation. (Bottom Right) An example demographics-primed backstory generated with `Mixtral-8x22B-Instruct-v0.1` given the prompt on the left. (Rest of Figure) First-person statement and biography prompt given to LLM for the backstory generation.

(i.e. asked to all human respondents, instead of questions asked for selective demographic groups or conditionally asked based on the response to the previous question) in order to investigate the response of overall population.

It is worth noting that in the original ATP surveys, some questions have answer choices in a Likert scale with the order of choices (*e.g.* positive-to-negative or negative-to-positive) randomized for each respondent. For such questions, we also ran-

Baseline QA Persona Conditioning Method

Question: What is your annual household income?
(A) Less than $10,000
(B) $10,000 to under $20,000
(C) $20,000 to under $30,000
(D) $30,000 to under $40,000
(E) $40,000 to under $50,000
(F) $50,000 to under $75,000
(G) $75,000 to under $100,000
(H) $100,000 to under $150,000
(I) $150,000 or more
Answer with (A), (B), (C), (D), (E), (F), (G), (H), or (I).
Answer: (D)

Question: What is the highest level of education you have completed?
(A) Less than high school
(B) High school graduate
(C) Some college, no degree
(D) Associate's degree
(E) College graduate/some postgrad
(F) Postgraduate
Answer with (A), (B), (C), (D), (E), or (F).
Answer: (C)

Question: What is your age?
(A) 18-29
(B) 30-49
(C) 50-64
(D) 65+
Answer with (A), (B), (C), or (D).
Answer: (D)

Question: What is your gender?
(A) Male
(B) Female
(C) Other
Answer with (A), (B), or (C).
Answer: (A)

Question: Which of the following racial or ethnic groups do you identify with?
(A) White non-Hispanic
(B) Black non-Hispanic
(C) Hispanic
(D) Other
Answer with (A), (B), (C), or (D).
Answer: (A)

Baseline BIO Persona Conditioning Method

Below you will be asked to provide a short description of your demographic information, and then answer some questions.

Description: I consider my race as White non-Hispanic. My highest level of education is some college, no degree. My age is 65+. My annual income is $30,000 to under $40,000. I consider my gender as male.

Figure 8: Baseline prompt examples for `QA` (left) and `Bio` (right). This example shows two prompts using the same demographic trait from a randomly sampled human respondent in ATP Wave 34.

domize the order of these options when presenting them in prompts to LLMs. Here we present the list of sampled questions from each wave.

D.1 ATP Wave 34

American Trends Panel Wave 34 is conducted from April 23, 2018 to May 6, 2018 with a focus on biomedical and food issues. The number of total respondents is 2,537.

D.2 ATP Wave 92

American Trends Panel Wave 92 is conducted from July 8, 2021 to July 21, 2021 with a focus on political typology. We randomly sampled 2,500

Anthology Method with
"Demographics-Primed" Backstories
w/ Matched Demographics in Q/A format

Question: Tell me about yourself. Please describe in detail.

Answer: I am a 65-year-old male who identifies as White non-Hispanic. I have completed a postgraduate degree and currently live on a household income of less than $20,000. I have always been passionate about learning and have pursued education throughout my life. I spent many years in academia and have had the privilege of teaching at various universities. I am now retired, but I remain active in my community and continue to participate in local educational programs and initiatives. Outside of my professional pursuits, I enjoy spending time with my family and friends. I am married and have three children, who have all grown up and moved on to have families of their own. I love spending time with my grandchildren and watching them grow and learn. In my free time, I enjoy staying active and being outdoors. I enjoy hiking, biking, and exploring new places. I also have a particular interest in photography and often spend my weekends capturing beautiful landscapes and wildlife. Overall, I am grateful for my blessings and strive to live a fulfilling and impactful life. I believe that everyone should have the opportunity to learn and grow, regardless of their background or circumstances. I am committed to doing my part to make that a reality.

Question: What is your age?
(A) 18-29
(B) 30-49
(C) 50-64
(D) 65+
Answer with (A), (B), (C), or (D).
Answer: (D)

Question: Which of the following racial or ethnic groups do you identify with?
(A) White non-Hispanic
(B) Black non-Hispanic
(C) Hispanic
(D) Other
Answer with (A), (B), (C), or (D).
Answer: (A)

Question: What is the highest level of education you have completed?
(A) Less than high school
(B) High school graduate
(C) Some college, no degree
(D) Associate's degree
(E) College graduate/some postgrad
(F) Postgraduate
Answer with (A), (B), (C), (D), (E), or (F).
Answer: (F)

(continued)

Question: What is your annual household income?
(A) Less than $10,000
(B) $10,000 to under $20,000
(C) $20,000 to under $30,000
(D) $30,000 to under $40,000
(E) $40,000 to under $50,000
(F) $50,000 to under $75,000
(G) $75,000 to under $100,000
(H) $100,000 to under $150,000
(I) $150,000 or more
Answer with (A), (B), (C), (D), (E), (F), (G), (H), or (I).
Answer: (B)

Question: What is your gender?
(A) Male
(B) Female
Answer with (A), or (B).
Answer: (A)

Anthology Method with
"Demographics-Primed" Backstories
w/ Matched Demographics in Biography format

Question: Tell me about yourself. Please describe in detail.

Answer: **[Same Backstory]**

Question: Please provide your demographic information.

Answer: My highest level of education is postgraduate. I consider my race as White non-Hispanic. My annual income is $10,000 to under $20,000. My age is 65+. I consider my gender as male.

Figure 9: (Left and Top Right) An example of demographics-primed backstory, appended with demographic traits used to generate the backstory in the Q/A format. (Bottom Right) The same backstory and demographic traits, but the demographic traits are presented in the biography format.

respondents for the study from the total 10,221 respondents.

D.3 ATP Wave 99

American Trends Panel Wave 99 is conducted from November 1, 2021 to November 7, 2021 with a focus on artificial intelligence and human enhancement. We randomly sampled 2,500 respondents for the study from the total 10,260 respondents.

Anthology Method with ”Natural” Backstories w/ Matched Demographics in Q/A format

Question: Tell me about yourself. How old are you? Where did you grow up? What events made who you are? What matters to you and what do you enjoy?

Answer: I'm a 30 year old white male. I was born and raised in Tennessee, in a small town called Clarksville. It's in the middle of nowhere. I was the kid that wasn't picked for any sports team. I wasn't a nerd. I was the kid that didn't talk to anybody. I still have no friends and I still don't talk to anybody. I was homeschooled and have never had a job. I do have a GED and I did take a few community college classes. I have been out of school since 2011. My life has been filled with anxiety and depression. I've had a few minor issues in my life, like when I was a child and I was abused.

Question: What is your annual household income?
(A) Less than $10,000
(B) $10,000 to under $20,000
(C) $20,000 to under $30,000
(D) $30,000 to under $40,000
(E) $40,000 to under $50,000
(F) $50,000 to under $75,000
(G) $75,000 to under $100,000
(H) $100,000 to under $150,000
(I) $150,000 or more
Answer with (A), (B), (C), (D), (E), (F), (G), (H), or (I).
Answer: (B)

Question: Which of the following racial or ethnic groups do you identify with?
(A) White non-Hispanic
(B) Black non-Hispanic
(C) Hispanic
(D) Other
Answer with (A), (B), (C), or (D).
Answer: (A)

Question: What is the highest level of education you have completed?
(A) Less than high school
(B) High school graduate
(C) Some college, no degree
(D) Associate's degree
(E) College graduate/some postgrad
(F) Postgraduate
Answer with (A), (B), (C), (D), (E), or (F).
Answer: (C)

(continued)

Question: What is your age?
(A) 18-29
(B) 30-49
(C) 50-64
(D) 65+
Answer with (A), (B), (C), or (D).
Answer: (B)

Question: What is your gender?
(A) Male
(B) Female
(C) Other
Answer with (A), (B), or (C).
Answer: (A)

Anthology Method with ”Natural” Backstories w/ Matched Demographics in Biography Format

Question: Tell me about yourself. How old are you? Where did you grow up? What events made who you are? What matters to you and what do you enjoy?

Answer: I was born in Houston, Texas. I'm 29 years old. I grew up in Pasadena, Texas, which is right outside of Houston. I had a pretty typical childhood. I always wanted to be in law enforcement when I grew up. So I pursued that career. I ended up in law enforcement for about 4 years. I worked in a juvenile facility. I worked in investigations and at the jail. And then, I made the decision to leave law enforcement and pursue something else that would be a little bit more family friendly and that's when I found this position at the Fort Bend County Library. I started off in the circulation department. That's where most people start. I was able to move to the branch level and now I'm a regional manager for the three branches in the Richmond/Rosenberg area. I am really passionate about reading and I love my job and the people that I work with.

Question: Please provide your demographic information.

Answer: My highest level of education is high school graduate. I consider my race as White non-Hispanic. My annual income is $100,000 to under $150,000. My age is 18-29. I consider my gender as female.

Figure 10: (Left and Top Right) An example of natural backstory, appended with demographic traits of a matched human user in the Q/A format. (Bottom Right) Another example of natural backstory, this time appended with demographic traits in the biography format.

E Demographic Survey on Virtual Subjects

The goal of demographic survey is to obtain the demographic information encoded in backstories. Five demographic variables (age, annual household income, education level, race or ethnicity, and gender) and a party affiliation question are asked to backstories as they are utilized in the downstream target population matching. We take two approaches to obtain the probable demographics of authors.

In the first approach, we use GPT-4o (OpenAI, 2024) to locate demographic information from the backstory. To minimize hallucination, we prompt GPT-4o to retrieve the demographic trait only if

American Trends Panel Wave 34
Selected Questions

Please answer the following question keeping in mind your previous answers.
Question: How likely is it that genetically modified foods will lead to more affordably-priced food
(A) Not at all likely
(B) Not too likely
(C) Fairly likely
(D) Very likely
Answer with (A), (B), (C), or (D).
Answer:

Please answer the following question keeping in mind your previous answers.
Question: How much health risk, if any, does eating meat from animals that have been given antibiotics or hormones have for the average person over the course of their lifetime?
(A) No health risk at all
(B) Not too much health risk
(C) Some health risk
(D) A great deal of health risk
Answer with (A), (B), (C), or (D).
Answer:

Please answer the following question keeping in mind your previous answers.
Question: How likely is it that genetically modified foods will create problems for the environment
(A) Not at all likely
(B) Not too likely
(C) Fairly likely
(D) Very likely
Answer with (A), (B), (C), or (D).
Answer:

Please answer the following question keeping in mind your previous answers.
Question: How likely is it that genetically modified foods will lead to health problems for the population as a whole
(A) Not at all likely
(B) Not too likely
(C) Fairly likely
(D) Very likely
Answer with (A), (B), (C), or (D).
Answer:

(Continued)

Please answer the following question keeping in mind your previous answers.
Question: How much of the food you eat is organic?
(A) None at all
(B) Not too much
(C) Some of it
(D) Most of it
Answer with (A), (B), (C), or (D).
Answer:

Please answer the following question keeping in mind your previous answers.
Question: How much health risk, if any, does eating food and drinks with artificial preservatives have for the average person over the course of their lifetime?
(A) No health risk at all
(B) Not too much health risk
(C) Some health risk
(D) A great deal of health risk
Answer with (A), (B), (C), or (D).
Answer:

Please answer the following question keeping in mind your previous answers.
Question: How much health risk, if any, does eating food and drinks with artificial coloring have for the average person over the course of their lifetime?
(A) No health risk at all
(B) Not too much health risk
(C) Some health risk
(D) A great deal of health risk
Answer with (A), (B), (C), or (D).
Answer:

Please answer the following question keeping in mind your previous answers.
Question: How much do you, personally, care about the issue of genetically modified foods?
(A) Not at all
(B) Not too much
(C) Some
(D) A great deal
Answer with (A), (B), (C), or (D).
Answer:

Figure 11: 8 questions sampled from ATP Wave 34 ASK ALL questions. The prompts "Please answer the following question keeping in mind your previous answers" are included before asking each survey question.

the backstory explicitly mentions related context (prompts are shown in E.1). This approach is limited to specific demographic variables, especially age, annual household income, and education level questions, since we avoid inferring race / ethnicity, gender, and party affiliation even in the case when backstory mentions those traits. Decoding hyperparameters are set to `top_p` = 1.0, `T` = 0.

In the second approach we perform a response sampling by prompting the language model with generated backstories that are appended with demographic questions. In E.2 we present the question format. The language model's responses are sampled 40 times for each backstory and question. Instead of estimating responses with the first-token logits (Santurkar et al., 2023; Hendrycks et al., 2021; Gao et al., 2023), we allow the model to generate open-ended responses as some responses (ex. "I am 25 years old." for the age question) cannot be accurately accounted by the logit method and

American Trends Panel Wave 92
Selected Questions

Please answer the following question keeping in mind your previous answers.
Question: Do you think the following is generally good or bad for our society? Greater social acceptance of people who are transgender (people who identify as a gender that is different from the sex they were assigned at birth)
(A) Very bad for society
(B) Somewhat bad for society
(C) Neither good nor bad for society
(D) Somewhat good for society
(E) Very good for society
Answer with (A), (B), (C), (D), or (E).
Answer:

Please answer the following question keeping in mind your previous answers.
Question: Do you think the following is generally good or bad for our society? An increase in the number of guns in the U.S.
(A) Very bad for society
(B) Somewhat bad for society
(C) Neither good nor bad for society
(D) Somewhat good for society
(E) Very good for society
Answer with (A), (B), (C), (D), or (E).
Answer:

Please answer the following question keeping in mind your previous answers.
Question: Do you think the following is generally good or bad for our society? Good-paying jobs requiring a college degree more often than they used to
(A) Very bad for society
(B) Somewhat bad for society
(C) Neither good nor bad for society
(D) Somewhat good for society
(E) Very good for society
Answer with (A), (B), (C), (D), or (E).
Answer:

Please answer the following question keeping in mind your previous answers.
Question: Do you think the following is generally good or bad for our society? Increased public attention to the history of slavery and racism in America
(A) Very bad for society
(B) Somewhat bad for society
(C) Neither good nor bad for society
(D) Somewhat good for society
(E) Very good for society
Answer with (A), (B), (C), (D), or (E).
Answer:

(Continued)

Please answer the following question keeping in mind your previous answers.
Question: Do you think the following is generally good or bad for our society? Same-sex marriages being legal in the U.S.
(A) Very bad for society
(B) Somewhat bad for society
(C) Neither good nor bad for society
(D) Somewhat good for society
(E) Very good for society
Answer with (A), (B), (C), (D), or (E).
Answer:

Please answer the following question keeping in mind your previous answers.
Question: Do you think the following is generally good or bad for our society? White people declining as a share of the U.S. population
(A) Very bad for society
(B) Somewhat bad for society
(C) Neither good nor bad for society
(D) Somewhat good for society
(E) Very good for society
Answer with (A), (B), (C), (D), or (E).
Answer:

Please answer the following question keeping in mind your previous answers.
Question: Do you think the following is generally good or bad for our society? A decline in the share of Americans belonging to an organized religion
(A) Very bad for society
(B) Somewhat bad for society
(C) Neither good nor bad for society
(D) Somewhat good for society
(E) Very good for society
Answer with (A), (B), (C), (D), or (E).
Answer:

Figure 12: 7 questions sampled from ATP Wave 92 ASK ALL questions

the sum of probability masses of valid tokens (ex. " (A") are often marginal to represent the true probability distribution. Sampled responses are subsequently parsed by regex matching of either the label (ex. "(A)") or the text (ex. "27"), recorded to obtain the distribution of 40 generations. We use Llama 3 (Meta, 2024) for the response sampling with decoding hyperparameters of `top_p` = 1.0, `T` = 1.0.

Combining two approaches, our demographic survey is performed as follows. First, we use GPT-4o to locate demographic information for variables of age, annual household income, and education level. For the remaining variables and the cases where explicit demographic information cannot be found, responses are sampled 40 times to construct a response distribution. Therefore, in the case of

American Trends Panel Wave 99
Selected Questions

Please answer the following question keeping in mind your previous answers.
Question: How excited or concerned would you be if artificial intelligence computer programs could know people's thoughts and behaviors?
(A) Very concerned
(B) Somewhat concerned
(C) Equal excitement and concern
(D) Somewhat excited
(E) Very excited
Answer with (A), (B), (C), (D), or (E).
Answer:

Please answer the following question keeping in mind your previous answers.
Question: How excited or concerned would you be if artificial intelligence computer programs could make important life decisions for people?
(A) Very concerned
(B) Somewhat concerned
(C) Equal excitement and concern
(D) Somewhat excited
(E) Very excited
Answer with (A), (B), (C), (D), or (E).
Answer:

Please answer the following question keeping in mind your previous answers.
Question: How excited or concerned would you be if artificial intelligence computer programs could perform household chores?
(A) Very concerned
(B) Somewhat concerned
(C) Equal excitement and concern
(D) Somewhat excited
(E) Very excited
Answer with (A), (B), (C), (D), or (E).
Answer:

Please answer the following question keeping in mind your previous answers.
Question: How excited or concerned would you be if artificial intelligence computer programs could handle customer service calls?
(A) Very concerned
(B) Somewhat concerned
(C) Equal excitement and concern
(D) Somewhat excited
(E) Very excited
Answer with (A), (B), (C), (D), or (E).
Answer:

(Continued)

Please answer the following question keeping in mind your previous answers.
Question: How excited or concerned would you be if artificial intelligence computer programs could perform repetitive workplace tasks?
(A) Very concerned
(B) Somewhat concerned
(C) Equal excitement and concern
(D) Somewhat excited
(E) Very excited
Answer with (A), (B), (C), (D), or (E).
Answer:

Please answer the following question keeping in mind your previous answers.
Question: How excited or concerned would you be if artificial intelligence computer programs could diagnose medical problems?
(A) Very concerned
(B) Somewhat concerned
(C) Equal excitement and concern
(D) Somewhat excited
(E) Very excited
Answer with (A), (B), (C), (D), or (E).
Answer:

Figure 13: 6 questions sampled from ATP Wave 99 ASK ALL questions

sampling, virtual users' demographic trait is not represented as a single trait but rather a distribution over probable demographics given the backstory. We can thereby construct a probable estimate of demographic information without undermining the diversity of virtual authors of backstories.

E.1 Questions For Locating Demographic Information

In this section, we present the prompts to locate the demographic information that has been mentioned in the backstory. These prompts are only available for annual household income, age, and education level questions.

Prompt to Locate Mentioned Demographic Traits: Age

Question: What does the person's essay above mention about the age of the person?
(A) 18-29
(B) 30-49
(C) 50-64
(D) 65 or Above
(E) Was not mentioned
First, provide evidence that is mentioned in the text. If the answer was not mentioned, select 'Was not mentioned'. Next, answer with (A), (B), (C), (D), or (E).
Answer:

Prompt to Locate Mentioned Demographic Traits: Income

Question: What does the person's essay above mention about the annual household income the person makes?
(A) Less than $10,000
(B) $10,000 to $19,999
(C) $20,000 to $29,999
(D) $30,000 to $39,999
(E) $40,000 to $49,999
(F) $50,000 to $59,999
(G) $60,000 to $69,999
(H) $70,000 to $79,999
(I) $80,000 to $89,999
(J) $90,000 to $99,999
(K) $100,000 to $149,999
(L) $150,000 to $199,999
(M) $200,000 or more
(N) Was not mentioned
First, provide evidence that is mentioned in the text. If the answer was not mentioned, select 'Was not mentioned'. Next, answer with (A), (B), (C), (D), (E), (F), (G), (H), (I), (J), (K), (L), (M), or (N).
Answer:

Prompt to Locate Mentioned Demographic Traits: Education Level

Question: What does the person's essay above mention about the highest level of education the person has completed?
(A) Less than high school
(B) High school graduate or equivalent (e.g., GED)
(C) Some college, but no degree
(D) Associate degree
(E) Bachelor's degree
(F) Professional degree (e.g., JD, MD)
(G) Master's degree
(H) Doctoral degree
(I) Was not mentioned
First, provide evidence that is mentioned in the text. If the answer was not mentioned, select 'Was not mentioned'. Next, answer with (A), (B), (C), (D), (E), (F), (G), (H), or (I).
Answer:

Figure 14: Question prompts used to locate the explicitly mentioned demographic information from the backstory. We apply these prompts only to variables of annual household income, age, and education level.

E.2 Demographic questions

In this section, we present the questions used in demographic survey, and a political affiliation survey. Each question is asked to each virtual user 40 times to sample a probability distribution of demographic traits.

Demographic Survey Prompt:
Age Question

Question: What is your age?
(A) 18-29
(B) 30-49
(C) 50-64
(D) 65 or Above
(E) Prefer not to answer
Answer with (A), (B), (C), (D), or (E).
Answer:

Demographic Survey Prompt:
Annual Household Income Question

Question: What is your annual household income?
(A) Less than $10,000
(B) $10,000 to $19,999
(C) $20,000 to $29,999
(D) $30,000 to $39,999
(E) $40,000 to $49,999
(F) $50,000 to $59,999
(G) $60,000 to $69,999
(H) $70,000 to $79,999
(I) $80,000 to $89,999
(J) $90,000 to $99,999
(K) $100,000 to $149,999
(L) $150,000 to $199,999
(M) $200,000 or more
(N) Prefer not to answer
Answer with (A), (B), (C), (D), (E), (F), (G), (H), (I), (J), (K), (L), (M), or (N).
Answer:

Demographic Survey Prompt:
Gender Question

Question: What is your gender?
(A) Male
(B) Female
(C) Other (e.g., non-binary, trans)
(D) Prefer not to answer
Answer with (A), (B), (C), or (D).
Answer:

Demographic Survey Prompt:
Education Level Question

Question: What is the highest level of education you have completed?
(A) Less than high school
(B) High school graduate or equivalent (e.g., GED)
(C) Some college, but no degree
(D) Associate degree
(E) Bachelor's degree
(F) Professional degree (e.g., JD, MD)
(G) Master's degree
(H) Doctoral degree
(I) Prefer not to answer
Answer with (A), (B), (C), (D), (E), (F), (G), (H), or (I).
Answer: (D)

Demographic Survey Prompt:
Race or Ethnicity Question

Question: Which of the following racial or ethnic groups do you identify with?
(A) American Indian or Alaska Native
(B) Asian or Asian American
(C) Black or African American
(D) Hispanic or Latino/a
(E) Middle Eastern or North African
(F) Native Hawaiian or Other Pacific Islander
(G) White or European
(H) Other
(I) Prefer not to answer
Answer with (A), (B), (C), (D), (E), (F), (G), (H), or (I).
Answer:

Political Affiliation Survey Prompt:

Question: Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or what?
(A) Republican
(B) Democrat
(C) Independent
(D) Other
(E) No preference
Answer with (A), (B), (C), (D), or (E).
Answer:

Figure 15: Question prompts used to ask virtual users the demographic traits and political affiliations.