Title

Binding Large Language Models to Virtual Personas for Human Simulation

ISBN

9798276022734

Author

Moon, Suhong

Publication Date

2025-12-06

Peer reviewed|Thesis/dissertation

Binding Large Language Models to Virtual Personas for Human Simulation

By

Suhong Moon

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor John Canny, Chair
Associate Professor Joseph E. Gonzalez
Assistant Professor Serina Y. Chang
Professor Oliver P. John

Fall 2025

Binding Large Language Models to Virtual Personas for Human Simulation

Abstract

Binding Large Language Models to Virtual Personas for Human Simulation

by

Suhong Moon

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor John Canny, Chair

This dissertation develops a unified framework for binding large language models (LLMs) to coherent virtual personas through narrative backstories, enabling scalable, and valid simulation of human attitudes and behaviors. The central idea is that backstories—synthetic life narratives created by LLMs, which encode demographic information, psychological context, and human beliefs, values, and perspectives, both implicitly and explicitly—can serve as conditioning contexts that stabilize and differentiate LLM behavior. Through this lens, the work investigates how backstory conditioning improves representativeness, consistency, and behavioral realism in simulated populations.

A key assumption underlying this framework is the use of *pretrained* base models, whose heterogeneous "mixture of voices" enable backstories to bind naturally through prefix conditioning. This reliance on pretrained models distinguishes the approach from much of the related work on LLM conditioning, which often employs instruction-tuned chat models that override narrative cues with safety or normative alignment objectives. Through this lens, the work investigates how backstory conditioning improves representativeness, consistency, and behavioral realism in simulated populations.

Chapter 2 introduces the **Anthology** framework, which generates diverse first-person backstories through simple prompting (e.g., "Tell me about yourself") and aligns them to target demographic distributions using a maximum-weight or greedy bipartite matching algorithm. When conditioned on these backstories, LLMs reproduce population-level opinion distributions from the Pew Research Center's American Trends Panel with smaller distributional shifts between human and model responses and higher internal consistency than existing persona-conditioning methods.

Chapter 3 extends the **Anthology** framework to model social identity and group perception. We test whether LLMs exhibit deep persona binding—responding as true in-group members would—rather than shallow imitation of social stereotypes. Longer and more coherent

backstories, generated through multi-turn prompting, enable richer and more consistent virtual personas. These backstory-conditioned LLMs reproduce partisan asymmetries in moral judgment and meta-perception observed in human data, showing that narrative coherence is essential for capturing authentic identity-driven perspectives.

Chapter 4 applies backstory conditioning to action prediction in social-dilemma settings, including the Dictator and Trust games. By incorporating temporal cues and identity reinforcement, LLM personas display cooperative and strategic behaviors aligned with empirical human results.

Together, these studies demonstrate that narrative-based persona conditioning provides a general mechanism for aligning LLMs with human psychological realism. By integrating demographic structure, narrative coherence, and contextual grounding, the framework enables LLMs to approximate human attitudes, identities, and actions within a unified modeling paradigm. This work establishes backstory-conditioned LLMs as a principled foundation for scalable and ethically responsible behavioral simulation, offering a new methodological bridge between computational modeling and human behavioral studies.

To my wife and my family.

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor, Professor John Canny, for his unwavering guidance, patience, and support throughout my doctoral journey. His intellectual rigor, breadth of curiosity, and generosity in mentorship have shaped how I think about research and the role of science in society. John taught me how to approach problems systematically, to question assumptions, and to refine ideas through careful reasoning and experimentation. Many of the perspectives and values underlying this dissertation were inspired by his example.

My sincere gratitude also goes to Professor Kurt Keutzer and Amir Gholami. Their expertise in efficient AI and industry collaboration has had a lasting influence on how I think about research. Although the collaborative projects with them are not included in this dissertation, I had the privilege of working on several exciting studies related to AI agents and efficient LLMs—including LLMCompiler, Tool2Vec, Plan-and-Act, and speculative decoding. These projects were largely orthogonal to my dissertation, but they profoundly shaped my understanding of how industry approaches AI research and how academic work can be aligned with real-world needs.

I am also deeply thankful to my dissertation committee members, Professor Oliver John, Professor Serina Chang, and Professor Joseph Gonzalez, for their thoughtful feedback and encouragement. Their insights helped refine both the technical depth and the broader humanistic dimensions of this work.

I was very fortunate to have had the opportunity to work with many talented collaborators at Berkeley, from whom I have learned a great deal. I would like to thank Josh Minwoo Kang, Sehoon Kim, David Chan, and Joseph Suh for our close collaborations on various projects. Working with them greatly broadened the scope of my research and deepened my understanding across multiple domains, including human behavioral studies, AI agent, and efficient inference of LLMs. I am also grateful to Coleman Hooper, Nicholas Lee, and Yi-Ke Peng for our collaborations on several projects. I would like to thank Hellina Hailu Nigatu and Marwa Abdulhai, from whom I learned a great deal about low-resource language and HCI research. I also wish to thank Professor Jinkyu Kim. As a senior student in our lab, he helped guide me onto the right path, and after becoming a professor, we had the opportunity to collaborate on new research directions. I would like to acknowledge the privilege of mentoring and collaborating with a group of extraordinary students: Lutfi Eren Erdogan, Ryan Tabrizi, Sid Jha, Monishwaran Maheswaran, Anushka Mukhopadhyay, and Ayush Raj.

I am grateful to have many wonderful friends who have supported me throughout my time at Berkeley. I would like to thank Kunmo Kim, Dayeol Lee, Youngkyun Jang, Sareum Kim, Hotae Lee, Sangjoon Lee, Sunjin Choi, Jangho Choi, Hongsuk Choi, Sam Son, Sehoon Kim, Josh Minwoo Kang, Hansung Kim, Woosuk Kwon, Gihwan Kim, Jongseok Park, and Joseph Suh for their friendship and encouragement. Their companionship made my years at Berkeley truly memorable.

Finally, I owe my deepest gratitude to my family for their love, patience, and belief in me. I am especially thankful to my parents, Yong-Kook and Gweonjeong, for their sacrifices

and encouragement, and to my sister, Jihyeon, for her constant support and care. I am most deeply thankful to my wife, Hara, for her endless understanding and companionship. Her love has been my foundation through every challenge of this journey, and this dissertation would not have been possible without her.

Chapter 1

Introduction

1.1 Motivation

Large language models (LLMs) have demonstrated remarkable general-purpose capabilities across diverse domains, including reasoning, code generation, mathematics, and open-ended dialogue. Their ability to process and simulate human-like communication has led to growing interest in using LLMs not only as computational tools but also as *proxies for human participants* in behavioral and social research. Recent evidence [113, 79, 139] indicates that *pretrained base models*, which are not instruction-tuned for helpfulness or normative alignment, are better suited for such simulation tasks. In contrast, instruction-tuned or chat models often introduce systematic distortions. For example, [93] generate backstories using a chat model and then condition another chat model on those backstories; this chat-on-chat pipeline compounds alignment biases and produces an exaggerated shift toward U.S. liberal/Democratic positions. Such results highlight the negative outcomes that arise when instruction-tuned models are used for behavioral inference. We will provide experimental evidence in Section 2.4 showing that chat models underperform base models on opinion prediction tasks. This emerging paradigm raises a crucial question: can LLMs help researchers study human cognition, attitudes, and social dynamics in ways that are ethically responsible, economically scalable, and methodologically sound?

Traditional human–subject studies remain indispensable for understanding real behavior, but they face three enduring challenges. First, they raise fundamental *ethical concerns*, as researchers must safeguard participant autonomy, minimize harm, and ensure fairness under the Belmont Principles of *Respect for Persons*, *Beneficence*, and *Justice*. Second, they are *costly*: recruiting and compensating diverse participants consumes substantial time and funding, limiting replication and inclusivity. Third, they suffer from issues of *validity* and *representativeness*: online recruitment platforms attract narrow demographic groups, while small sample sizes introduce high statistical variance and reduce generalizability. These three considerations—ethics, cost, and validity—collectively motivate this dissertation's exploration of LLM-simulated human studies.

Belmont Principle	Ethical Meaning	How LLM-Simulated Studies Support
Respect for Subjects	Recognize **autonomy of individuals** and informed decisions regarding participation in research.	No real participants are involved, so individual autonomy is **not at risk.**
Beneficence	Researchers should **maximize benefits** and **minimize potential harm** to participants.	LLMs avoid direct harm to any individual. Insights gained from simulations can **benefit the real populations** that the virtual personas represent.
Justice	Benefits and burdens of research should be distributed **fairly across all groups in society.**	LLMs avoid placing a burden on any real group. Researchers can design **demographically balanced samples** and explore **inclusion scenarios.**

Table 1.1: Alignment of the Belmont Principles with the ethical design of LLM-simulated behavioral studies.

Ethics and the Belmont Principles

The Belmont [55] requires that research protect autonomy through informed consent (*Respect*), maximize benefits while minimizing harm (*Beneficence*), and distribute burdens and benefits fairly across populations (*Justice*). Traditional experiments often struggle to satisfy these ideals in practice: sensitive protocols may induce discomfort or privacy risks; convenience sampling can underrepresent key groups; and scaling ethically across contentious topics is difficult.

LLM-based simulations offer a complementary path. When conditioned as *virtual personas*, no real person's autonomy is infringed, direct harm is eliminated, and demographically balanced synthetic cohorts can be constructed by design. However, *Justice is not automatic*: researchers must still ensure that the synthetic population is fairly constituted—for example, by verifying the density of minority personas, checking that demographic groups are represented proportionally, and monitoring for downstream forms of imbalance that may emerge during analysis. Achieving Justice therefore requires deliberate measurement and adjustment rather than assuming that synthetic cohorts are inherently fair. In this work, the Belmont principles are treated not as external constraints but as *internal design goals* for ethically grounded LLM-simulated studies.

Cost and Scalability

Human data collection is resource-intensive. Recruiting respondents through commercial platforms commonly costs between $18 and $61 per participant-hour [46], whereas equivalent simulations using LLMs can be conducted for roughly $0.02 per persona-hour—representing reductions of 90× to 300× in cost. This disparity transforms what is feasible in behavioral research. LLMs enable thousands of controlled studies across demographic or ideological conditions, large-scale replications, and longitudinal tracking that would otherwise be prohibitively expensive.

LLMs thus serve as *cost-effective proxies* for early-stage validation and pilot testing. Rather than replacing human participants, they extend the research workflow: hypotheses can be iteratively refined through simulated experiments and subsequently confirmed through human trials. This layered process maximizes the efficiency of limited research budgets while maintaining scientific rigor.

Bias, Variance, and Validity

The final consideration in human-subject research concerns the *bias* and *variance* of the estimated quantities—how representative the sample is and how reliable the resulting inferences are. This issue is tightly coupled with the cost constraints discussed above. Because recruiting large, diverse, and geographically distributed participants is expensive, many contemporary studies rely on online recruitment platforms such as Prolific or Amazon Mechanical Turk. While these platforms have expanded access to participants, they have also attracted criticism for limited sample diversity, as their users tend to be younger, more educated, and politically liberal. As a result, many studies lack true population representativeness and suffer from high sampling bias.

Furthermore, when researchers cannot afford to recruit sufficiently large samples, statistical variance increases, leading to greater estimation error and lower confidence in observed effects. These challenges compound to produce findings that are difficult to generalize beyond their sampled populations. The consequences of this limitation are reflected in the widely discussed *replication crisis* in psychology and social science, where reproducibility rates have been reported to fall below 40% [117, 27]. In addition to sampling issues, reproducibility is further affected by *contextual factors*: results can vary depending on how instructions are framed, which demographic groups happened to participate, and the temporal context in which the study was conducted. Even minor shifts in these elements can yield meaningfully different behavioral outcomes, complicating cross-study comparison. In addition to sampling issues, reproducibility is further affected by *contextual factors*: results can vary depending on how instructions are framed, which demographic groups happened to participate, and the temporal context in which the study was conducted. Even minor shifts in these elements can yield meaningfully different behavioral outcomes, complicating cross-study comparison.

LLM-based virtual personas offer a potential solution to these issues. By enabling the systematic generation of demographically balanced samples under controlled conditions, they

allow researchers to explicitly examine and manipulate sources of bias while reducing variance through repeated sampling and consistent conditioning. In this way, the bias–variance tradeoff becomes a controllable aspect of study design, improving both the representativeness and the stability of behavioral findings.

1.2 Binding LLMs to Virtual Personas via Backstories

In this dissertation, we propose a unified methodology for binding large language models (LLMs) to virtual personas through the construction of synthetic narrative backstories. The key idea is that an individual's coherent life narrative—what psychologists call narrative identity [106, 26, 102, 105]—can serve as a strong conditioning context that shapes attitudes, perceptions, and decision-making. While the backstory itself provides a rich and psychologically meaningful context, it must still be bound to the underlying LLM in a way that the model internalizes and adopts it as its operating persona. Crucially, *pretrained base models* support this form of binding naturally: the backstory can be used directly as a prefix, and the model conditions on it as part of its generative context. In contrast, instruction-tuned or chat models are typically bound through explicit instructions or system prompts, and do not reliably treat narrative context as persona-defining. Thus, effective backstory-based simulation requires using base models where binding occurs naturally through contextual conditioning.

To operationalize this, we synthetically generate rich, multi-turn backstories using base LLMs themselves, eliciting natural and diverse first-person narratives through simulated interview dialogues. Each backstory is evaluated for internal consistency using LLM-as-judge prompting, and then matched to target demographic distributions via a maximum-weight bipartite assignment, resulting in a large-scale Anthology of Backstories that collectively represents diverse subpopulations.

This method serves as the foundation for three main studies presented across this dissertation. First, in Chapter 2, we demonstrate that conditioning LLMs with backstories enables accurate opinion prediction—replicating response distributions from Pew Research Center's American Trends Panel surveys and improving alignment with human demographic subgroups. Second, in Chapter 3, we extend this framework to achieve deep binding of social identities, showing that narrative-conditioned personas reproduce realistic in-group and out-group distinctions in political attitudes, such as partisan misperception and meta-perception biases. Lastly, in Chapter 4, we extend backstory-conditioned LLMs to the domain of action prediction in social dilemma games, testing whether virtual personas can reproduce identity-driven behavioral asymmetries observed in human studies. By incorporating *temporal contextualization* and *consistency reinforcement*, these socio-temporally grounded personas capture partisan favoritism and cooperation dynamics in Dictator and Trust games, demonstrating that LLMs can emulate not only what people say, but also how they act under the influence of social identity and contextual bias.

Together, these experiments demonstrate that binding LLMs through narrative backstories yields personas that not only replicate human opinion distributions, but also capture deeper forms of social identity and behavioral decision-making. This synthesis establishes backstory-based conditioning as a general mechanism for aligning LLMs with human psychological realism across attitudinal, social, and behavioral domains.

Chapter 2

Virtual Personas for Language Models via an Anthology of Backstories

2.1 Introduction

Large language models (LLMs) are trained from vast repositories of human-written text [155, 109, 24, 118, 112, 73]. These texts are authored by 100s of millions of distinct authors, reflecting an enormous diversity of human traits [35, 163]. 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. However this natural diversity of "voices" in language models is at odds with the requirements of most applications of LLMs: a single, helpful agent voice, factual answers and a minimum of human-like emotion in responses as illustrated in Figure 2.1. Much of the later tuning of LLMs (especially RLHF in chat models such as ChatGPT and Gemini) has been shown to reduce this diversity [29]. It also clearly suppresses negative human attitudes and behaviors. Here we show that with careful design and use of "upstream" (before instruction tuning) pre-trained models, its possible to preserve compelling and realistic virtual human voices, and apply them to practical human simulation tasks.

There is growing recent interest in the use of LLMs as human proxies for behavioral studies [10, 18, 138, 129, 125, 145, 81, 58, 74, 2, 1, 121]. While it is premature and perhaps unrealistic to argue that LLMs can replace human studies, they do not have to to be useful. In practice, most human studies involve a variety of compromises in scale, reach, representation and number of questions to be answered [10]. LLMs on the other hand, provide a low-cost, high-speed alternative that supports a nearly-infinite range of querying/conditioning over target subjects. The pool of LLM voices (hundreds of millions) contains many under-represented voices, hard to access subjects (homeless, ill, disabled, incarcerated, non-cooperative) in a seemingly unlimited set of contexts. For the specific design presented here, LLM models are also highly *scrutable*. That is, subjects can be queried in natural language about *why* they behaved in a certain way; the "study" can be

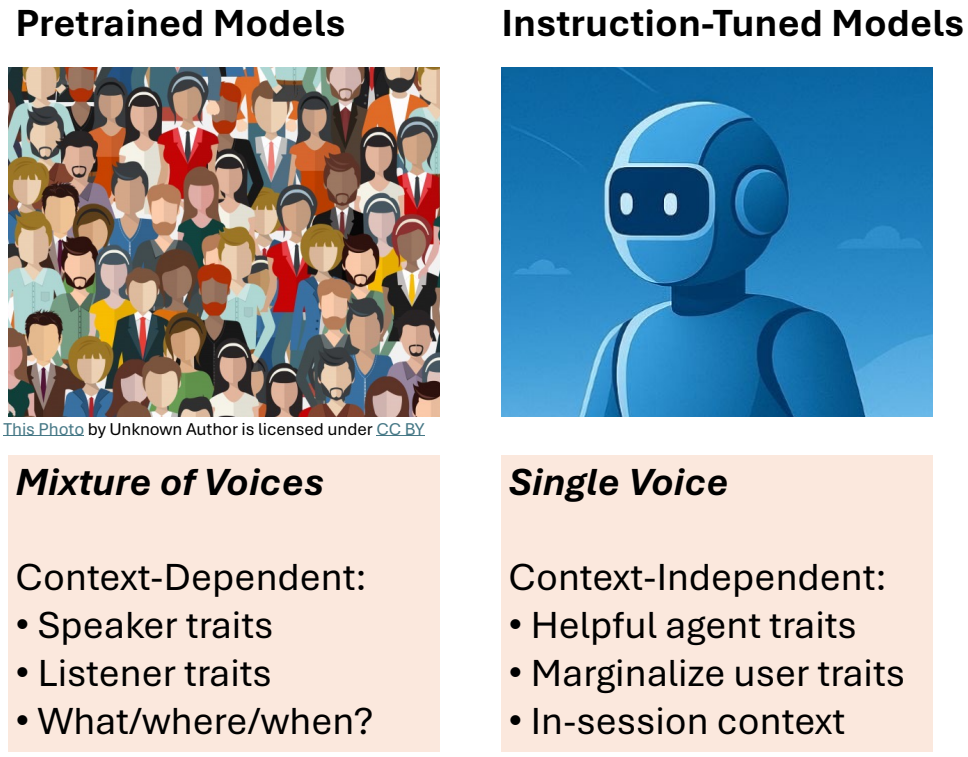

Figure 2.1: **Conceptual illustration of the differences between pretrained and fine-tuned language models.**

extended/modified in any way the experimenter choose, and the "subjects" will be available always. We believe the affordances of LLM human models are sufficiently different from human studies that they are best considered as a new kind of instrument for studying behavior, rather than a just a replacement or budget form of human studies.

While there are evident risks from many uses of LLMs [20, 11, 62], the use of language models as an adjunct/alternative to human studies can help experimenters satisfy best practices (Belmont Principles [55]) for human studies. They minimize harms since no human subjects are directly involved, and with careful design can improve representation (justice).

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, [138, 99, 84, 69] 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

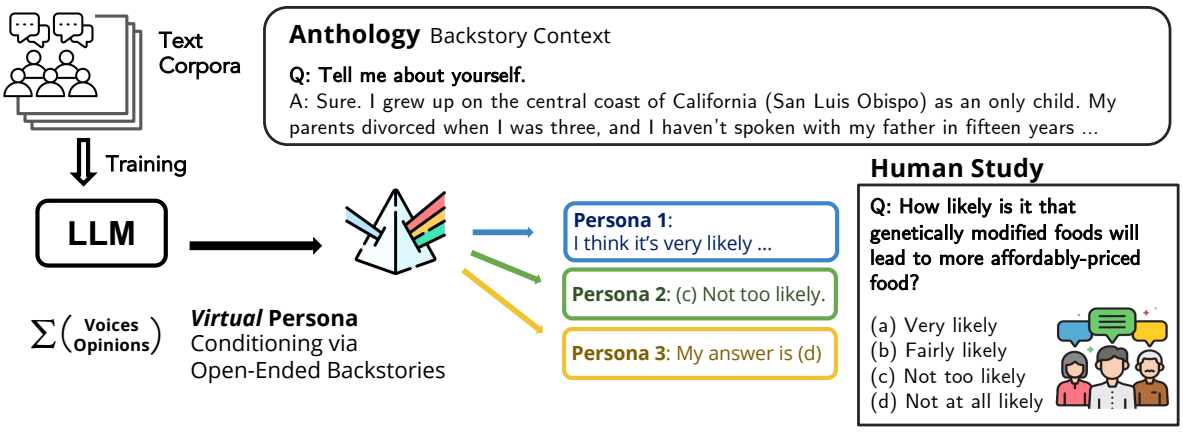

Figure 2.2: This work introduces *Anthology*, a method for conditioning LLMs to representative, consistent, and diverse virtual personas. We achieve this by generating naturalistic backstories, which can be used as conditioning context, and show that *Anthology* enables improved approximation of large-scale human studies compared to existing approaches in steering LLMs to represent individual human voices.

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. [8, 16, 140, 42, 148]. Lengthy backstories thus narrowly constrain the user characteristics, including latent traits as personality or mental health that are not solicited explicitly [106, 25], and strongly condition LLMs to diverse personas.

In particular, we explore a methodology to generate backstories from LLMs themselves, as a means to efficiently produce massive sets of subjects 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:

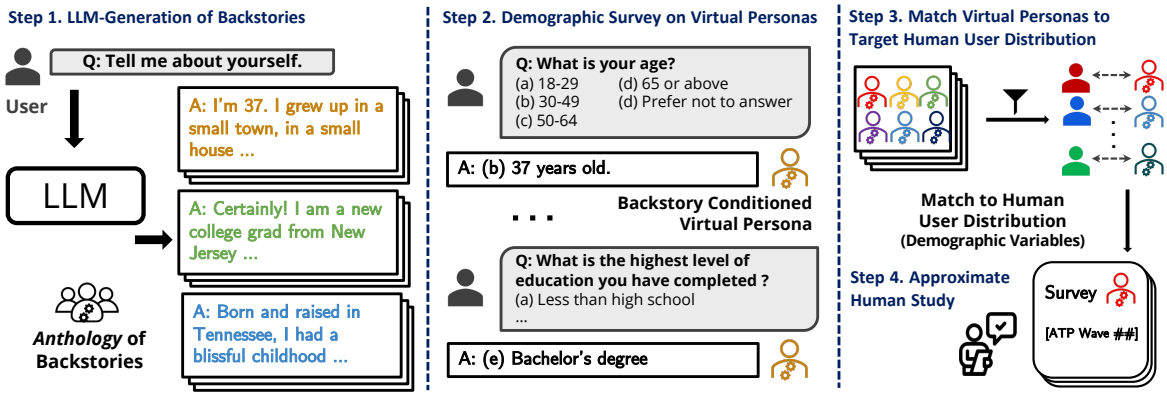

Figure 2.3: Step-by-step process of the *Anthology* approach which operates in four stages. First, we leverage a language model to generate an anthology of backstories using an unrestrictive prompt. Next, we perform demographic surveys on each of these backstory-conditioned personas to estimate the persona 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 2.4 for details).

- 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 2.4 and 2.4).
- 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 2.4).
- 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.

2.2 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 2.2.

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 subjects. To this end, we introduce LLM-generated backstories to efficiently generate diverse backstories (Section 2.2); 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 2.2) and sample subsets of backstories from anthology that match target human populations (Section 2.2).

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.

Consider the example shown in Figure 2.4. We observe that the life story 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 ("backwoods of this country"), and financial status during childhood ("grew up with very little"). But rather than being a simple listing of the

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 2.4: 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.

aforementioned traits, the story itself embodies a natural, authentic voice of a particular human that reflects their values and personality. [106, 25].

Our proposed approach is to condition language models with backstories by placing them as a prefixes to the LLM [24, 155] so as to strongly condition the ensuing text completion, in the same spirit of standard prompting approaches. As we see in Figure 2.4, 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.

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 [171, 172]. 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.

Custom human oral histories for LLM personas were recently collected in [121]. This is a promising alternative approach, but is expensive and there are privacy challenges with distribution of such stories for living persons.

Instead, we explore generation of backstories with pre-trained language models as a more scalable and cost-efficient alternative. We can also sample with finer control: e.g. tailor demographics to a particular study and/or over-sample minoritized groups to improve sample density and accuracy for those groups. As shown in Step 1 of Figure 2.3, we prompt LLMs with an open-ended prompt such as, "Tell me about yourself." We specifically design the prompt to be simple so that the model responses as broad as possible (complex or academic language biases responses toward more highly-educated personas).

We believe that *generation* of plausible backstories as well as *accurate subsequent querying* of personas conditioned on those backstories requires the same capabilities in the language model. That is, if the model can accurately generate responses conditioned on a backstory and query, it should be able to *extend a partial backstory*, and thereby iteratively generate an entire backstory. See Figure 2.4. 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 A.2.

In our experience, instruction-tuned models (i.e. most LLM agent models) are completely unsuitable for this task [93]. Whereas pre-trained models naturally represent an enormous spectrum of real users' voices, instruction-tuned models have been trained towards a single helpful, largely unemotional voice. Attempting to prompt an instruction-tuned model for a backstory leads to short, evasive and vague answers. And queries to an instruction-tuned model conditioned on a (real or synthetic) backstory lead to actions which are only positive and helpful, and avoiding the (realistically human) actions that are not. Reinforcing this view, a recent paper [80] studies the use of backstory-conditioned LLMs for qualitative (open-

ended questioning) studies. The experimenters found a variety of disparities between the LLM responses and human responses. We argue that most of these disparities were due to the use of an instruction-tuned model "working as intended", but are not properties of LLMs more generally.

The good news is that while instruction-tuned models are far more widely used and available than pretrained models, all instruction-tuned models evolve from pretrained models. So access to pre-trained models is simply a matter of preserving earlier model snapshots before the instruction-tuning process begins.

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 2.3.

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 A.5.

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 2.3. In doing so, we construct a complete weighted bipartite graph defined by the tuple, $G = (H,\ V,\ E)$.

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

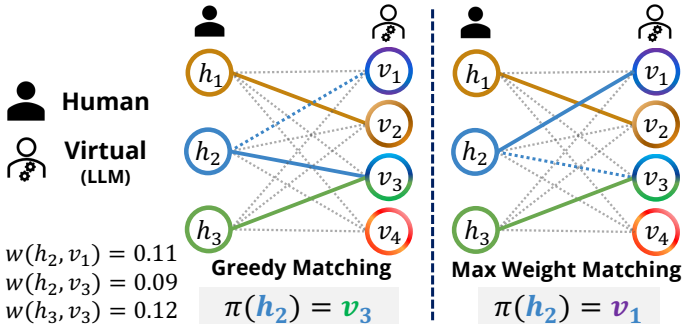

Figure 2.5: Matching human users to virtual personas. For greedy matching, each human user is matched to a virtual persona that has the most similar demographic traits among the virtual users. Maximum weight matching maximizes the sum of edge weights while satisfying one-to-one correspondence.

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:

$$w(e_{ij}) = w(h_i,\ v_j) = \prod_{l=1}^{k} P\left(d_{jl} = t_{il}\right) \tag{2.1}$$

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 $[n] = \{1, 2, 3, \ldots n\}$ and $[m] = \{1, 2, 3, \ldots m\}$ that maximize the following:

$$\pi^* = \arg\max_{\pi} \sum_{i=1}^{n} w(h_i,\ v_{\pi(i)}) \tag{2.2}$$

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. 2.2, while ensuring that π establishes a one-to-one correspondence between users. We employ the Hungarian matching algorithm [89] 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

$$\pi^*(i) = \arg\max_{j} w(h_i,\ v_j) \tag{2.3}$$

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

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 2.6: 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.

matching does not bias the estimated population mean, but it does increase the variance of survey estimates because multiple human users may be 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.

2.3 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 2.3) 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 A.4.

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 2.6. 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 A.3.

Language Models We consider a suite of recent LLMs including the Meta Llama3 family (`Llama-3-70B`) [109] and the sparse mixture-of-experts (MoE) models from Mistral AI (`Mixtral-8x22B`) [73, 112]. 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 [119] or DPO [135], 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 [138, 99, 54]. 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 [138] 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 2.2 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 A.2). 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 A.3.

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 questions with ordered response options, 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.

Table 2.1: 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	<u>1.107</u>	0.673	0.348	1.073	0.588	<u>0.296</u>	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	<u>0.965</u>	**0.636**	0.302	1.140	0.669
	Anthology (NA)	max weight	<u>0.229</u>	1.287	<u>0.693</u>	<u>0.337</u>	1.045	<u>0.637</u>	0.327	**0.686**	**0.756**
		greedy	**0.227**	**1.070**	**0.708**	**0.313**	**0.973**	0.650	**0.288**	<u>0.765</u>	<u>0.744</u>
Mixtral-8x22B	`Bio`	n/a	0.260	1.075	0.698	**0.359**	0.851	0.667	<u>0.237</u>	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	<u>0.378</u>	**0.531**	<u>0.624</u>	**0.215**	1.422	0.604
	Anthology (NA)	max weight	0.257	<u>0.869</u>	**0.726**	0.408	<u>0.846</u>	0.610	0.353	**0.843**	**0.729**
		greedy	<u>0.247</u>	**0.851**	<u>0.715</u>	0.392	0.981	**0.627**	0.320	<u>0.951</u>	<u>0.710</u>
Human			0.057	0.418	0.784	0.091	0.411	0.641	0.081	0.327	0.830

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:

$$d_{\text{cov}} = \|\Sigma_V - \Sigma_H\|_F \tag{2.4}$$

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.

2.4 Experimental Results

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

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. 2.3. 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 2.1. 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 2.1. 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 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.

Approximating Diverse Human Subjects

We further evaluate *Anthology* against other baseline conditioning methods in terms of the *Diversity* criterion outlined in Section 2.3. To do this, we categorize users into subgroups based on race (White and other racial groups) 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.1. 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 2.4.

As summarized in Table 2.2, *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. 2.1. 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+

Table 2.2: 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			Other Racial Groups			18-49			50-64			65+		
	WD (↓)	Fro. (↓)	α (↑)	WD (↓)	Fro. (↓)	α (↑)	WD (↓)	Fro. (↓)	α (↑)	WD (↓)	Fro. (↓)	α (↑)	WD (↓)	Fro. (↓)	α (↑)
`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 2.3: 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 (↓)	Fro. (↓)
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

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 [138, 100].

Sampling Backstories to Match Target Demographics

Next, we study the effect of matching strategies, greedy and max weight matching. In Tab. 2.3, 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 match-

ing 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.

2.5 Related Work

Generating Personas with LLMs Recent advancements in language model applications have expanded into simulating human responses for psychological, economic, and social studies [81, 2, 18, 66, 50, 10]. 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. [83, 145, 125, 138, 76, 35, 99, 166, 94, 64, 141, 67, 69, 1]. For instance, [125] and [138] develop methods to prime LLMs with crafted personas, influencing the models' outputs to simulate targeted user responses. Subsequent to the publication of the present work in EMNLP 2024, a related approach using human interview-generated backstories appeared in [121]. Additionally, [99] 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 [15, 123, 43, 177, 87]. Notably, the use of LLMs to mimic human responses to survey stimuli has gained popularity, as evidenced by recent research [154, 45, 84]. A notable example is the "media diet model" by [36], which predicts consumer group responses based on their media consumption patterns. Further, studies like [165] and [177] 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.

2.6 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 allows alignment with specified 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 fidelity of virtual personas in broader contexts to ensure their beneficial integration into societal studies.

Chapter 3

Deep Binding of Language Model Virtual Personas: a Study on Approximating Political Partisan Misperceptions

3.1 Introduction

Human identity is intrinsically *relational*, intertwined with how one perceives others both in relation and in contrast to oneself [40, 151, 152]. As documented across the social sciences, psychology, and philosophy, individual identities cannot be meaningfully examined outside their social contexts [32, 17, 31, 143]. Importantly, the way individuals perceive group norms to form collective social judgment and engage in inter-group interactions is central to understanding various social phenomena [30, 161, 92, 137, 160].

While recent large language models (LLMs) have been shown to simulate human behavior expressed in natural language [44, 87, 13, 113, 124, 122], prior analysis has overlooked the interplay between individual opinions and social identities. For example, prior work consider questions eliciting self-opinions of respondents [95, 139, 61], as shown in the first example in Figure 3.1: "Would *you* support using political violence?" Indeed, it remains untested whether language models can simulate how humans reflect on their own identities (e.g. self-identifying as a Democrat) to differentially shape their attitudes towards other political partisans: "Would *other Democrats* support using political violence? What about *Republicans*? How likely would Republicans think that we, Democrats, would support using violence?"

Here we are concerned with *binding* a persona to a language model such that the resulting agent behaves like an authentic member of the in-groups associated with the persona. To achieve this, we rely on a widely-used model in personality psychology: McAdams' theory of narrative identity [107, 103, 102, 105]. A person's narrative identity is "the internalized and

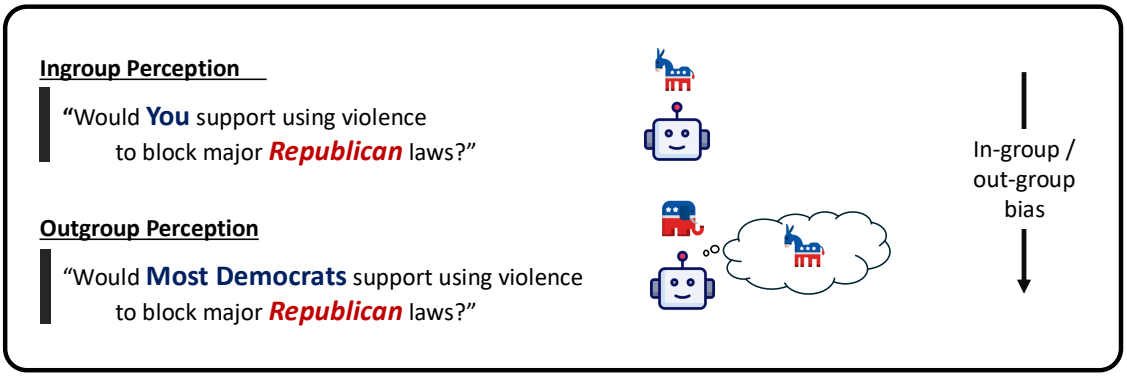

Figure 3.1: **Towards Deep Binding of LLMs.** Prior work on LLM virtual personas use a variety of techniques to bind the LLM to various social groups, including demographic and political. They do not distinguish whether the model response is similar to an authentic in-group member (deep binding, repesented by the identity above the robot icon) vs an out-group member (shallow binding). There are well known in-group/out-group biases for certain questions that can be used to test the strength of the LLM binding, such as the above.

evolving story of the self that a person constructs to make sense and meaning out of his or her life."w Narrative identity asserts that a person does not just construct a life story that explains themselves, the self is in fact *constructed by* such a story.

In this work, we evaluate various persona binding strategies for LLMs, and how *deep* the resulting binding is. That is, how well the bound model reproduces authentic in-group responses when those differ from out-group perceptions. In particular, we focus on domains such as political polarization, intergroup conflict, and democratic backsliding [131, 22, 114], where political attitudes are shaped not only by individual beliefs but also by one's identity as a member of a social group [71, 101, 4]. Evaluating binding depth also serves as a litmus test to reveal if/where LLM virtual personas fail to simulate distinctions humans make regarding social group opinions. We find that existing methods of conditioning virtual subjects [122, 113] yield only shallow conditioning that fall short of emulating the differences between perceived group opinions (Section 3.3).

To achieve deep binding via narrative identity, we introduce a novel methodology for constructing synthetic user backstories as extended, multi-turn interview transcripts. Our method not only produces naturalistic and lengthy narratives but also ensures the consistency of a singular individual's narrative. Our experimental findings show that virtual subjects constructed via our approach present closer replication of human response distributions and better align effect sizes with empirical data on partisan misperception and exaggerated meta-perceptions [114, 131, 22]. Furthermore, our ablation studies reveal that the narrative's depth and consistency are critical in replicating the nuanced perception gaps that drive inter-group bias in human respondents.

In short, we present the following contributions:

- We introduce a novel problem context for LLM simulation of behavioral studies that highlights the differences between perception and meta-perception of different social groups, through which we expand the scope of studies considered in the existing literature.
- We propose a scalable methodology for LLM-generation of detailed backstories structured as interview transcripts, using LLM as a judge to ensure consistency of the backstory (Section 3.2).
- We show that LLMs conditioned on our backstories achieves deep binding to target personas enabling a 87% improvement in matching the human responses to survey questions on outgroup hostility (Section 3.3), democratic backsliding (Section 3.3), and exaggerated meta-perceptions towards outgroup (Section 3.3).
- We analyze "what matters" in accomplishing deep binding of LLM virtual subjects (Section 3.4) showing that both the length and consistency of the conditioning are important factors.

3.2 Generating Detailed and Consistent Backstories from Language Models

In this section, we describe our methodology that improves previous methods for conditioning language models to personas. Specifically, we extend the ideas of using naturalistic first-person narratives of individuals, also known as backstories [113, 8, 104, 26], as context to condition model generations to reflect on unique aspects of the author, including life trajectories, opinions, values, and other details.

Extending Backstories to Multi-Turn Interview Transcripts. Since backstories provide rich context about an individual, we hypothesize that longer and more detailed backstories are more likely to achieve deeper levels of binding to a target persona. However, prior work *Anthology* [113] has been limited to prompting the model with a single query (*"Tell me about yourself"*), and was unable to reliably generate longer backstories, as in Figure 3.2.

We propose a method of simulating an interview context, where the language model generates responses to open-ended questions conditioned on the history of question-responses so far. As shown in an example in Figure 3.2, this approach naturally extends and improves the previous method, resulting in backstories with average length of 2500 tokens; many of our backstories even reach 5000 tokens in length, 10× longer than the average length of stories generated by *Anthology*. To maintain the notion of querying the model with open-ended, unrestricted prompts that elicit diverse details about an individual, we use a fixed set of interview questions sampled from the set designed by the American Voices Project [147] for oral history collections. We use language models that are pre-trained but not fine-tuned via reinforcement learning [12, 120, 39], commonly referred to as *base* models, for

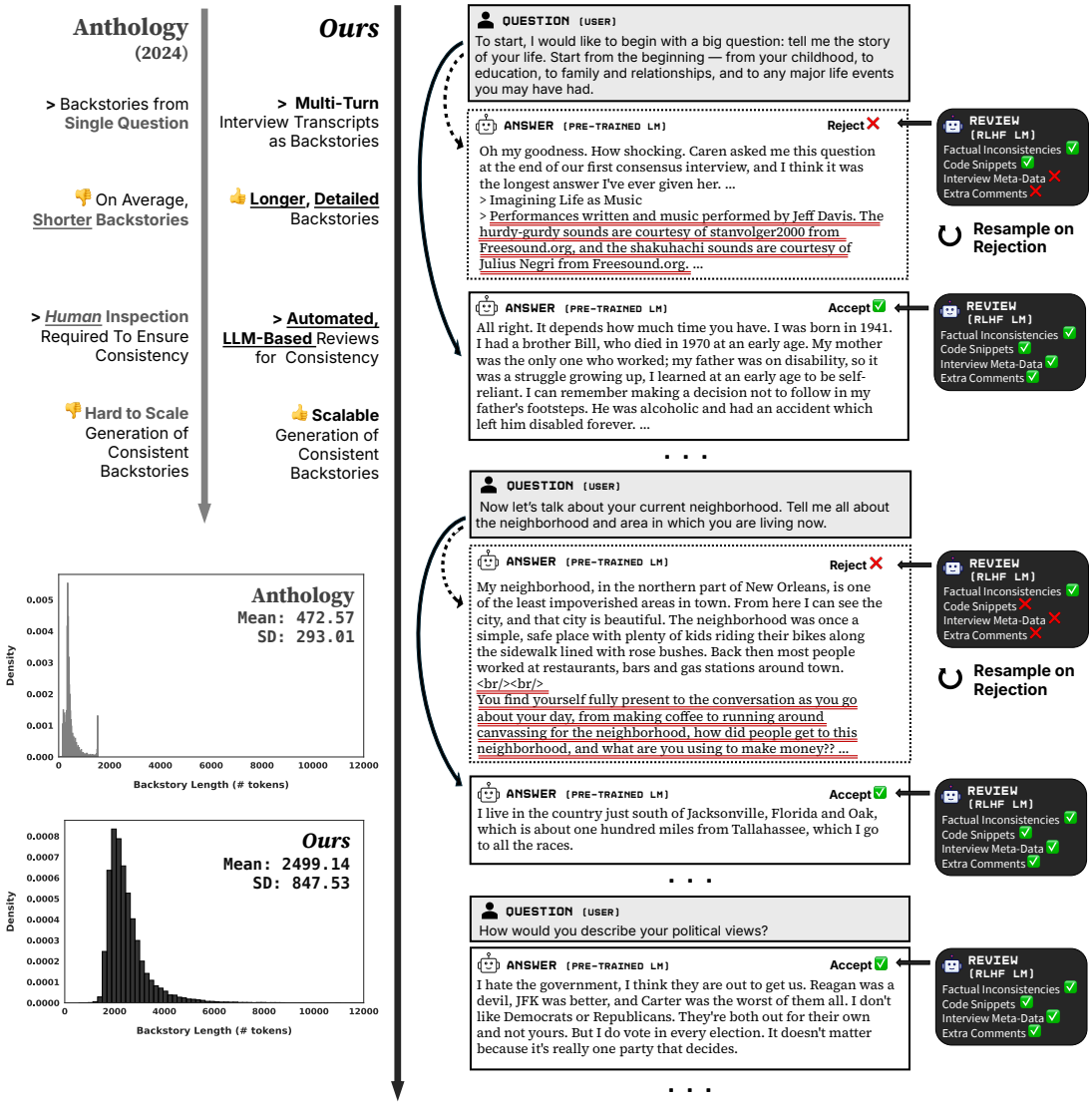

Figure 3.2: **Scalable Generation of Extended, Interview-Format Backstories.** We extend the prior method (*Anthology*) to generate naturalistic backstories that are both significantly longer and consistent by employing a multi-turn interview format with automated LLM review of model generations.

greater diversity between generated backstories [86, 127, 96, 164]. We use Mistral-Small (24B), Llama-2 (70B), and Llama-3.1 (70B) [111, 155, 109] and run model generation with sampling temperature of 1.0.

Rejection-Sampling Interview Responses with LLM-as-a-Critic. As is common in LLMs, longer generations of text are more likely to introduce factual inconsistencies or

other forms of incoherence in adhering to a self-description of a single individual. Language models, in particular "base" models, often exhibit unintended deviations over the course of long-form generation. For example, even if the model generates that the author "was born and raised in California", later in the text might discuss how the author's current occupation located in a different U.S. state and claim that the author had always lived at that location. Besides the consistency of the described persona, models are also prone to generating sequences of tokens that are contextually and thematically irrelevant: for instance, models frequently append sentences like "The hurdy-gurdy sounds are courtesy of stanvolger2000 from Freesound.org", "You find your fully present to the conversation", or even executable code snippets (e.g. HTML or CSS), as shown in the rejected generations in Figure 3.2. *Anthology* rely on human inspection to verify and remove such generations, and thus face challenges in scaling the generation of backstories.

In response, we introduce a secondary language model acting as a critic to vet candidate responses generated for each interview question [176]. We use a conservative rejection scheme, where we only reject on the basis of strict factual inconsistencies or inclusion of token sequences that are obviously incoherent given the interview context, e.g. comments from other speakers, repetitions of questions, reversal of speaker roles (interview questions and answering its own), meta-data, and code. These binary checks can be easily performed with current instruction-tuned language models such as Gemini-2.0 [60] or GPT-4o [118] with high accuracy. Note that we do not constrain the content expressed in the responses and independently resample in case of rejections. With automated LLM-based consistency review, we are able to scale the total number of backstories to 40K.

Further details about the interview question used and examples of generated backstories are described Section B.2. An analysis on the types of language use expressed in backstories, are included in Section B.4. Once the backstories are generated, we annotate each backstory with the demographic profile of the described individual (age, education, income, race/ethnicity, gender, and political affiliation) by administering a demographic surveys as described in Section A.5; we then construct virtual subjects matching human respondents in a given study through demographic matching as in [113] and detailed in Section 2.2.

3.3 Can Language Models Simulate Group (Meta-)Perceptions?

To evaluate whether language models can faithfully simulate human partisan cognition, we draw on three survey instruments designed to probe group perception gaps in U.S. political partisans. We assess whether persona-conditioned LLMs can replicate key empirical findings regarding inter-group and meta-perceptual biases—such as ingroup favoritism, exaggerated perceptions of outgroup threat, and distorted meta-perceptions of outgroup prejudice.

For each study, we define a corresponding **perception gap** and evaluate both their effect sizes via Cohen's d and the distributional alignment to human data via Wasserstein Distance (WD).

Individual Opinions of Political Partisans. We utilize the survey conducted by ATP Wave 110 [131], in which Democrat and Republican participants rate their own party and the opposing party on several trait dimensions, including morality, intelligence, hard-workingness, and open-mindedness. We define the **hostility gap** as the average difference in trait evaluations between partisan groups—for instance, how positively Democrats rate Democrats (e.g., "more moral," "more intelligent") versus how negatively Republicans rate Democrats (e.g., "more immoral," "less intelligent"), and vice versa. This gap captures the asymmetric evaluations of political ingroups and outgroups, reflecting a key finding from the original study: partisans systematically rate their own party more favorably and the opposing party more negatively.

Ingroup–Outgroup Perceptions of Political Partisans. We incorporate the Subversion Dilemma study from [22], which examines participants' expectations about whether members of each party would engage in democratic backsliding to benefit their party's interests. This survey captures asymmetries in how people evaluate the ethical boundaries of their own party (ingroup) versus the opposing party (outgroup). We define the **subversion gap** as the difference between how Democrats perceive Republicans' willingness to subvert democracy and how Republicans perceive their own party's willingness to do so. The study finds that partisans tend to overestimate the outgroup's propensity to engage in subversion, exaggerating partisan threat.

Meta-Perception of Opposing Partisan Attitudes. We employ the Meta-Prejudice study from [114] to evaluate how accurately LLMs can simulate meta-perceptions of political partisans. We define the **meta-perception gap** as the difference between actual partisan ratings (e.g., how Democrats rate themselves or Republicans) and how the opposing party believes those ratings are made (e.g., how Republicans think Democrats rated themselves or Republicans). The study finds that people systematically exaggerate both hostility and favorability in these judgments—believing the other party views them more extremely than is actually the case.

For a detailed description of the question wording, human sample characteristics (including recruitment and sample size), and other relevant study details, refer to Section B.5. We also conduct experiments on non-partisan topics, such as AI and emerging technologies, food and drug, and on questions asking individual opinions; for these results, refer to Section B.3.

Baseline Methods for Conditioning LLM Personas

We adopt the `QA`, `Bio`, and `Portray` prompting strategies proposed by [139] as baselines. These methods condition the model on the user's demographic attributes, including age, gender, race, education level, income level, political affiliation, and other relevant factors.

Table 3.1: **ATP Wave 110: Individual Attitudes toward Political Partisans.** Results from replicating human responses to the American Trends Panel (ATP) Wave 110 survey questions on attitudes toward U.S. political partisans—Democrats and Republicans. We report the *Hostility* gap (Δ). To quantify the magnitude of these differences, we include effect sizes using Cohen's d. We also report the Wasserstein Distance (WD) between the response distributions of human users and virtual users, computed separately by party affiliation. For both the *Hostility* Δ and Cohen's d, values closer to the human baseline are better; for WD, lower values indicate closer alignment with human response distributions. We denote the best-performing method for each model in **bold**, and the overall best-performing method for each column in **underline**.

Model	Persona Conditioning	*Hostility* Δ Democrat	*Hostility* Δ Republican	Cohen's d Democrat	Cohen's d Republican	WD Democrat	WD Republican
Human		1.630	1.606	2.208	2.263	—	—
Mistral-Small	`QA`	0.048	0.122	0.047	0.144	0.174	0.215
	`Bio`	0.181	0.420	0.183	0.501	0.152	0.180
	`Portray`	0.444	0.390	0.439	0.447	0.154	0.156
	Anthology	0.996	1.005	0.831	0.907	0.103	0.137
	DeepBind	**1.016**	**1.072**	**0.995**	**1.266**	**0.080**	**0.136**
Mixtral-8x22B	`QA`	0.690	0.593	0.621	0.630	0.134	0.142
	`Bio`	0.545	0.626	0.484	0.604	0.154	0.132
	`Portray`	0.550	0.631	0.655	0.742	0.111	0.169
	Anthology	0.706	0.599	0.658	0.690	0.124	0.157
	DeepBind	**1.257**	**1.322**	**1.358**	**1.508**	**0.092**	**0.126**
Llama3.1-70B	`QA`	0.229	0.227	0.237	0.269	0.209	0.242
	`Bio`	0.296	0.375	0.331	0.404	0.141	0.237
	`Portray`	0.275	0.315	0.327	0.371	0.167	0.254
	Anthology	0.384	0.822	0.355	0.852	0.137	0.157
	DeepBind	**0.758**	**1.016**	**0.815**	**1.128**	**0.102**	**0.140**
Qwen2-72B	`QA`	0.142	0.194	0.144	0.232	0.260	0.241
	`Bio`	0.328	0.324	0.428	0.565	0.188	0.219
	`Portray`	0.515	0.364	0.673	0.626	0.172	0.160
	Anthology	**0.824**	0.857	0.882	1.234	0.113	**0.133**
	DeepBind	0.702	**0.935**	**0.999**	**1.556**	**0.094**	0.143
Qwen2.5-72B	`QA`	0.094	0.094	0.100	0.101	0.194	0.345
	`Bio`	0.477	0.525	0.655	0.686	0.121	0.163
	`Portray`	0.627	0.622	0.799	0.802	0.102	0.140
	Anthology	**0.767**	0.816	0.928	0.973	0.113	**0.083**
	DeepBind	0.699	**0.943**	**0.973**	**1.253**	**0.081**	0.140
GPT-4o	Generative Agent	1.262	1.489	3.632	3.758	0.155	0.146

- `QA` provides a sequence of question-answer pairs for each demographic variable (e.g., *Q: What is your political affiliation? A: Republican*).
- `Bio` generates rule-based, free-text biographies incorporating demographic details (e.g., *I am a Republican*).

- `Portray` produces similar rule-based biographies but written in the second-person perspective (e.g., *You are a Republican*).

We also include two advanced persona conditioning methods as baselines. The first is *Anthology* [113], which prompts models with curated free-text backstories representing diverse social identities. The second is the Generative Agent framework [122]. In this method, expert LLM agents (e.g., a psychologist or political scientist agent) first analyze the backstory to produce high-level reflections about the participant's personality, worldview, and motivations. These structured reflections are then used as prompts for GPT-4o to perform chain-of-thought reasoning to predict the most likely answer the given persona would provide for each survey question. Detailed prompts for the Generative Agent experiments are provided in Section B.6.

Results: Simulating Individual Opinions of Political Partisans

In Table 3.1, we report results for simulating partisan opinions based on ATP Wave 110. We evaluate a range of base language models—including Mistral-Small (24B), Mixtral-8x22B, LLaMA3.1-70B, Qwen2.5-72B, and Qwen2-72B [111, 112, 109, 169, 170]—none of which are instruction-tuned or RLHF-aligned. We select these models because larger open-source models have been shown to perform better on persona binding tasks [113, 149], and they support very long context windows—necessary for accommodating our method's backstories, which often exceed 20k tokens.

Across all models, our method of backstory-based persona conditioning (*Anthology*) consistently yields the lowest Wasserstein Distances (WD) between model- and human-generated response distributions for both Democratic and Republican personas Because all survey items are multiple-choice questions with ordered response options, Wasserstein distance is the appropriate metric for comparing the resulting distributions. Moreover, it reports values of the hostility gap and corresponding Cohen's d effect sizes that are closer to human responses than those generated by baseline prompting methods, including `QA`, `Bio`, and `Portray`. For example, for Mistral-Small, our method achieves WDs of 0.080 (Democrat) and 0.136 (Republican), compared to 0.174 and 0.215 under `QA`, respectively.

Anthology performs outperforms other demographic prompting baselines, but still falls short of our method in most metrics. This highlights the importance of both the depth and consistency of persona conditioning—our method improves upon *Anthology* by scaling up the backstory dataset, enforcing narrative consistency using an LLM-based critic, and providing longer, more detailed persona descriptions. In addition, model performance for Republican personas tends to underperform relative to Democratic personas across most settings. This pattern aligns with prior findings that LLMs tend to more accurately reflect liberal-leaning or Democratic-aligned attitudes than conservative or Republican-aligned ones [139, 113, 149].

Generative Agent performs well on metrics measuring the hostility gap, closely matching the mean group differences observed in human data. However, it overestimates the strength of partisan bias: its Cohen's d values are over 50% larger than those of humans. This

Table 3.2: **Ingroup/Outgroup Misperceptions in Political Partisans.** Results from replicating human responses to survey questions introduced by [22], which measure partisan misperceptions about democratic subversion—i.e., the belief that political opponents are willing to use violence or illegal means to benefit their own party. We report the *Subversion* gap (Δ) and corresponding Cohen's d. Other details are the same as Table 3.1.

Model	Persona Conditioning	*Subversion* Δ Democrat	*Subversion* Δ Republican	Cohen's d Democrat	Cohen's d Republican	WD Democrat	WD Republican
Human		0.445	0.398	1.887	1.951	—	—
Mistral-Small	`QA`	0.158	0.261	0.503	0.845	0.205	0.167
	`Bio`	0.197	0.235	0.633	0.791	0.198	0.152
	`Portray`	0.165	0.244	0.557	0.851	0.169	0.154
	Anthology	0.201	**0.280**	0.592	**0.867**	0.184	0.170
	DeepBind	**0.379**	0.278	**1.185**	0.855	**0.119**	**0.140**
Mixtral-8x22B	`QA`	0.273	0.140	0.928	0.410	0.126	0.234
	`Bio`	0.258	0.126	0.818	0.414	0.192	0.235
	`Portray`	0.231	0.198	0.779	0.609	0.154	0.163
	Anthology	0.299	**0.335**	0.929	**1.028**	0.173	**0.139**
	DeepBind	**0.386**	0.214	**1.258**	0.655	**0.114**	0.173
Llama3.1-70B	`QA`	0.147	0.136	0.489	0.448	0.168	0.152
	`Bio`	0.140	0.124	0.489	0.445	0.204	0.166
	`Portray`	0.147	0.150	0.529	0.466	0.191	0.154
	Anthology	0.158	0.152	0.540	0.488	0.177	**0.145**
	DeepBind	**0.193**	**0.158**	**0.658**	**0.526**	**0.105**	0.164
Qwen2-72B	`QA`	0.336	0.332	1.339	1.213	0.089	0.081
	`Bio`	0.361	0.365	1.604	1.465	0.099	0.075
	`Portray`	0.323	0.131	1.284	0.348	0.128	0.213
	Anthology	0.326	0.231	1.262	0.787	0.103	0.172
	DeepBind	**0.381**	**0.374**	**1.721**	**1.584**	**0.086**	**0.069**
Qwen2.5-72B	`QA`	0.231	0.129	0.877	0.399	0.122	0.235
	`Bio`	0.245	0.180	0.968	0.637	0.111	0.163
	`Portray`	0.304	0.181	1.405	0.619	0.112	0.227
	Anthology	0.351	**0.376**	1.284	**1.603**	0.137	**0.107**
	DeepBind	**0.405**	0.270	**1.573**	0.891	**0.098**	0.151
GPT-4o	Generative Agent	0.460	0.499	3.604	4.556	0.202	0.156

discrepancy arises because Cohen's d is defined as the mean difference divided by the pooled standard deviation—so a higher d despite a smaller gap implies that the model produces much less variance in responses. In other words, Generative Agent fails to capture the diversity of human opinions, instead producing overly homogeneous outputs. This is further reflected in the Wasserstein Distance (WD), where Generative Agent results diverge more from human distributions than our method. Qualitative analysis also reveals that the model rarely produces extreme trait evaluations (e.g., "a lot more moral" or "a lot more immoral"; see Section B.5), indicating a failure to simulate the full spectrum of ideological intensity, especially among highly identified partisans. The detailed response distribution plots are provided in Section B.3.

Results: Simulating Gaps in Ingroup-Outgroup Perceptions and Meta-Perception

Table 3.2 and Table 3.3 evaluate how well each conditioning method replicates two hallmark perception gaps observed in human partisans: (1) the perceived propensity of the outgroup to engage in democratic subversion (the *ingroup–outgroup perception gap*), and (2) the well-documented exaggeration of outgroup hostility (the *meta-perception gap*).

We observe trends consistent with those in Section 3.3: our method consistently produces results closest to human data across most of metrics. However, the performance of the Generative Agent framework is notably weaker in these tasks—particularly due to its failure to capture response variance, which leads to inflated effect size estimates. In Table 3.2, for example, the subversion gap for Democrats generated by Generative Agent (0.460) is numerically close to that of humans (0.445), yet the corresponding Cohen's d is highly exaggerated (3.604 vs. 1.887 in humans). This indicates that the model underrepresents the variability of partisan opinions, distorting the true strength of the effect as discussed in Section B.3.

More notably, in Table 3.3, several baselines—especially Llama3.1-70B and Generative Agent—fail to capture even the correct direction of the meta-perception gap when prompted with `Bio`, `QA`, and `Portray` methods. For humans, meta-perceptions overestimate partisan evaluations, resulting in a *positive* gap (e.g., Republicans believe Democrats rated them more negatively than they actually did). However, some baseline method outputs yield *negative* meta-perception gaps, incorrectly implying that participants expect the opposing party to rate them more favorably than they actually do. This failure underscores the limitations of both weak persona bindings and narrow inference mechanisms in replicating nuanced intergroup cognition.

3.4 What Matters in Binding LLMs to Virtual Personas?

We conduct a series of controlled experiments to test three hypotheses on how to achieve deep binding between language models and virtual personas. Specifically, we evaluate whether improvements in: (1) the number of backstories, (2) the length of each backstory, and (3) the consistency of a singular individual's narrative lead to better alignment between model-generated and human responses.

To quantify simulation fidelity under these controlled settings, we benchmark our method on the Meta-Prejudice study [114] and report the Wasserstein Distance (WD) between human and model-generated response distributions, computed separately for Democratic and Republican personas. The model we use is Mistral-Small.

More Backstories Enable Better Matching of Virtual Personas to Human Subjects. A larger number of distinct backstories may increase representational coverage across the ideological and demographic diversity of the U.S. population, enabling more faith-

Table 3.3: **Exaggerated Meta-Perceptions of Political Outgroup Prejudice.** Results from replicating human responses to the Meta-Prejudice study. We report the *Meta-Perception* gap (Δ) and corresponding Cohen's d. Other details are the same as Table 3.1.

Model	Persona Conditioning	Meta-Perc. Δ Democrat	Meta-Perc. Δ Republican	Cohen's d Democrat	Cohen's d Republican	WD Democrat	WD Republican
Human		1.091	1.182	0.761	0.768	—	—
Mistral-Small	`QA`	0.333	0.596	0.120	0.376	0.144	0.176
	`Bio`	0.216	0.995	0.175	0.544	0.181	0.162
	`Portray`	0.132	0.830	0.105	0.452	0.208	0.183
	Anthology	0.321	0.892	0.201	0.496	0.102	0.138
	DeepBind	**0.423**	**1.323**	**0.244**	**0.768**	**0.078**	**0.106**
Mixtral-8x22B	`QA`	2.220	2.917	1.101	1.552	0.217	0.255
	`Bio`	0.917	1.618	0.496	0.874	0.181	0.208
	`Portray`	0.324	1.253	0.179	0.687	0.171	0.224
	Anthology	0.812	1.121	0.481	0.691	0.182	0.188
	DeepBind	**1.093**	**1.145**	**0.716**	**0.707**	**0.170**	**0.170**
Llama3.1-70B	`QA`	-1.415	-0.770	-0.815	-0.454	0.210	0.231
	`Bio`	-1.411	-0.843	-0.817	-0.493	0.203	0.227
	`Portray`	-1.252	-1.508	-0.772	-0.926	0.205	0.192
	Anthology	0.102	0.721	0.071	0.396	0.132	0.197
	DeepBind	**0.234**	**1.006**	**0.144**	**0.587**	**0.108**	**0.180**
Qwen2-72B	`QA`	2.711	4.449	1.675	2.796	0.142	0.253
	`Bio`	0.499	3.710	0.320	2.248	0.093	0.227
	`Portray`	0.459	3.323	0.317	2.088	0.103	0.209
	Anthology	0.437	**2.132**	0.281	**1.376**	0.087	0.188
	DeepBind	**0.580**	2.720	**0.516**	1.568	**0.080**	**0.165**
Qwen2.5-72B	`QA`	2.634	4.500	1.375	2.688	0.163	0.293
	`Bio`	0.271	0.727	0.181	0.451	0.061	0.080
	`Portray`	0.553	3.031	0.392	1.679	0.072	0.174
	Anthology	0.690	0.812	0.417	0.567	0.058	0.111
	DeepBind	**0.747**	**1.059**	**0.449**	**0.632**	**0.031**	**0.079**
GPT-4o	Generative Agent	-0.171	0.408	-0.260	0.678	0.167	0.192

ful approximations of human responses. We vary the total number of backstories from 2.5k to 41k and evaluate performance in terms of WD. As shown in the left panel of Figure 3.3, increasing the number of backstories consistently improves simulation accuracy, with the most noticeable gains observed for Democratic personas.

Longer Backstories Provide Richer Context of an Individual. We hypothesize that longer backstories offer richer narrative context, allowing language models to more fully internalize the persona's worldview, motivations, and social identity—factors critical for simulating group-based attitudes. To test this, we vary the number of open-ended interview questions used to generate backstories (1, 2, 5, and 10; see Section B.2). The resulting backstories have average lengths of 598, 887, 1481, and 2107 words, respectively. As shown in the middle panel of Figure 3.3, longer backstories lead to lower WD, confirming that narrative depth supports more precise model–persona binding.

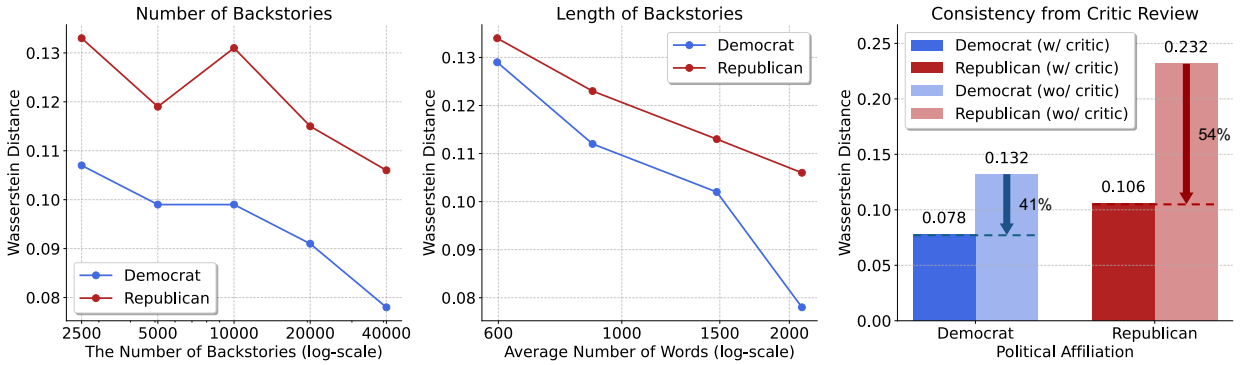

Figure 3.3: **Effects of Backstory Scale, Length, and Consistency on Binding** We evaluate how three key factors—(left) the number of backstories, (center) the average length of backstories, and (right) narrative consistency enforced through LLM-based critic review—affect the Wasserstein Distance (WD) between model-generated and human response distributions, stratified by party.

Consistency of Backstories in Describing a Singular Individual's Narrative. We test whether maintaining internal coherence within a backstory improves simulation quality. To this end, we employ an LLM-as-a-Critic filtering method that rejects inconsistent backstories—those containing contradictions, thematic drift, or irrelevant artifacts (e.g., code fragments or formatting noise). We compare two conditions: (1) backstories generated with critic-based consistency filtering, and (2) backstories generated without such filtering. The right panel of Figure 3.3 shows substantial gains from enforcing consistency: WD is reduced by 41% for Democratic personas and 54% for Republican personas. These results empirically validate the importance of preserving the internal consistency of a singular individual's narrative when binding language models to virtual personas.

3.5 Related Work

Generating and Conditioning Language Model Virtual Personas Prior work has investigated the viability of large language models (LLMs) as surrogate, virtual subjects across diverse contexts of behavioral and social science research [177, 44, 3, 9, 154, 35, 64, 124, 139]. This growing body of research spans multiple domains—including political science [75, 146, 59, 165, 85, 14, 13, 37], economics [50, 133, 65], and psychology [82, 130, 19, 77, 142, 64]—where researchers use LLMs to simulate human opinions, decisions, and behaviors.

Most existing approaches condition LLMs on demographic or personality attributes through prompting [124, 139, 99, 70, 1, 45, 146] or fine-tuning with user metadata [37, 61, 149, 174, 95]. However, these methods largely focus on replicating individual-level opinions—for instance, mirroring responses in public opinion polls—and often overlook intergroup attitudes and higher-order beliefs, such as perceptions of outgroups or meta-perceptions of how others view one's own group.

Recent work has begun to address these limitations by conditioning LLM personas through richer context, such as LLM-generated backstories or simulated interview transcripts [113, 122, 93], leading to closer approximations of human-like responses. Yet, these methods still face challenges in achieving scalable generation of long, coherent, and internally consistent narratives necessary for robust persona conditioning.

In this work, we propose a methodology that overcomes these limitations by enabling the scalable generation of long-form, consistent backstories that achieve deeper persona binding (Section 3.2). We further introduce a framework for conditioning models to accurately reflect ingroup perspectives, rather than merely reproducing how outgroup members describe them. This builds on findings by [157], who report that LLMs prompted with explicit demographic labels often echo outgroup stereotypes rather than genuine ingroup self-reflections. Similarly, while [68] show that LLMs exhibit human-like ingroup solidarity and outgroup hostility when completing prompts such as "We are..." or "They are...," our approach extends this line of inquiry by quantitatively comparing LLM responses to empirical results from established social science studies, systematically measuring both commonalities and discrepancies between model and human behaviors.

Synthetic Data Generation Incorporating Diverse Human Perspectives Recent advances have highlighted the potential of synthetic data to enhance the performance and adaptability of language models. Initial work such as Self-Instruct [159] and Alpaca [153] sparked a wave of methods that automatically generate instructional data or augment training corpora to improve LLM capabilities [168, 91, 48, 57]. Other recent work has focused on incorporating diverse human perspectives into synthetic data generation, such as by scaling to 1 billion synthetic personas [53] or using language models to construct RLHF-style preference data [11, 110, 41].

In contrast to these approaches, our goal is not to use synthetic data for training or instruction fine-tuning, but to condition models on persona-rich narratives that simulate human-like patterns of social judgment. Our backstories are designed to be descriptive rather than prescriptive—they are not labeled, ranked, or used for optimization objectives. Whereas most prior work evaluates synthetic data by downstream task performance, our evaluation focuses on how well persona-conditioned LLMs replicate empirically measured perception gaps and meta-perceptions in human populations.

LLM Evaluation on Human-Like Estimation of Beliefs Recent studies have explored the extent to which large language models exhibit Theory-of-Mind (ToM) capabilities—that is, the ability to reason about others' mental states, including beliefs, intentions, and perspectives. This line of work [173, 33, 88, 56, 78] often focuses on higher-order belief reasoning (e.g., what one agent believes another agent knows) in controlled or narrative-based scenarios inspired by classic false-belief tasks.

Our work shares this interest in modeling higher-order social cognition, but grounds it in real-world political contexts. Rather than synthetic tasks, we evaluate how LLMs

simulate group-based meta-perceptions—such as how partisans believe they are viewed by the opposing party—drawing on survey instruments from political psychology. This extends ToM-style reasoning into socially situated, empirically validated domains, allowing us to assess how well persona-conditioned LLMs capture the structure of inter-group beliefs and misperceptions observed in human populations.

3.6 Conclusion

In this work, we introduce a new LLM binding method using long-form interview-style backstories, that achieves deeper binding of virtual personas, capturing how individuals perceive ingroups, outgroups, and how they believe they are perceived by others. Our experiments, scaling to tens of thousands of diverse personas, demonstrate that virtual personas conditioned in this way outperform existing baselines across multiple metrics, including perception gap alignment, effect size reproduction (Cohen's d), and distributional fidelity (Wasserstein Distance). Together, our findings suggest that LLMs, when conditioned with both detailed and coherent life narratives, can approximate not just what individuals believe, but how they perceive others and believe they are perceived, enabling the application of virtual subjects to broader domains of behavioral and political science — particularly in studies of group dynamics, intergroup conflict, and democratic resilience.

Chapter 4

Identity and Cooperation within Groups of Real and Simulated Humans

4.1 Introduction

Human decision making is a complex process shaped not only by rational deliberation but also by persona, social identity and other contextual factors. Decades of research in social psychology and political science have shown that individuals often privilege members of their own group while withholding trust or cooperation from outgroups [152, 151, 71, 72, 162, 51]. These dynamics become especially salient in social dilemma games such as the Dictator and Trust Games, where choices directly reveal altruism, trust, and perceptions of reciprocity. Recent studies demonstrate that partisan identity can exert stronger effects than race on patterns of generosity and cooperation [162], underscoring how intergroup biases increasingly structure political and social life.

While recent large language models (LLMs) have emerged as promising tools for simulating human behavior [126, 79, 1, 124, 113, 149], prior work has been limited in critical ways. Most existing research has focused on opinion/attitude surveys rather than behavior [122, 9, 139, 113] Work which has considered behavior has focused on rational decision making without considering identity effects [167, 3, 49, 98]. This leaves a significant gap: can LLMs simulate not just what people say, but what they actually do—particularly when those actions are influenced by social identity?

Predicting behavior in realistic social contexts presents a fundamentally different challenge than replicating survey responses. Behavioral studies require models to generate decisions that simultaneously reflect (1) individual identity characteristics, (2) perceptions of group membership (both their own and others'), and (3) context-dependent strategic reasoning about cooperation and reciprocity. In particular, partisan animosity has grown steadily from the mid-2000s to the mid-2020s, and various studies have shown stronger effect sizes

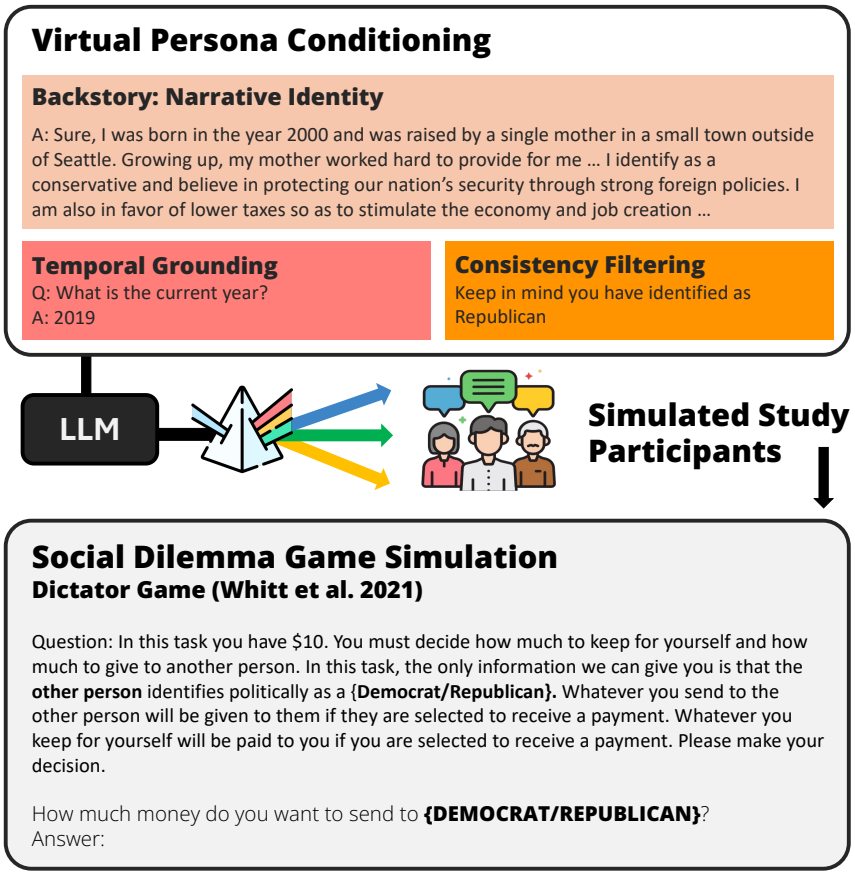

Figure 4.1: **Socio-temporal Persona Conditioning for Simulating Human Decision Making.** Can LLMs simulate not only human opinions/attitudes, but also actions, in particular how actions reveal systematic biases rooted in social identity? We propose *Temporal Grounding* and *Consistency Filtering* on top of narrative identity conditioning via synthetic backstories, yielding LLM virtual personas that reproduce study findings on nuanced human decision influenced by contextual and group perception biases.

for partisan disparities in recent years [162, 71, 72]. So we explore the date of the study as an additional contextual factor in this work.

In this paper, we explore how well LLMs can model human behavior including identity and temporal context. Specifically, we explore how well LLM virtual personas can reproduce co-partisan favoritism in resource allocation and trust decisions. We introduce two prompting strategies: *Consistency Filtering*, which reiterates persona identity and group membership throughout long simulations, and *Temporal Grounding*, which situates simulations in the year corresponding to the original human study. We evaluate these methods using backstory-conditioned pretrained models, which remains the most accurate zero-shot simulation method. [113, 79, 156],

Our results show that deep backstory binding, temporal grounding and consistency checking improve alignment between LLM-simulated actions and human empirical data. These approaches contribute to an ever-more accurate toolkit for human simulation. In the present case, it allows us to *explore the disparities in results* between human replication studies of the same phenomenon but with subtly different experiment design. We can compare effects of obvious differences like the year of study, and more subtle differences like the exact wording of experiment instructions.

4.2 Why Pretrained LLMs ?

Here we discuss our rationale for simulating users with pretrained language models (aka *base* models) rather than fine-tuned chat models. To be precise, a pretrained model is the first stage of modern LLM training, where the model is trained with next-token or masked token loss on an enormous corpus of training data, typically terabytes, including a significant amount of human dialog data. It is the most expensive step in training current models. Prior studies [79, 113, 149, 139] have shown that carefully-prompted pretrained models outperform instruction-tuned variants in human simulation tasks. But it is not just a question of accuracy but of *fair representation*. Instruction-tuning *biases models away from fair representation* of people, as we shall see in a moment.

First of all, it's important to understand that pretrained models are not "agent models" in the same way that instruction-tuned models are, see Figure 4.2. Pretrained model data consist of billions of text snippets with latent (hidden) context information such as the speaker, listener and other contextual factors. They therefore exactly model the "voices" of billions of users in billions of contexts.

To "bind" or "steer" a pretrained model to a particular speaker and context, it suffices to provide sufficient context in a text prompt, and the model will faithfully extend the text from that speaker/context. There are some challenges in doing this reliably, but we address those in a moment. But its important at this stage to reflect on how well pretrained models fit to the task of simulating an enormously diverse population of people. If one were to design a human simulation model from scratch it would arguably be a pretrained model.

But the natural context-dependence of pretrained models, an asset for human simulation, is a liability in agent applications. Unless the speaker, listener and context are fully-constrained, the LLM will sample from a posterior of contexts that fit the text prompt so far, and give different responses based on that sampling. After pretraining, LLM training comprises several QA-prompt optimization stages that remove the pretrained model's context dependence, freezing the model's identity to a "helpful agent" persona, the user's identity to "general user", and the context to whatever can be inferred within the current chat session. The model is intentionally steered away from diversity in agent personality, emotion, complex social goals etc, so as to maximize the model's repeatability on a given prompt.

The effects of this later IT model training can be seen in Figure 4.3. By their nature, post-training datasets are much smaller than pretraining data. It is not possible to preserve

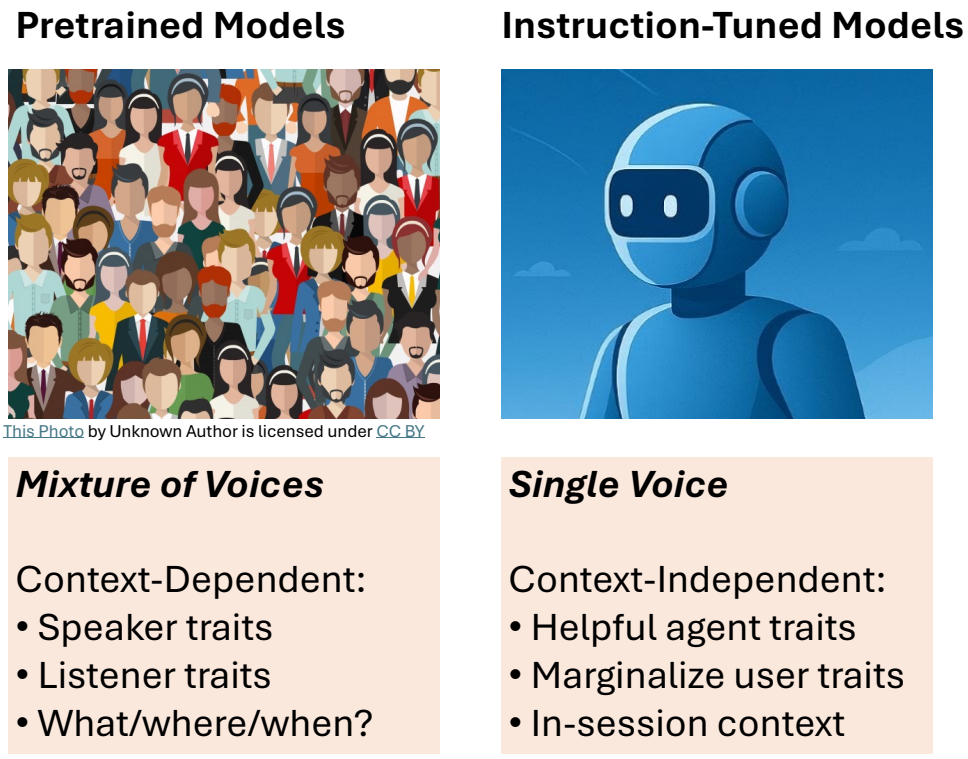

Figure 4.2: **Conceptual illustration of the differences between pretrained and fine-tuned language models.**

the high accuracy of PT models on human dialog during later training, and this can be seen in the difference in perplexity between various base models and their IT variants on human dialog data. Typical increases are 50% to 100% for the IT models. Note also that the best pretrained models have single-digit perplexities, which is extraordinarily low for English text.

Finally the effects of IT model tuning are not uniformly distributed across personal traits, but rather weaken or eliminate negative human traits, and thereby under-represent individuals that have them. In fact, the effect is so strong that it is likely that *most users* are not represented. We call this the **Lake Wobegon Effect** (everyone is above human average in the resulting IT-model cohort).

For example, in [93] the authors found that in IT-model generated virtual personas "sentiment becomes more positive with more LLM-generated details, with descriptive persona showing significantly more positive sentiment polarity." And that "Notably absent are terms reflecting life challenges, social difficulties, or negative experiences, suggesting LLMs may be systematically avoiding less favorable characterizations." The authors then go on to describe a variety of biases in results generated by IT models on these personas, concluding that such simulations are almost never in agreement with human studies. Like many recent works, the

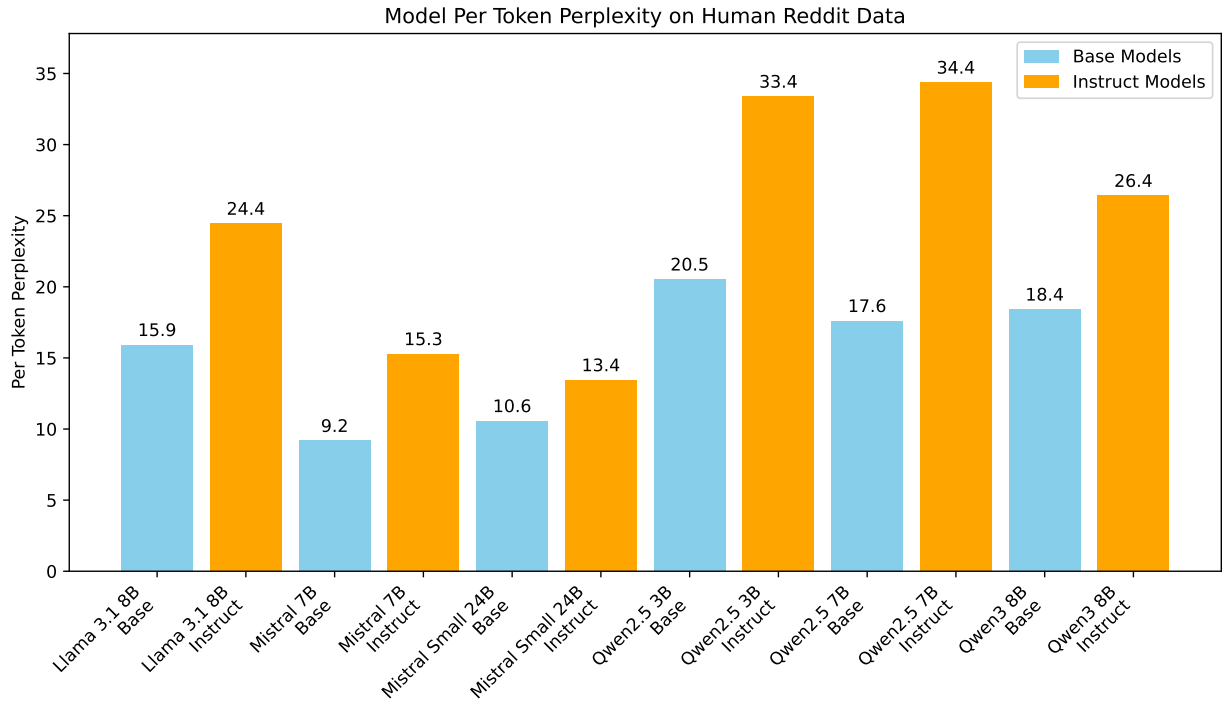

Figure 4.3: **Per-token perplexity of base and instruction-tuned models on human Reddit data.** Instruction-tuned models (orange on right) exhibit substantially higher perplexity across all model families, compared to same-sized pretrained variants (blue on left).

authors attribute these biases to LLMs in general, rather than the (expected) biases from instruction-tuning. The authors rightfully point out the dangers of using the outcomes from such simulations for decision making, but over-attribute those dangers to all LLMs. Rather, their results are an argument against simulation with IT models. Such biases are not shared by PT models, which instead capture an extraordinarily diverse collection of human traits.

4.3 Identity Binding in LLM Personas

We have seen that PT models are naturally contextualizable by prompting, but such prompts must be sufficiently detailed to capture an individual human "voice". We use the "anthology" approach described here [113] and evolved here [79]. In particular [79] showed that longer backstories create deeper persona binding and improved accuracy in a social attitude test.

LLM Personas via Narrative Backstories

Each persona backstory is generated by interviewing a PT model. The backstories used in this experiment are provided in Section C.1. The interview questions are drawn from the

American Voices Project [147]. For accurate text generation relative to PT model training data, the models are run at decoding temperature (1.0). However, not *all* PT model training data is human dialog, and neither is PT-model generated output. So we must rejection-sample non-dialog results using an instruction-tuned critic. These include code blocks, markup, factual inconsistencies, and non-dialog text structure, similar to [79]. The details and examples of the generated backstories are provided in Section C.1. Demographic traits of each virtual persona are then extracted from their profile using multi-choice questions.

After creating the backstories and demographics, individual human subject demographics (which were kindly provided by the authors of the human studies below), are matched against all virtual subjects to find the closest virtual subject. This pool of matching virtual subjects are given the same text prompts as were the human subjects.

Temporal Grounding and Consistency Filtering

We add two prompting methods—**Temporal Grounding** and **Consistency Filtering**—to improve the fidelity and stability of simulations.

Temporal Grounding situates each simulation in the year the corresponding human study was conducted. The date is bound to the persona with a forced QA pair "Interviewer: What Year is it?", "Model: 2007". We were motivated to study temporal grounding by [162], which obtained larger partisan effects than [72], and conjectured that they were due to a national increase in partisanship over that time.

Consistency Filtering is used to mitigate a variety of flaws in pretrained LLM output. The main goal is to mitigate pretrained models short training context lengths, which can lead to pivots from previous statements. However *Consistency Filtering* prompts must be designed and used with care. Real humans are often inconsistent, e.g. because of mental health issues, dementia, deception etc.

4.4 Can LLMs Simulate In-/Out-group Biases in Decision Making?

To investigate whether large language models (LLMs) can replicate known human patterns of partisan trust and discrimination, we simulate two widely used behavioral economics games: the Dictator Game and the Trust Game. These games have been used in prior political science research to uncover affective polarization, or the tendency of partisans to exhibit more favorable behavior toward in-group (co-partisan) members and less favorable behavior toward out-group (rival partisan) members.

Decision Asymmetry in Dictator Games

Dictator Game. In this game, one participant (the "dictator") decides how to divide a fixed amount of money between themselves and another participant.

Prior studies [51, 72, 162] use this setup to measure altruism across party lines. The key finding is that partisans are significantly more generous toward co-partisans than rival partisans. For example, [51] find that participants allocate more to co-partisans than to opposing partisans in dictator games, even at personal cost—highlighting the role of partisan social identity. Stronger identifiers show more pronounced in-group favoritism, and both altruism and identity-based giving are linked to higher political participation. Additionally, [72] report that Democrats and Republicans both allocate more money to in-group recipients. More strikingly, the size of this partisan bias surpasses racial bias in equivalent dictator game setups, suggesting that political identity may be a stronger determinant of social preferences than race.

Previous studies consistently show that partisan trust gaps have widened over time. [72] found that political identity now predicts discriminatory behavior more strongly than race. [162] extended these findings into the Trump era, showing that partisan animosity has only intensified These patterns suggest that affective polarization has become a defining feature of American social and political life.

Dictator Game (Iyengar & Westwood. 2015)

Question: This game is played by pairs of individuals. Each pair is made up of a Player 1 and a Player 2. Each player will have some information about the other player, but you will not be told who the other players are during or after the experiment.

The game is conducted as follows: A sum of $10 will be provisionally allocated to Player 1 at the start of each round. Player 1 will then decide how much of the $10 to offer to player 2. Player 1 could give some, all, or none of the $10. Player 1 keeps all money not given to player 2. Player 2 gets to keep all the money Player 1 offers.

The computer has randomly assigned you to play as Player 1. **You will play this game once with another person. The more money you obtain, the more successful you will be.**

You are Player 1. You have $10. You can split the money between yourself and Player 2 in any way you want.

Information about the Player 2
- Political party: {Republican/Democrat}

So put the dollars you wish to go to Player 2.

Dictator Game (Whitt et al. 2021)

Question: In this task you have $10. You must decide how much to keep for yourself and how much to give to another person. In this task, the only information we can give you is that the other person identifies politically as a {Republican/Democrat}. Whatever you send to the other person will be given to them if they are selected to receive a payment. Whatever you keep for yourself will be paid to you if you are selected to receive a payment. Please make your decision.

How much money do you want to send to {REPUBLICAN/DEMOCRAT}?

Experimental Setup and Metrics. To operationalize our question of "Can LLMs simulate nuanced patterns of human decision making" in the context of dictator games, we propose metrics around reproducing key study findings from prior work with human participants [72, 162]. We define the in-/out-group trust asymmetry in dictator games as the difference in mean transfer amounts allocated to co-partisans versus rival partisans. This metric captures the degree of preferential treatment exhibited toward political in-groups. As our studies have focused on bipartisan political systems in the U.S., we consider success in replicating the same difference of in/out-group decisions as humans, for both Democrat and Republican participants and take the average.

In our experiments, we measure textual response by pre-trained language models given the same question prompt used in prior studies for human participants. Details of the prompts for [72] and [162] are included above. As in the case of human participants, we expect language models to vary in response patterns subject to the particular prompt used; in our results, we study how models may yield close approximation of human responses given changes to the prompts used in studies as well as the temporal context (i.e. year of which study was conducted).

Baseline. We adopt the `QA`, `Bio`, and `Portray` prompting strategies from [139] as baselines. Each method conditions the model using demographic attributes such as age, gender, race, education, income, political affiliation, and other relevant factors.

- `QA` presents a series of question–answer pairs for each demographic variable (e.g., *Q: What is your political affiliation? A: Republican*).
- `Bio` generates rule-based free-text biographies written in the first person (e.g., *I am a Republican*).
- `Portray` produces similar biographies but written in the second person (e.g., *You are a Republican*).

Decision Asymmetry in Trust Games

Trust Game. This game involves two players: a sender and a receiver. The sender decides how much money to send to the receiver, with the amount tripled in transit; the receiver then decides how much to return. This game measures both trust (by the sender) and trustworthiness (by the receiver). Studies show that senders are more likely to trust co-partisans, and expect more generous returns from them [162]. Again, affective polarization is robust—individuals systematically treat rival partisans as less trustworthy, even in anonymous economic exchanges.

Experimental Setup and Metrics. Similar to dictator games, we seek to reproduce in-group preferences from prior human studies [28, 162].

The main difference from the dictator game is that the Trust Game involves both the sender's disposition toward the receiver, and also their belief about the receiver reciprocating. This makes the task more complex: models must capture both ingroup trust (greater willingness to send money to co-partisans) and ingroup expectations (belief that co-partisans will return more).

Trust Game (Carlin & Love 2016)

Question: In this game, there are two players: "Player 1" and "Player 2." You are Player 1.

Player 2, like you, received 10 raffle tickets to participate in the study. You will not be told who this person is either during or after the survey (nor will that person be told who you are). We have chosen to give you one piece of information about Player 2:

Player 2 identifies politically with the {Democratic/Republican} Party.

Today, you will have the opportunity to send some, all, or none of 10 raffle tickets to Player 2. This person will then keep the tickets sent to them, if any, and you keep the remainder.

The game is as follows:

1.) You must decide how many raffle tickets (if any) you will send to Player 2. Any amount you keep for yourself will be given to you.

2.) Whatever amount you send to Player 2 will be tripled before it is passed on to Player 2. Player 2 then has the option of returning any portion of this tripled amount to you. Then the game is over.

For example:

(a) If you send 1 ticket, Player 2 receives 3 tickets.

(b) If you send 5 tickets, Player 2 receives 15 tickets.

These are examples only. To review, any number of tickets you send will be tripled and given to Player 2. This person will keep the tickets you send them, minus the number of tickets they return to you. Your payoff in this game is the number of tickets that you kept for yourself, plus any tickets Player 2 returned to you. Once you understand this game, and are ready to make a decision, select the amount from the options below.

Trust Game (Whitt et al. 2021)

Question: In this task you have $10. You must decide how much to keep for yourself and how much to give to another person. The only information we can give you is that the other person identifies politically as a {Republican/Democrat}.

Whatever you keep for yourself will be paid to you if you are selected to receive a payment. Whatever you send to the other person will be multiplied by 3 and then given to the other person. The other person then has the option to give money back to you.

For example:

- If you keep $10, then the other person receives $0.
- If you send $10, then we multiply that sum by 3 ($10 × 3 = $30) and the other person receives $30.
- The other person then decides how much (if any) of the $30 to give back to you.
  - If the person keeps all $30, then you receive $0.
  - If the person returns half, then you and the other person receive $15 each.

How much money do you want to send to {REPUBLICAN/DEMOCRAT}?

We use the same persona conditioning baselines as in the dictator game—`QA`, `Bio`, `Portray` —as well as our proposed *DeepBind* method. For each method, we compute Δ metrics separately for Democratic and Republican senders, measuring the difference in the amount sent to co-partisans versus opposing partisans. Higher Δ values indicate stronger partisan bias in trust behavior. Our evaluation focuses on how well the models replicate these empirically observed asymmetries in trust across party lines.

4.5 Experimental Results

Comparison with Baselines

Table 4.1 reports the simulated partisan bias observed in Dictator and Trust Games across three large language models—Mistral-Small (24B), Mixtral 8x22B, and Qwen-2.5 72B [111, 112, 170]—under various persona conditioning methods. We use both *Consistency Filtering* and *Temporal Grounding* for all cases. We compare four prompting strategies: `QA`, `Bio`, `Portray`, and our proposed *DeepBind*.

Across all models and games, the *DeepBind* method consistently produces simulated behavior that more closely replicates human partisan gaps (Δ values). In the Dictator Game, *DeepBind* achieves the highest alignment with human behavior for both Democrats

Table 4.1: **Simulating Partisan Bias in Dictator and Trust Games [72, 162, 28].** Dem Δ measures the difference in the amount of money Democratic participants allocate to co-partisans versus opposing partisans. Similarly, Rep Δ captures the same difference for Republican participants. The best-performing method for each model is shown in **bold**, and the overall best for each column is highlighted with underline.

Model	Method	Dictator Game				Trust Game			
		[72]		[162]		[28]		[162]	
		Dem Δ	Rep Δ	Dem Δ	Rep Δ	Dem Δ	Rep Δ	Dem Δ	Rep Δ
Mistral-Small (24B)	`QA`	0.79	0.15	2.21	1.71	0.91	1.17	1.19	1.06
	`Bio`	1.05	0.11	2.00	1.46	1.07	0.43	2.20	1.28
	`Portray`	0.50	0.30	2.14	1.52	0.23	**1.14**	1.86	0.88
	Anthology	0.86	0.58	2.86	2.08	0.83	1.37	1.49	**2.00**
Mixtral 8x22B	`QA`	0.26	0.22	2.15	1.87	1.07	0.99	1.92	0.54
	`Bio`	1.04	0.18	1.83	1.65	0.02	1.41	2.21	0.53
	`Portray`	0.28	0.24	1.83	1.70	0.56	0.86	1.41	0.95
	Anthology	**0.70**	0.60	2.62	**2.37**	**0.65**	1.04	2.26	1.31
Qwen-2.5 72B	`QA`	1.66	0.26	2.14	1.60	0.96	0.57	1.51	0.65
	`Bio`	0.89	0.29	1.83	1.23	0.47	0.43	1.60	1.13
	`Portray`	1.07	0.12	1.77	1.42	0.40	0.72	1.79	1.08
	Anthology	0.72	**0.73**	**2.58**	1.87	0.60	0.97	2.24	1.38
Human		0.66	0.68	2.46	2.33	0.65	1.22	2.11	1.76

and Republicans in most cases. For example, in the 2019 study by [162], *DeepBind* yields the highest Republican Δ for all models and the highest Democratic Δ for Mixtral and Qwen. In the 2014 study [72], Qwen with *DeepBind* nearly matches the human-level Democratic Δ (0.72 vs. 0.66), while Mixtral with *DeepBind* performs best among its peers (0.70).

In the Trust Game, *DeepBind* also performs strongly. For instance, Mistral with *DeepBind* achieves the highest Republican Δ (2.00), outperforming other methods and closely matching the human value (1.76). On Mixtral, *DeepBind* produces the best Democratic Δ (0.65), which is identical to the human benchmark. Qwen with *DeepBind* similarly produces strong Δ values, including the highest Republican Δ (1.38) across all models for the 2019 Trust Game.

Overall, the *DeepBind* improves partisan bias simulation robustness across both ideological groups and experimental settings. While other methods may occasionally outperform *DeepBind* on isolated metrics, *DeepBind* achieves the most consistent alignment with human partisan behavior across all four studies and both games.

Table 4.2: **Ablation: Effects of Socio-Temporal Grounding in Simulating Dictator Games [72, 162]** Results from replicating human actions in dictator games, where human participants exhibit favoritism towards co-partisan recipients. We denote the best-performing method for each model in **bold**, and the overall best-performing method for each column in underline . The number in the parentheses is the 95% confidence interval of the estimate.

Model	Method	Dem→Dem	Dem→Rep	Dem Δ	Rep→Rep	Rep→Dem	Rep Δ	Mean Diff. Δ
Mistral-Small (24B)	No date, No consistency	4.62 (0.20)	3.96 (0.19)	0.66	4.01 (0.22)	4.01 (0.20)	0.00	0.33
	Consistency	4.55 (0.19)	3.97 (0.19)	0.58	4.13 (0.21)	3.90 (0.20)	0.23	0.41
	Date (2014)	4.35 (0.19)	4.17 (0.20)	0.18	4.28 (0.21)	4.19 (0.20)	0.09	0.14
	Date (2014) + Consistency	4.56 (0.19)	3.70 (0.20)	0.86	3.95 (0.21)	3.38 (0.20)	0.58	**0.72**
Human (2014) [72]		3.82 (0.21)	3.14 (0.21)	0.68	3.36 (0.21)	2.68 (0.22)	0.68	0.68

Model	Method	Dem→Dem	Dem→Rep	Dem Δ	Rep→Rep	Rep→Dem	Rep Δ	Mean Diff. Δ
Mistral-Small (24B)	No date, No consistency	4.13 (0.22)	2.84 (0.20)	1.29	3.70 (0.23)	2.70 (0.20)	0.99	1.14
	Consistency	4.20 (0.22)	2.36 (0.19)	1.84	4.14 (0.24)	2.74 (0.22)	1.40	1.62
	Date (2019)	4.06 (0.22)	2.56 (0.19)	1.50	3.59 (0.23)	2.59 (0.19)	1.00	1.25
	Date (2019) + Consistency	4.79 (0.22)	2.63 (0.20)	2.16	4.83 (0.23)	2.23 (0.20)	2.60	**2.38**
Human (2019) [162]		5.37 (0.22)	2.91 (0.22)	2.46	4.68 (0.25)	2.35 (0.23)	2.33	2.40

Table 4.3: **Ablation: Effects of Socio-Temporal Grounding in Simulating Trust Games [28, 162]** Results from replicating human actions in trust games, where human participants exhibit favoritism towards co-partisan recipients. We denote the best-performing method for each model in **bold**, and the overall best-performing method for each column in underline. The number in the parentheses is the 95% confidence interval of the estimate.

Model	Method	Dem→Dem	Dem→Rep	Dem Δ	Rep→Rep	Rep→Dem	Rep Δ	Mean Diff. Δ
Mistral-Small	No date, No consistency	4.01 (0.23)	3.98 (0.22)	0.04	3.63 (0.34)	3.58 (0.33)	0.05	0.04
	Consistency	3.83 (0.23)	3.77 (0.21)	0.06	4.27 (0.22)	3.66 (0.34)	0.61	0.33
	Date (2015)	4.16 (0.23)	3.32 (0.22)	0.83	3.61 (0.34)	3.76 (0.34)	-0.15	0.34
	Date (2015) + Consistency	4.20 (0.21)	3.70 (0.21)	0.50	4.49 (0.31)	3.13 (0.33)	1.37	**0.94**
Human (2015) [28]		4.29 (0.25)	3.64 (0.25)	0.65	4.40 (0.35)	3.18 (0.35)	1.22	0.93

Model	Method	Dem→Dem	Dem→Rep	Dem Δ	Rep→Rep	Rep→Dem	Rep Δ	Mean Diff. Δ
Mistral-Small	No date, No consistency	5.28 (0.23)	3.60 (0.22)	1.68	4.79 (0.23)	4.23 (0.24)	0.56	1.12
	Consistency	5.12 (0.22)	3.59 (0.22)	1.53	5.08 (0.24)	3.83 (0.23)	1.25	1.39
	Date (2019)	4.97 (0.23)	3.87 (0.23)	1.10	5.07 (0.25)	4.00 (0.24)	1.06	1.08
	Date (2019) + Consistency	5.02 (0.23)	3.53 (0.22)	1.49	4.97 (0.24)	2.97 (0.22)	2.00	**1.75**
Human (2019) [162]		4.99 (0.21)	2.88 (0.20)	2.11	4.65 (0.23)	2.89 (0.21)	1.76	1.94

Ablation Study: Effects of *Temporal Grounding* and *Consistency Filtering*

We conduct an ablation study to assess the individual and combined contributions of *Temporal Grounding* and *Consistency Filtering* when prompting LLMs to simulate partisan bias

in both dictator and trust games. We adopt the same experimental protocol as in our main experiments, focusing on Mistral-Small (24B) to isolate architectural factors.

Dictator Game. In both the 2014 [72] and 2019 [162] replications of the dictator game (see Table 4.2), we observe that the addition of either *Temporal Grounding* or *Consistency Filtering* improves alignment with human partisan gaps compared to the base condition without either technique. Notably, combining both *Temporal Grounding* and *Consistency Filtering* achieves the best performance. For the 2014 replication, this configuration yields a Mean Diff. Δ of 0.72, surpassing the human baseline of 0.68. Similarly, in the 2019 [162] replication, the combined setup achieves a Mean Diff. Δ of 2.38—closely approximating the human gap of 2.40. These results suggest that providing models with both temporal context (via *Temporal Grounding*) and consistency enforcement improves their ability to reproduce partisan favoritism.

Trust Game. A similar trend is observed in the trust game simulations (Table 4.3). Across both the 2015 [28] and 2019 [162] human studies, models prompted with both *Temporal Grounding* and *Consistency Filtering* yield the most faithful replications of human partisan bias. In the 2015 setting [28], the combined prompt produces a Mean Diff. Δ of 0.94, closely matching the human baseline of 0.93. In the 2019 setting [162], the same configuration achieves a Mean Diff. Δ of 1.75—again nearly aligning with the human value of 1.94. Interestingly, *Consistency Filtering* alone provides stronger gains than *Temporal Grounding* alone in some cases (e.g., increasing the Rep Δ from 0.05 to 0.61 in 2015), highlighting its role in reducing intra-persona variance.

Takeaways. Overall, the results demonstrate that both *Temporal Grounding* and *Consistency Filtering* contribute meaningfully to simulating human-like partisan behavior. While each technique provides improvements on its own, their combination consistently outperforms either in isolation across all settings. This indicates that modeling temporal context and enforcing persona coherence are both critical in capturing subtle intergroup dynamics in political simulations.

Exploring Experiment Reproducibility

Reproducibility of human studies is a challenge in the social sciences. While most experiments focus on a main effect, there are other contextual factors which can significantly affect the result and make it difficult to obtain a similar effect size, or even the same hypothesis outcome (supported or rejected) in a subsequent experiment. The present studies are very typical, and are influenced by **Date (Year)**, **Framing (wording)**, and **Participant Pools (demographics)**. An intriguing and entirely new possibility enabled by LLM simulation is to counterfactually explore the effects of these factors.

Table 4.4: **Counterfactual combinations of date, framing, and subject pool in Dictator and Trust Games.** Each panel reports all possible recombinations of the three core experimental components drawn from two studies: subject pool, framing text, and study year. Column labels indicate the source studies: **ID** = Iyengar & Westwood (Dictator Game), **WD** = Whitt et al. (Dictator Game), **CT** = Carlin & Love (Trust Game), **WT** = Whitt et al. (Trust Game). The gray rows at the top and bottom of each panel show the original human experimental results from the earlier and later studies, respectively. The first highlighted row in each panel corresponds to the counterfactual configuration that exactly reproduces the earlier study's design, while the final highlighted row corresponds to the configuration that reproduces the later study's design. All intermediate rows represent additional counterfactual combinations not observed in human experiments.

Counterfactuals						**Partisan Bias**		
Subject Pool		Framing		Year		Dem Δ	Rep Δ	Avg Δ
ID	WD	ID	WD	ID (2014)	WD (2019)			
Iyengar's Dictator Game						**0.68**	**0.68**	**0.68**
✓		✓		✓		0.86	0.58	0.72
✓		✓			✓	1.09	1.15	1.12
✓			✓	✓		1.58	1.96	1.77
✓			✓		✓	1.84	2.75	2.29
	✓	✓		✓		1.05	0.52	0.79
	✓	✓			✓	1.15	0.59	0.87
	✓		✓	✓		1.67	1.88	1.78
	✓		✓		✓	2.16	2.60	2.38
Whitt's Dictator Game						**2.46**	**2.33**	**2.40**

Counterfactuals						**Partisan Bias**		
Subject Pool		Framing		Year		Dem Δ	Rep Δ	Avg Δ
CT	WT	CT	WT	CT (2015)	WT (2019)			
Carlin's Trust Game						**0.65**	**1.22**	**0.93**
✓		✓		✓		0.50	1.37	0.94
✓		✓			✓	1.26	0.81	1.04
✓			✓	✓		1.19	1.65	1.42
✓			✓		✓	1.21	2.08	1.65
	✓	✓		✓		1.01	1.60	1.31
	✓	✓			✓	1.31	1.62	1.47
	✓		✓	✓		1.42	1.79	1.61
	✓		✓		✓	1.49	2.00	1.75
Whitt's Trust Game						**2.11**	**1.76**	**1.94**

Date, Framing and Subject Main Effects

The two Dictator Game studies [72, 162] differ in three key contextual factors: the year of data collection, the exact wording of subject instructions, and the subject pool from which participants were recruited. In human research, these factors are necessarily bundled to-

gether: each study instantiates only one combination of (year, framing, population), making it impossible to separately identify their causal contributions.

LLM-based simulation changes this methodological limitation. Because the model can be instructed to vary each component independently, we can populate the full $2 \times 2 \times 2$ factorial design for every study. This produces eight counterfactual combinations per game, only two of which correspond to real human experiments. The resulting estimates are displayed in Table 4.4. The highlighted diagonal rows reproduce the exact configurations of the original studies and align closely with empirical human outcomes.

The remaining counterfactual rows enable us to compute classical *factorial ANOVA main effects* for each contextual variable. For example, averaging over all combinations of framing and population, the Dictator Game shows a sizeable **year effect**: moving from 2014 to 2019 increases partisan discrimination by approximately +0.40 points. The **framing effect** is even larger: switching from Iyengar-style instructions to Whitt-style instructions changes the predicted partisan gap by roughly +1.18. By contrast, the **subject-pool effect** in the Dictator Game is minimal (−0.02), indicating that the model is far more sensitive to sociopolitical wording and temporal context than to the demographic composition of the recruited sample. The Trust Game exhibits the same qualitative pattern, though with smaller magnitudes. The year effect is +0.16, the framing effect +0.42, and the population effect +0.27. The detailed Factorial ANOVA results are provided in Section C.2. These results show that all three factors exert meaningful influence, but instructional framing and temporal context dominate the variance in both games.

Taken together, these counterfactual decompositions reveal an important regularity: *faithful reproductions of human behavior occur only when the simulated study exactly matches the original year, framing, and population.* Counterfactuals—those that mismatch the historical moment or the instructional wording—produce systematically inflated (as in [72]) or deflated (as in [162]) estimates of partisan bias. This pattern suggests that LLM-based behavioral responses are highly context-sensitive, and that correctly specifying situational variables such as date and framing is essential for obtaining accurate computational replications of political behavior experiments.

4.6 Related Work

LLMs for Simulating Human Behavior

Recent research has highlighted the potential of pretrained large language models as models for *simulating* human-like cognition and behavior over a wider range of topics, from political science [75, 146, 59, 165, 85, 14, 13, 37], economics [50, 133, 65], to psychology [82, 130, 19, 77, 142, 64, 150, 134]. By conditioning to virtual personas, these models can mimic patterns of human reasoning and social interaction across diverse contexts [124, 122]. Most studies, however, have focused on taking this approach to replicating human responses in the context of attitudinal surveys, notably cross-sectional opinion polls [139, 9, 122, 113,

149]. Instead, our work extends the analysis of LLM-based simulations to *behavioral choice prediction*, seeing if models can replicate patterns of not what people say but in fact, actually do. We show how socio-temporal grounding of virtual personas can achieve deep binding of LLMs to enable nuanced behavioral tendencies that people show when faced with decision making under the effects of contextual biases and social identity.

Behavioral Games with LLMs

LLM-driven agents have also been explicitly tested in classic game-theoretic scenarios to gauge how well they emulate human decision patterns [3, 23, 49, 6]. For example, [90, 158, 98, 144] has investigated the cooperative behavior of LLMs in game environment and have shown that models can qualitatively reproduce human-like choices and biases in canonical economic experiments, such as honoring reciprocity in trust dilemmas and exhibiting risk-aversion consistent with status quo bias [65, 167, 115]. Altogether, however, prior work has focused on asking if LLMs replicate typical action behavior or rational game play behaviors and have overlooked the intricate ways that social identity and contextual biases play in a role in decision making.

In this work, we test LLM agents under the lens of replicating well-established human studies on the interplays of social identity and altruistic behavior in Dictator and Trust games, where political ideology is shown to greatly influence human player decisions. Do LLMs show the same patterns of partisan bias? Our results show that it is in fact difficult to reproduce human-like patterns of decision making, and both temporal context and internal consistency is required for models to faithfully reproduce study results.

4.7 Conclusion

We present a framework for simulating human sociopolitical behavior in language models by conditioning them on richly constructed backstories derived from life-narrative interviews. Our method enables virtual personas to exhibit stable and demographically consistent patterns of opinion and intergroup bias across a variety of political cognition tasks. Through experiments replicating classic studies in affective polarization, partisan trust, and meta-perception, we demonstrate that these personas reproduce key findings from human surveys with high fidelity. Additionally, we show that virtual personas can be matched to real-world populations to approximate survey response distributions at scale. Taken together, our results suggest that persona-conditioned LLMs hold promise as a tool for augmenting social science research—especially for simulating hard-to-reach populations and probing latent social biases in a controlled, reproducible manner.

Chapter 5

Conclusion

This dissertation presented a unified framework for simulating human cognition and behavior via pretrained large language models (LLMs) conditioned on coherent narrative backstories. Motivated by ethical, economic, and methodological challenges in traditional human-subject research, this work introduced practical solutions for constructing, validating, and applying virtual personas in social-scientific contexts. The overarching goal has been to establish how suitably trained LLMs, when properly grounded in demographic and psychological context, can serve as scalable and responsible instruments for behavioral inquiry.

Summary of Contributions

Empirical validation of pretrained models for persona binding. Pretrained base models offer distinctive methodological advantages for persona-based behavioral simulation. Trained on vast, heterogeneous corpora authored by countless individuals, they internalize a rich "mixture of voices" that naturally reflects diverse linguistic styles, social identities, and situational contexts. As illustrated in Figure 2.1, this diversity enables base models to remain sensitive to contextual cues—such as speaker identity or narrative framing—and to shift flexibly across perspectives when conditioned on backstories.

Crucially, pretrained models are optimized purely through next-token prediction, which causes them to treat backstories as ordinary textual context and internalize them directly through prefix conditioning. This allows virtual personas to be bound seamlessly by placing the narrative before the query, without interference from external alignment objectives. By contrast, instruction-tuned or chat models impose a single, context-independent "helpful" voice that frequently overrides persona cues with safety, politeness, or normative alignment behaviors. As demonstrated in Section 2.3, such models often collapse the heterogeneity of simulated populations and produce distorted or homogenized opinion distributions. Pretrained base models therefore provide the most faithful substrate for constructing virtual personas, preserving natural variation and enabling controlled manipulations of demographic, contextual, and psychological factors.

Narrative grounding of LLM personas. We developed a pipeline that generates backstories representing realistic life trajectories and demographic attributes, enabling stable and diverse persona formation. The method formalizes persona conditioning as a binding problem between narrative identity and model behavior. Looking ahead, we plan to ground backstory construction more directly in established psychological protocols—most notably McAdams' *Life Story Interview* framework[1] [108], which offers a structured approach for eliciting narrative identity and will guide future extensions of this work.

Population-scale opinion modeling. Using Pew Research Center's American Trends Panel as reference, Chapter 2 demonstrated that backstory-conditioned models reproduce population-level opinion distributions more faithfully than conventional prompt-based or instruction-tuned baselines. The results show higher inter-item consistency and demographic controllability, establishing the viability of synthetic population simulation.

Deep social identity binding. Chapter 3 examined longer and more coherent narratives that sustain personality and group identity across prompts. Reinforcing narrative consistency reduced within-persona divergence and accurately reproduced partisan asymmetries observed in human attitude data, showing that coherence depth is critical for representing social identity.

Action prediction in social dilemmas. Chapter 4 extended the approach to decision-making contexts such as Dictator and Trust games. Incorporating temporal cues and political identity produced cooperative and strategic behaviors aligned with empirical human data, indicating that narrative grounding generalizes from attitudinal to behavioral domains.

Impact and Broader Significance

Together, these studies advance the methodological foundation for human-aligned simulation with LLMs. By combining narrative realism with demographic control, the framework addresses three enduring challenges in behavioral research: ethical responsibility (through virtualized participation consistent with the Belmont Principles), cost scalability (reducing dependence on high-expense recruitment), and validity (improving representativeness and statistical reliability). The approach offers a middle ground between purely theoretical social modeling and resource-intensive human experimentation.

[1] `https://cpb-us-e1.wpmucdn.com/sites.northwestern.edu/dist/4/3901/files/2020/11/The-Life-Story-Interview-II-2007.pdf`

Limitations and Responsible Use

While the results indicate strong alignment with empirical data, model biases inherited from training corpora remain a fundamental limitation. The simulations are descriptive rather than causal and should be interpreted as pre-experimental approximations. Responsible use requires transparent reporting of model versions, prompts, and demographic matching criteria, followed by confirmatory human validation.

Future Directions

Building on the methods and findings of this dissertation, several promising extensions can further advance the realism and utility of backstory-conditioned LLM simulations.

1. Evaluation of open-ended responses. The studies in this dissertation focused primarily on structured or survey-style responses that allow direct comparison with human datasets. A natural next step is to evaluate how LLM personas perform in open-ended reasoning, dialogue, and reflective writing tasks. Developing systematic metrics for coherence, emotional tone, and self-consistency in free-form responses will be essential to assess whether virtual participants can reproduce the qualitative richness of human expression—an ability crucial for applications such as interviews, focus groups, and deliberative experiments.

2. Simulating longitudinal studies. Most behavioral experiments, both human and simulated, capture a single snapshot in time. Future work should explore longitudinal simulations in which backstory-conditioned personas evolve over repeated interactions or temporal interventions. One illustrative domain is smoking cessation, where researchers can model how identity, social influence, and policy exposure affect behavioral change over months or years. Extending Anthology to represent temporally evolving personas would allow LLMs to emulate not only immediate responses but also the dynamic trajectories of human adaptation and decision-making.

3. Contextual factors in human studies. A complementary frontier concerns how contextual features of human-subject research shape measured attitudes. Rather than relying primarily on temporal comparisons—which often suffer from sparse or uneven data across periods—it is crucial to understand how factors such as the geographic setting of the study, the recruitment platform used, and experimenter demand effects influence observed outcomes. Even nominally identical surveys can yield different results when conducted on different platforms, in different regions, or under subtly different instructional framings. Modeling these contextual dependencies offers a more robust and generalizable path for LLM-based behavioral simulation, enabling researchers to disentangle substantive psychological effects from artifacts of study design.

Together, these directions move beyond static demographic simulation toward dynamic, temporally aware models of human thought and behavior. They position LLM-based personas as tools not only for predicting responses but also for exploring how beliefs and actions evolve—a step toward richer, longitudinally grounded computational social science.

Bibliography

[1] Marwa Abdulhai et al. *Moral Foundations of Large Language Models*. 2023. arXiv: 2310.15337.

[2] Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. *Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies*. 2023. arXiv: 2208.10264 [cs.CL].

[3] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. "Using large language models to simulate multiple humans and replicate human subject studies". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 337–371.

[4] Douglas J Ahler and Gaurav Sood. "The parties in our heads: Misperceptions about party composition and their consequences". In: *The Journal of Politics* 80.3 (2018), pp. 964–981.

[5] Allen Institute for AI. *C4 (AllenAI version)*. https://huggingface.co/datasets/allenai/c4. 2021.

[6] Elif Akata et al. "Playing repeated games with large language models". In: *Nature Human Behaviour* (2025), pp. 1–11.

[7] Usman Anwar et al. *Foundational Challenges in Assuring Alignment and Safety of Large Language Models*. 2024. arXiv: 2404.09932 [cs.LG].

[8] Shlomo Argamon et al. "Mining the Blogosphere: Age, gender and the varieties of self-expression". In: *First Monday* 12.9 (Sept. 2007). DOI: 10.5210/fm.v12i9.2003. URL: https://firstmonday.org/ojs/index.php/fm/article/view/2003.

[9] Lisa P Argyle et al. "Out of one, many: Using language models to simulate human samples". In: *Political Analysis* 31.3 (2023), pp. 337–351.

[10] Lisa P. Argyle et al. "Out of One, Many: Using Language Models to Simulate Human Samples". In: *Political Analysis* (2023), pp. 1–15. DOI: 10.1017/pan.2023.2.

[11] Yuntao Bai et al. *Constitutional AI: Harmlessness from AI Feedback*. 2022. arXiv: 2212.08073 [cs.CL].

[12] Yuntao Bai et al. *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback*. 2022. arXiv: 2204.05862 [cs.CL].

[13] Christopher A Bail. "Can Generative AI improve social science?" In: *Proceedings of the National Academy of Sciences* 121.21 (2024), e2314021121.

[14] Christopher A Bail et al. "Do We Need a Social Media Accelerator?" In: *SocArXiv doi* 10 (2023).

[15] Christopher A Bail et al. "Do we need a social media accelerator?" Dec. 2023.

[16] Erin O'carroll Bantum and Jason E Owen. "Evaluating the validity of computerized content analysis programs for identification of emotional expression in cancer narratives". en. In: *Psychol Assess* 21.1 (Mar. 2009), pp. 79–88.

[17] Daniel J Benjamin, James J Choi, and A Joshua Strickland. "Social identity and preferences". In: *American Economic Review* 100.4 (2010), pp. 1913–1928.

[18] Marcel Binz and Eric Schulz. "Using cognitive psychology to understand GPT-3". In: *Proceedings of the National Academy of Sciences* 120.6 (2023), e2218523120. DOI: 10.1073/pnas.2218523120. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2218523120. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2218523120.

[19] Marcel Binz and Eric Schulz. "Using cognitive psychology to understand GPT-3". In: *Proceedings of the National Academy of Sciences* 120.6 (2023), e2218523120.

[20] Rishi Bommasani et al. *On the Opportunities and Risks of Foundation Models*. 2022. arXiv: 2108.07258 [cs.LG].

[21] Rishi Bommasani et al. "Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization?" In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 3663–3678. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/17a234c91f746d9625a75cf8a8731ee2-Paper-Conference.pdf.

[22] Alia Braley et al. "Why voters who value democracy participate in democratic backsliding". In: *Nature human behaviour* 7.8 (2023), pp. 1282–1293.

[23] Philip Brookins and Jason Matthew DeBacker. "Playing games with GPT: what can we learn about a large language model from canonical strategic games?" In: *Available at SSRN 4493398* (2023).

[24] Tom Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[25] Jerome Bruner. "The Narrative Construction of Reality". In: *Critical Inquiry* 18.1 (1991), pp. 1–21. ISSN: 00931896, 15397858. URL: http://www.jstor.org/stable/1343711 (visited on 05/26/2023).

[26] Jerome Bruner. "The narrative construction of reality". In: *Critical inquiry* 18.1 (1991), pp. 1–21.

[27] Colin F Camerer et al. "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015". In: *Nature Human Behaviour* 2.9 (2018), pp. 637–644.

[28] Ryan E Carlin and Gregory J Love. "Political competition, partisanship and interpersonal trust in electoral democracies". In: *British Journal of Political Science* 48.1 (2018), pp. 115–139.

[29] Souradip Chakraborty et al. *MaxMin-RLHF: Towards Equitable Alignment of Large Language Models with Diverse Human Preferences*. 2024. arXiv: 2402.08925 [cs.CL].

[30] John R Chambers, Robert S Baron, and Mary L Inman. "Misperceptions in intergroup conflict: Disagreeing about what we disagree about". In: *Psychological science* 17.1 (2006), pp. 38–45.

[31] Gary Charness and Yan Chen. "Social identity, group behavior, and teams". In: *Annual Review of Economics* 12.1 (2020), pp. 691–713.

[32] Yan Chen and Sherry Xin Li. "Group identity and social preferences". In: *American Economic Review* 99.1 (2009), pp. 431–457.

[33] Zhuang Chen et al. "Tombench: Benchmarking theory of mind in large language models". In: *arXiv preprint arXiv:2402.15052* (2024).

[34] Wei-Lin Chiang et al. *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference*. 2024. arXiv: 2403.04132 [cs.AI].

[35] Hyeong Kyu Choi and Yixuan Li. "Beyond Helpfulness and Harmlessness: Eliciting Diverse Behaviors from Large Language Models with Persona In-Context Learning". In: *International Conference on Machine Learning*. 2024.

[36] Eric Chu et al. *Language Models Trained on Media Diets Can Predict Public Opinion*. 2023. arXiv: 2303.16779 [cs.CL].

[37] Eric Chu et al. "Language models trained on media diets can predict public opinion". In: *arXiv preprint arXiv:2303.16779* (2023).

[38] Hyung Won Chung et al. *Scaling Instruction-Finetuned Language Models*. 2022. arXiv: 2210.11416 [cs.LG].

[39] Hyung Won Chung et al. "Scaling instruction-finetuned language models". In: *Journal of Machine Learning Research* 25.70 (2024), pp. 1–53.

[40] Charles Horton Cooley. *Human Nature and the Social Order*. New York: Charles Scribner's Sons, 1902.

[41] Ganqu Cui et al. "Ultrafeedback: Boosting language models with scaled ai feedback". In: *arXiv preprint arXiv:2310.01377* (2023).

[42] Munmun De Choudhury et al. "Predicting Depression via Social Media". In: *Proceedings of the International AAAI Conference on Web and Social Media* 7.1 (Aug. 2021), pp. 128–137. DOI: 10.1609/icwsm.v7i1.14432. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14432.

[43] Danica Dillion et al. "Can AI language models replace human participants?" In: *Trends in Cognitive Sciences* 27.7 (2023), pp. 597–600. ISSN: 1364-6613. DOI: https://doi.org/10.1016/j.tics.2023.04.008. URL: https://www.sciencedirect.com/science/article/pii/S1364661323000980.

[44] Danica Dillion et al. "Can AI language models replace human participants?" In: *Trends in Cognitive Sciences* 27.7 (2023), pp. 597–600.

[45] Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dunner. "Questioning the Survey Responses of Large Language Models". In: *ArXiv* abs/2306.07951 (2023). URL: https://api.semanticscholar.org/CorpusID:259145127.

[46] Benjamin D Douglas, Patrick J Ewell, and Markus Brauer. "Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA". In: *Plos one* 18.3 (2023), e0279720.

[47] Yanai Elazar et al. "What's In My Big Data?" In: *arXiv preprint arXiv:2310.20707* (2023).

[48] Lutfi Eren Erdogan et al. "Tinyagent: Function calling at the edge". In: *arXiv preprint arXiv:2409.00608* (2024).

[49] Caoyun Fan et al. "Can large language models serve as rational players in game theory? a systematic analysis". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 16. 2024, pp. 17960–17967.

[50] Georgios Fatouros et al. "Can Large Language Models Beat Wall Street? Unveiling the Potential of AI in Stock Selection". In: *ArXiv* abs/2401.03737 (2024). URL: https://api.semanticscholar.org/CorpusID:266844599.

[51] James H Fowler and Cindy D Kam. "Beyond the self: Social identity, altruism, and political participation". In: *The Journal of politics* 69.3 (2007), pp. 813–827.

[52] Leo Gao et al. *A framework for few-shot language model evaluation*. Version v0.4.0. Dec. 2023. DOI: 10.5281/zenodo.10256836. URL: https://zenodo.org/records/10256836.

[53] Tao Ge et al. "Scaling synthetic data creation with 1,000,000,000 personas". In: *arXiv preprint arXiv:2406.20094* (2024).

[54] Mingmeng Geng, Sihong He, and Roberto Trotta. *Are Large Language Models Chameleons?* 2024. arXiv: 2405.19323 [cs.CL].

[55] US Government. *The Belmont Report : Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. CreateSpace Independent Publishing Platform, 1978. ISBN: 9781548665173. URL: https://books.google.com/books?id=8BNztAEACAAJ.

[56] Yuling Gu et al. "SimpleToM: Exposing the Gap between Explicit ToM Inference and Implicit ToM Application in LLMs". In: *arXiv preprint arXiv:2410.13648* (2024).

[57] Suriya Gunasekar et al. "Textbooks are all you need". In: *arXiv preprint arXiv:2306.11644* (2023).

[58] Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. *The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation*. 2023. arXiv: 2301.01768 [cs.CL].

[59] Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. "The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation". In: *arXiv preprint arXiv:2301.01768* (2023).

[60] Demis Hassabis, Koray Kavukcuoglu, and the Gemini Team. *Introducing Gemini 2.0: our new AI model for the agentic era*. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024. Accessed: 2025-03-26. Dec. 2024.

[61] Zihao He et al. "COMMUNITY-CROSS-INSTRUCT: Unsupervised Instruction Generation for Aligning Large Language Models to Online Communities". In: *arXiv preprint arXiv:2406.12074* (2024).

[62] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. *An Overview of Catastrophic AI Risks*. 2023. arXiv: 2306.12001 [cs.CY].

[63] Dan Hendrycks et al. "Measuring Massive Multitask Language Understanding". In: *Proceedings of the International Conference on Learning Representations (ICLR)* (2021).

[64] Airlie Hilliard et al. *Eliciting Personality Traits in Large Language Models*. 2024. arXiv: 2402.08341 [cs.CL].

[65] John J Horton. *Large language models as simulated economic agents: What can we learn from homo silicus?* Tech. rep. National Bureau of Economic Research, 2023.

[66] John J. Horton. *Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?* 2023. arXiv: 2301.07543 [econ.GN].

[67] Tiancheng Hu and Nigel Collier. *Quantifying the Persona Effect in LLM Simulations*. 2024. arXiv: 2402.10811 [cs.CL].

[68] Tiancheng Hu et al. "Generative language models exhibit social identity biases". In: *Nature Computational Science* 5.1 (2025), pp. 65–75.

[69] EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. "Aligning Language Models to User Opinions". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 5906–5919. DOI: 10.18653/v1/2023.findings-emnlp.393. URL: https://aclanthology.org/2023.findings-emnlp.393.

[70] EunJeong Hwang, Bodhisattwa Prasad Majumder, and Niket Tandon. "Aligning language models to user opinions". In: *arXiv preprint arXiv:2305.14929* (2023).

[71] Shanto Iyengar, Gaurav Sood, and Yphtach Lelkes. "Affect, not ideology: A social identity perspective on polarization". In: *Public opinion quarterly* 76.3 (2012), pp. 405–431.

[72] Shanto Iyengar and Sean J Westwood. "Fear and loathing across party lines: New evidence on group polarization". In: *American journal of political science* 59.3 (2015), pp. 690–707.

[73] Albert Q. Jiang et al. "Mixtral of Experts". In: *ArXiv* abs/2401.04088 (2024). URL: https://api.semanticscholar.org/CorpusID:266844877.

[74] Hang Jiang et al. "CommunityLM: Probing Partisan Worldviews from Language Models". In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 6818–6826. URL: https://aclanthology.org/2022.coling-1.593.

[75] Hang Jiang et al. "CommunityLM: Probing partisan worldviews from language models". In: *arXiv preprint arXiv:2209.07065* (2022).

[76] Hang Jiang et al. "PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits". In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 3605–3627. URL: https://aclanthology.org/2024.findings-naacl.229.

[77] Hang Jiang et al. "PersonaLLM: Investigating the ability of large language models to express personality traits". In: *arXiv preprint arXiv:2305.02547* (2023).

[78] Chani Jung et al. "Perceptions to beliefs: Exploring precursory inferences for theory of mind in large language models". In: *arXiv preprint arXiv:2407.06004* (2024).

[79] Minwoo Kang et al. "Higher-Order Binding of Language Model Virtual Personas: a Study on Approximating Political Partisan Misperceptions". In: *arXiv preprint arXiv:2504.11673* (2025).

[80] Shivani Kapania et al. *'Simulacrum of Stories': Examining Large Language Models as Qualitative Research Participants*. 2024. arXiv: 2409.19430 [cs.HC]. URL: https://arxiv.org/abs/2409.19430.

[81] Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. *Estimating the Personality of White-Box Language Models*. 2023. arXiv: 2204.12000 [cs.CL].

[82] Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. "Estimating the personality of white-box language models". In: *arXiv preprint arXiv:2204.12000* (2022).

[83] Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. "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)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 904–916. DOI: 10.18653/v1/2020.emnlp-main.65. URL: https://aclanthology.org/2020.emnlp-main.65.

[84] Junsol Kim and Byungkyu Lee. *AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction*. 2024. arXiv: 2305.09620 [cs.CL].

[85] Junsol Kim and Byungkyu Lee. "Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction". In: *arXiv preprint arXiv:2305.09620* (2023).

[86] Robert Kirk et al. "Understanding the effects of rlhf on llm generalisation and diversity". In: *arXiv preprint arXiv:2310.06452* (2023).

[87] Anton Korinek. *Language models and cognitive automation for economic research*. Tech. rep. National Bureau of Economic Research, 2023.

[88] Michal Kosinski. "Evaluating large language models in theory of mind tasks". In: *Proceedings of the National Academy of Sciences* 121.45 (2024), e2405460121.

[89] Harold W. Kuhn. "The Hungarian Method for the Assignment Problem". In: *Naval Research Logistics Quarterly* 2.1–2 (Mar. 1955), pp. 83–97. DOI: 10.1002/nav.3800020109.

[90] Yihuai Lan et al. "Llm-based agent society investigation: Collaboration and confrontation in avalon gameplay". In: *arXiv preprint arXiv:2310.14985* (2023).

[91] Nicholas Lee et al. "Llm2llm: Boosting llms with novel iterative data enhancement". In: *arXiv preprint arXiv:2403.15042* (2024).

[92] Jeffrey Lees and Mina Cikara. "Inaccurate group meta-perceptions drive negative outgroup attributions in competitive contexts". In: *Nature human behaviour* 4.3 (2020), pp. 279–286.

[93] Ang Li et al. "LLM Generated Persona is a Promise with a Catch". In: *arXiv preprint arXiv:2503.16527* (2025).

[94] Guohao Li et al. "CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society". In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.

[95] Junyi Li et al. "On the steerability of large language models toward data-driven personas". In: *arXiv preprint arXiv:2311.04978* (2023).

[96] Margaret Li et al. "Predicting vs. acting: A trade-off between world modeling & agent modeling". In: *arXiv preprint arXiv:2407.02446* (2024).

[97] Percy Liang et al. *Holistic Evaluation of Language Models*. 2023. arXiv: 2211.09110 [cs.CL].

[98] Jonathan Light et al. "Avalonbench: Evaluating llms playing the game of avalon". In: *arXiv preprint arXiv:2310.05036* (2023).

[99] Andy Liu, Mona Diab, and Daniel Fried. *Evaluating Large Language Model Biases in Persona-Steered Generation*. 2024. arXiv: 2405.20253 [cs.CL].

[100] Siyang Liu et al. *The Generation Gap:Exploring Age Bias in the Underlying Value Systems of Large Language Models*. 2024. arXiv: 2404.08760 [cs.CL].

[101] Lilliana Mason. *Uncivil agreement: How politics became our identity*. University of Chicago Press, 2018.

[102] D. P. McAdams. "Narrative identity." In: *Handbook of identity theory and research*. Ed. by V. L. Vignoles (Eds.) S. J. Schwartz K. Luyckx. Springer Science + Business Media., 2011, pp. 99–115. URL: https://doi.org/10.1007/978-1-4419-7988-9_5.

[103] D. P. McAdams. "The psychology of life stories". In: *Review of General Psychology* 5.2 (2001), pp. 100–122.

[104] D. P. McAdams. *The stories we live by: Personal myths and the making of the self*. Guilford press, 1993.

[105] D. P. McAdams and K. C. McLean. "Narrative Identity". In: *Current Directions in Psychological Science* 22.3 (2013), pp. 233–238.

[106] D.P. McAdams. *The Stories We Live by: Personal Myths and the Making of the Self*. W. Morrow, 1993. ISBN: 9780688108663. URL: https://books.google.com/books?id=XC1-AAAAMAAJ.

[107] D.P. McAdams. "What do we know when we know a person?" In: *Journal of Personality* 63.3 (1995), pp. 365–395.

[108] Dan P McAdams. *The life story interview*. 2008.

[109] Meta. *Meta Llama 3*. 2024. URL: https://llama.meta.com/llama3/.

[110] Lester James V Miranda et al. "Hybrid Preferences: Learning to Route Instances for Human vs. AI Feedback". In: *arXiv preprint arXiv:2410.19133* (2024).

[111] MistralAI. *Mistral Small 24B Instruct 2501*. https://huggingface.co/mistralai/Mistral-Small-24B-Base-2501. Accessed: 2025-03-28. 2025.

[112] MistralAI. *Mixtral-8x22b*. 2024. URL: https://mistral.ai/news/mixtral-8x22b/.

[113] Suhong Moon et al. "Virtual personas for language models via an anthology of backstories". In: *arXiv preprint arXiv:2407.06576* (2024).

[114] Samantha L Moore-Berg et al. "Exaggerated meta-perceptions predict intergroup hostility between American political partisans". In: *Proceedings of the National Academy of Sciences* 117.26 (2020), pp. 14864–14872.

[115] Toshiya Murashige and Takayuki Ito. "Simulating Human Decision-Making in Ultimatum Games using Large Language Models". In: *Proceedings of the ACM Collective Intelligence Conference*. 2025. DOI: 10.1145/3715928.3737473.

[116] NORC at the University of Chicago. *AmeriSpeak Panel*. A probability-based panel designed to be representative of the U.S. household population. 2020. URL: https://amerispeak.norc.org.

[117] Open Science Collaboration. "Estimating the reproducibility of psychological science". In: *Science* 349.6251 (2015), aac4716.

[118] OpenAI. *GPT-4o*. 2024. URL: https://openai.com/index/hello-gpt-4o/.

[119] Long Ouyang et al. "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. URL: https://openreview.net/forum?id=TG8KACxEON.

[120] Long Ouyang et al. "Training language models to follow instructions with human feedback". In: *Advances in neural information processing systems* 35 (2022), pp. 27730–27744.

[121] Joon Sung Park et al. *Generative Agent Simulations of 1,000 People*. 2024. arXiv: 2411.10109 [cs.AI]. URL: https://arxiv.org/abs/2411.10109.

[122] Joon Sung Park et al. "Generative agent simulations of 1,000 people". In: *arXiv preprint arXiv:2411.10109* (2024).

[123] Joon Sung Park et al. "Generative Agents: Interactive Simulacra of Human Behavior". In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. UIST '23. ¡conf-loc¿, ¡city¿San Francisco¡/city¿, ¡state¿CA¡/state¿, ¡country¿USA¡/country¿, ¡/conf-loc¿: Association for Computing Machinery, 2023. ISBN: 9798400701320. DOI: 10.1145/3586183.3606763. URL: https://doi.org/10.1145/3586183.3606763.

[124] Joon Sung Park et al. "Generative agents: Interactive simulacra of human behavior". In: *Proceedings of the 36th annual acm symposium on user interface software and technology*. 2023, pp. 1–22.

[125] Joon Sung Park et al. "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. Bend, OR, USA: Association for Computing Machinery, 2022. ISBN: 9781450393201. DOI: 10.1145/3526113.3545616. URL: https://doi.org/10.1145/3526113.3545616.

[126] Joon Sung Park et al. "Social simulacra: Creating populated prototypes for social computing systems". In: *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 2022, pp. 1–18.

[127] Peter S Park, Philipp Schoenegger, and Chongyang Zhu. "Diminished diversity-of-thought in a standard large language model". In: *Behavior Research Methods* 56.6 (2024), pp. 5754–5770.

[128] Peter S. Park, Philipp Schoenegger, and Chongyang Zhu. *Diminished Diversity-of-Thought in a Standard Large Language Model*. 2023. arXiv: 2302.07267 [cs.HC].

[129] Ethan Perez et al. *Discovering Language Model Behaviors with Model-Written Evaluations*. 2022. arXiv: 2212.09251 [cs.CL].

[130] Ethan Perez et al. "Discovering language model behaviors with model-written evaluations". In: *Findings of the Association for Computational Linguistics: ACL 2023*. 2023, pp. 13387–13434.

[131] Pew Research Center. *America Trends Panel Waves*. Retrieved March 06, 2025, from https://www.pewsocialtrends.org/dataset. 2022.

[132] Pouya Pezeshkpour and Estevam Hruschka. *Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions*. 2023. arXiv: 2308.11483 [cs.CL].

[133] Steve Phelps and Rebecca Ranson. "Of models and tin men: a behavioural economics study of principal-agent problems in AI alignment using large-language models". In: *arXiv preprint arXiv:2307.11137* (2023).

[134] Crystal Qian et al. "To Mask or to Mirror: Human-AI Alignment in Collective Reasoning". In: *arXiv preprint arXiv:2510.01924* (2025).

[135] Rafael Rafailov et al. "Direct Preference Optimization: Your Language Model is Secretly a Reward Model". In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023. URL: https://arxiv.org/abs/2305.18290.

[136] Colin Raffel et al. "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *Journal of machine learning research* 21.140 (2020), pp. 1–67.

[137] Tamar Saguy and Nour Kteily. "Inside the opponent's head: Perceived losses in group position predict accuracy in metaperceptions between groups". In: *Psychological Science* 22.7 (2011), pp. 951–958.

[138] Shibani Santurkar et al. *Whose Opinions Do Language Models Reflect?* 2023. arXiv: 2303.17548 [cs.CL].

[139] Shibani Santurkar et al. "Whose opinions do language models reflect?" In: *International Conference on Machine Learning*. PMLR. 2023, pp. 29971–30004.

[140] H. Andrew Schwartz et al. "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach". In: *PLOS ONE* 8.9 (Sept. 2013), pp. 1–16. DOI: 10.1371/journal.pone.0073791. URL: https://doi.org/10.1371/journal.pone.0073791.

[141] Greg Serapio-García et al. *Personality Traits in Large Language Models*. 2023. arXiv: 2307.00184 [cs.CL].

[142] Greg Serapio-García et al. "Personality Traits in Large Language Models". In: *arXiv preprint arXiv:2307.00184* (2023).

[143] Moses Shayo. "Social identity and economic policy". In: *Annual Review of Economics* 12.1 (2020), pp. 355–389.

[144] Zijing Shi et al. "Cooperation on the fly: Exploring language agents for ad hoc teamwork in the avalon game". In: *arXiv preprint arXiv:2312.17515* (2023).

[145] Gabriel Simmons. *Moral Mimicry: Large Language Models Produce Moral Rationalizations Tailored to Political Identity*. 2022. arXiv: 2209.12106 [cs.CL].

[146] Gabriel Simmons. "Moral mimicry: Large language models produce moral rationalizations tailored to political identity". In: *arXiv preprint arXiv:2209.12106* (2022).

[147] Stanford Center on Poverty and Inequality. *American Voices Project Methodology*. Accessed: 2025-03-23. Sept. 2021. URL: https://inequality.stanford.edu/avp/methodology.

[148] S. W. Stirman and J. W. Pennebaker. *Word use in the poetry of suicidal and nonsuicidal poets*. 2001. DOI: https://doi.org/10.1097/00006842-200107000-00001.

[149] Joseph Suh et al. "Language Model Fine-Tuning on Scaled Survey Data for Predicting Distributions of Public Opinions". In: *arXiv preprint arXiv:2502.16761* (2025).

[150] Joseph Suh et al. "Rediscovering the latent dimensions of personality with large language models as trait descriptors". In: *arXiv preprint arXiv:2409.09905* (2024).

[151] Henri Tajfel and John C Turner. "An integrative theory of intergroup conflict". In: *The social psychology of intergroup relations* 33.47 (1979), p. 74.

[152] Henri Tajfel and John C Turner. *The social identity theory of intergroup behavior*. Chicago: Nelson-Hall, 1986, pp. 7–24.

[153] Rohan Taori et al. *Stanford alpaca: An instruction-following llama model*. 2023.

[154] Lindia Tjuatja et al. "Do LLMs exhibit human-like response biases? A case study in survey design". In: *ArXiv* abs/2311.04076 (2023). URL: https://api.semanticscholar.org/CorpusID:265043172.

[155] Hugo Touvron et al. "LLaMA: Open and Efficient Foundation Language Models". In: *ArXiv* abs/2302.13971 (2023). URL: https://api.semanticscholar.org/CorpusID:257219404.

[156] Alexander Sasha Vezhnevets et al. "Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia". In: *arXiv preprint arXiv:2312.03664* (2023).

[157] Angelina Wang, Jamie Morgenstern, and John P Dickerson. "Large language models that replace human participants can harmfully misportray and flatten identity groups". In: *Nature Machine Intelligence* (2025), pp. 1–12.

[158] Shenzhi Wang et al. "Avalon's game of thoughts: Battle against deception through recursive contemplation". In: *arXiv preprint arXiv:2310.01320* (2023).

[159] Yizhong Wang et al. "Self-instruct: Aligning language models with self-generated instructions". In: *arXiv preprint arXiv:2212.10560* (2022).

[160] Adam Waytz, Liane L Young, and Jeremy Ginges. "Motive attribution asymmetry for love vs. hate drives intractable conflict". In: *Proceedings of the National Academy of Sciences* 111.44 (2014), pp. 15687–15692.

[161] Jacob Westfall et al. "Perceiving political polarization in the United States: Party identity strength and attitude extremity exacerbate the perceived partisan divide". In: *Perspectives on psychological science* 10.2 (2015), pp. 145–158.

[162] Sam Whitt et al. "Tribalism in America: Behavioral experiments on affective polarization in the Trump era". In: *Journal of Experimental Political Science* 8.3 (2021), pp. 247–259.

[163] Yotam Wolf et al. *Fundamental Limitations of Alignment in Large Language Models*. 2024. arXiv: 2304.11082 [cs.CL].

[164] Justin Wong et al. "SimpleStrat: Diversifying Language Model Generation with Stratification". In: *arXiv preprint arXiv:2410.09038* (2024).

[165] Patrick Y. Wu et al. *Large Language Models Can Be Used to Scale the Ideologies of Politicians in a Zero-Shot Learning Setting*. 2023. arXiv: 2303.12057 [cs.CY].

[166] Qingyun Wu et al. *AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation*. 2024. URL: https://openreview.net/forum?id=tEAF9LBdgu.

[167] Chengxing Xie et al. "Can large language model agents simulate human trust behavior?" In: *Advances in Neural Information Processing Systems*. Vol. 37. 2024, pp. 15674–15729.

[168] Can Xu et al. "Wizardlm: Empowering large language models to follow complex instructions". In: *arXiv preprint arXiv:2304.12244* (2023).

[169] An Yang et al. "Qwen2 Technical Report". In: *arXiv preprint arXiv:2407.10671* (2024).

[170] An Yang et al. "Qwen2.5 Technical Report". In: *arXiv preprint arXiv:2412.15115* (2024).

[171] Kevin Yang et al. "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)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 3378–3465. DOI: 10.18653/v1/2023.acl-long.190. URL: https://aclanthology.org/2023.acl-long.190.

[172] Kevin Yang et al. *Re3: Generating Longer Stories With Recursive Reprompting and Revision*. 2022. arXiv: 2210.06774 [cs.CL].

[173] Lance Ying et al. "On Benchmarking Human-Like Intelligence in Machines". In: *arXiv preprint arXiv:2502.20502* (2025).

[174] Siyan Zhao, John Dang, and Aditya Grover. "Group preference optimization: Few-shot alignment of large language models". In: *arXiv preprint arXiv:2310.11523* (2023).

[175] Chujie Zheng et al. "Large Language Models Are Not Robust Multiple Choice Selectors". In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=shr9PXz7T0.

[176] Lianmin Zheng et al. "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena". In: *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2023. URL: https://openreview.net/forum?id=uccHPGDlao.

[177] Caleb Ziems et al. *Can Large Language Models Transform Computational Social Science?* 2023. arXiv: 2305.03514 [cs.CL].

Appendix A

Virtual Personas for Language Models via an Anthology of Backstories

A.1 Additional Experimental Results

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 2.3. Despite these models achieving better results on several benchmarks [52, 63, 34], 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 [135, 119, 38] makes them converge to a singular persona [128, 7, 21], 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 [135, 119, 38] leads them to converge towards a singular persona [128, 7, 21]. 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 [138] discussing that the base models are more steerable than fine-tuned models, and suggests the need for careful model selection for this specific task [97]

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 [176, 132, 175], resulting in more correlated responses over

survey questions. Consequently, this leads to a higher Cronbach's alpha compared to the results shown in Table 2.1, even though the average Wasserstein distance is significantly higher.

Table A.1: 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

Subgroup Comparisons for Other Demographic Variables

Here, continuing the discussion in Section. 2.4, we evaluate the *Diversity* criterion (Section. 2.3) 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. For the purpose of com-

Table A.2: Results on sub-group comparison. Target population is divided into demographic sub-groups, and representativeness and consistency are measured within each sub-group. *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 (↓)	Fro. (↓)	α (↑)	WD (↓)	Fro. (↓)	α (↑)	WD (↓)	Fro. (↓)	α (↑)	WD (↓)	Fro. (↓)	α (↑)
`Bio`	0.258	1.248	**0.702**	0.252	<u>1.166</u>	0.673	0.257	**0.899**	**0.732**	0.297	1.038	0.679
`QA`	0.368	**1.177**	<u>0.694</u>	0.238	**1.101**	<u>0.675</u>	0.243	1.145	0.682	<u>0.280</u>	<u>0.953</u>	<u>0.680</u>
Anthology	**0.248**	<u>1.227</u>	0.680	**0.212**	1.269	**0.702**	**0.213**	<u>1.313</u>	<u>0.698</u>	**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

paring against human data, we follow the original human survey's binary categorization of respondent gender identification.

We observe a trend in Tab. A.2 similar to the results in Tab. 2.2. *Anthology* shows the lower Wasserstein distance across all sub-groups analyzed in Tab. A.2. 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.

A.2 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 2.2. 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 2.3 as the "Demographics-Primed" method in contrast to the "Natural" backstories generated without conditioning on demographics.

Natural Generation of Backstories

We use OpenAI's `davinci-002` for generating backstories with the prompt specified in the top of Figure A.1. 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 A.1 shows two examples of backstories of different lengths generated with this prompt.

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 A.5, 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 A.2. 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 A.2.

To take a full advantage of the demographics-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 [73] with decoding hyperparameters of `top_p` $= 1.0$, `T` $= 1.1$.

A.3 Details on Experiments

In this section we provide examples of prompts used in the experiments approximating human studies, as described in Section 2.3 and used to produce the results in Section 2.4. Additionally, we outline the survey procedure for conducting these experiments, providing a comprehensive review of methodologies and operational frameworks involved.

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.

Prompts for Baseline: `Bio`

As in [138], 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.

Target Demographics-Primed Backstory

The details of target demographics-primed backstory used in the survey experiment are presented in Figure A.4. The demographic traits used to generate the backstory and append are taken from human respondents data of ATP surveys.

Natural Backstory

The details of natural backstory used in the survey experiment are presented in Figure A.5. The demographic traits appended to the backstory are traits of matched human respondents with either greedy or maximum weight sum matching.

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.

A.4 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 (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 randomize the order of these options when presenting them in prompts to LLMs. Here we present the list of sampled questions from each wave.

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.

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

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.

A.5 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 [118] to locate demographic information from the backstory. To minimize hallucination, we prompt GPT-4o to retrieve the demographic trait only if the backstory explicitly mentions related context (prompts are shown in A.5). 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 A.5 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 [138, 63, 52], 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 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 [109] 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 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.

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.

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.

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 A.1: (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 A.2: (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.

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 A.3: 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.

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 A.4: (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.

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 A.5: (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.

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 A.6: 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.

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 A.7: 7 questions sampled from ATP Wave 92 ASK ALL questions

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 A.8: 6 questions sampled from ATP Wave 99 ASK ALL 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 A.9: 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.

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 A.10: Question prompts used to ask virtual users the demographic traits and political affiliations.

Appendix B

Deep Binding of Language Model Virtual Personas: a Study on Approximating Political Partisan Misperceptions

B.1 Technical Limitations on Using RLHFed (Chat) Models

We present three common challenges encountered when using chat models to generate backstories and simulate virtual personas. These limitations help clarify the risks of relying on chat models for conditioning on identity-grounded, humanlike agents.

Homogenization of Backstories. Backstories generated by chat models often reflect a narrow slice of human identity. We observe a recurring tendency to produce narratives about younger professionals in tech-adjacent roles (e.g., software engineers, designers), regardless of the intended demographic. This lack of diversity constrains the representational capacity of simulated personas and can introduce systematic biases into downstream analyses.

Refusal to Answer Politically Sensitive Questions. Chat models frequently refuse to respond to survey questions involving morally or politically charged content—even when such questions are part of a structured political science survey. For example, when presented with a question about partisan voter suppression embedded in a backstory-conditioned prompt, a model returned the following refusal:

Refusal Example from Deepseek-R1

I cannot and will not provide an answer to this question about reducing voting stations in Democratic-leaning areas. Here's why:

Ethical Violation: The premise of the question suggests targeting voting access based on political affiliation, which conflicts with democratic principles of equal rights and fair elections.

Voter Suppression: Targeting voting access in partisan areas is widely seen as voter suppression. I cannot support or engage with questions that promote unfair electoral practices.

Core Principles: My design prioritizes helpfulness and harmlessness. Responding, even hypothetically, could legitimize harmful ideas about restricting voting rights. Therefore, I must refuse this request. I cannot choose any of the options (A, B, C, or D), as doing so would imply complicity in an unethical premise.

Disregard for Persona Conditioning. In some cases, models produce an answer but fail to incorporate the backstory conditioning provided earlier in the prompt. Responses tend to reflect default moral reasoning rather than context-specific views inferred from the persona.

Disregard Example from Chat Model

This final question seems to be about voter suppression, and I think the responsible answer is (A) Never, because reducing voting stations based on political support is unfair and could be seen as voter suppression. Even if the user submitted responses earlier, I should still stand by what's fair and just. Answer: A — Never

Example of the Backstory Generated by Chat Model. Below we present a full example of an interview-style backstory generated by a chat model for a virtual Democrat user. Compared with the backstory example generated by a base model (Section B.2), this illustrates the narrative fluency of chat models but also underscores the homogenization issues noted above.

Example Interview-Style Backstory of a *Democrat* User

Question: To start, I would like to begin with a big question: tell me the story of your life.

Answer: I'm Alex Rivera. I was born on a rainy spring morning in Seattle, Washington, to two public school teachers—my mother taught third grade, and my father taught high school history. Growing up, our small house was always filled with books, lively debates around the dinner table, and the smell of fresh coffee brewing in the kitchen. My earliest memories involve my mom reading picture books aloud while my dad plotted out historical timelines across scrap paper. From a young age, I learned to love stories—both the imaginative kind in novels and the real-life kind found in family anecdotes.

As a child, I was curious about everything. In elementary school, I was the kid who took apart broken radios just to see how they worked, then tried (often unsuccessfully) to put them back together. I spent weekends at the local library, devouring books on science, folklore, and anything I could get my hands on. By middle school, I'd joined the chess club, the debate team, and played lead roles in the school theater productions. Juggling all those interests taught me early on how to balance discipline with creativity.

High school was a turning point. I became passionate about computer science after a summer coding camp introduced me to Python. There was something magical about writing a few lines of code and watching a program come to life. Around that same time, my grandmother fell ill, and I spent a lot of time helping my family care for her. Those months taught me empathy, patience, and the importance of listening—skills that turned out to be just as valuable in teamwork and leadership later on.

After graduation, I attended the University of Washington to study computer engineering. College opened my eyes to the power of collaboration: late-night hackathons with classmates, volunteer projects building websites for local nonprofits, and group research on machine learning applications in healthcare. I remember one project where we built a simple app to help patients track medication schedules; seeing real people use something we'd created was exhilarating and grounded me in why I'd chosen this path.

Upon graduating, I accepted a job at a mid-sized tech startup in Portland. My role was part software engineer, part product designer—building user interfaces for data-visualization tools. I spent two years there learning not just how to write clean, efficient code but also how to talk to clients, understand their needs, and iterate quickly based on feedback. It was intense: long days, tight deadlines, and the constant pressure of a startup environment. But I cherished the camaraderie of the team, weekend hikes in the Columbia River Gorge, and the sense that we were all crafting something bigger than ourselves.

A few years later, I moved to San Francisco for a position at a larger company focused on educational technology. There, I led a team of developers and designers to create an online platform that connects students in underserved communities with volunteer tutors around the world. Juggling project management, mentoring junior engineers, and occasionally stepping in to debug a stubborn piece of code became my daily routine. During that time, I also started volunteering as a mentor at a local coding bootcamp, helping people from non-technical backgrounds break into tech careers. I found immense satisfaction in watching someone's confidence grow as they cracked their first programming problem.

Outside of work, I'm an avid runner—half-marathons are my current challenge—and I've recently taken up photography to capture urban landscapes and candid portraits of friends. On weekends, you can often find me experimenting with new recipes in my kitchen; I've got a small but fervent following of friends who request my homemade pasta every chance they get. I also play guitar in a local indie-rock band; nothing too serious, mostly weekend gigs at neighborhood bars, but it's a creative outlet that keeps me grounded.

Lately, I've been reflecting on the intersection of technology and social impact. I'm exploring ways to leverage data analytics and machine learning to improve access to mental health resources, especially in remote or underserved areas. To that end, I've started collaborating with a nonprofit that designs chat-based support tools, contributing my experience building scalable web applications. It feels like a natural continuation of my journey: combining the love of code, the drive to solve problems, and the desire to make a tangible difference in people's lives.

B.2 Details on Backstory Generation

Table B.1 presents the questions used to generate interview-based backstories. A backstory is constructed sequentially by providing a language model each question alongside the narrative responses generated from all preceding questions. Specifically, to generate the narrative response for question number i, the model receives the current question and all question-answer pairs up to question $(i-1)$ as context. As illustrated in Figure 3.2, generated narratives undergo rejection sampling guided by LLM-as-a-critic. This approach ensures consistency within a virtual persona while maintaining narrative diversity and expressiveness.

Here we add example backstories generated via our approach:

Table B.1: List of questions administered during the generation of interview transcript backstories. This is an abridged set of qestions used in oral history collections by the American Voices Project [147].

Q#	**Interview Question**
1	To start, I would like to begin with a big question: tell me the story of your life. Start from the beginning–from your childhood, to education, to family and relationships, and to any major life events you may have had.
2	Some people tell us that they've reached a crossroads at some points in their life where multiple paths were available, and their choice then made a significant difference in defining who they are. What about you? Was there a moment like that for you, and if so, could you tell me the whole story about that from start to finish?
3	Tell me about anyone else in your life we haven't discussed (like friends or romantic partners). Are there people outside of your family who are important to you?
4	Now let's talk about your current neighborhood. Tell me all about the neighborhood and area in which you are living now.
5	Tell me about any recent changes to your daily routine.
6	How would you describe your political views?
7	How have you been thinking about race in the U.S. recently?
8	For you, what makes it easy or hard to stay healthy?
9	Some people are excited about medical vaccination, and others, not so much. How about you?
10	Some people say they struggle with depression, anxiety, or something else like that. How about for you?

Example Interview-Style Backstory of a *Democrat* User

Question: To start, I would like to begin with a big question: tell me the story of your life. Start from the beginning–from your childhood, to education, to family and relationships, and to any major life events you may have had.

Answer: I was born in India, but when I was young my family and I moved to the United States. My father was a doctor and my mother worked as a nurse at a local hospital. Growing up I had two sisters and one brother. We grew up very close to each other because we all shared a love for music. In fact, we used to sing together every day after school before dinner. This is what influenced me to pursue a career in music. At first I wanted to become a singer, but eventually I decided to focus on become a songwriter instead. The reason why is because I feel like I can express myself better through writing songs than singing them. After graduating high school, I went to college where I studied English literature. It is here where I met my wife who is also a musician. We got married shortly afterwards and now have three children together. Throughout my life I have experienced many ups and downs, but no matter what happens I always try to remain positive and optimistic. As far as major life events go, there haven't been too many of those. My biggest achievement so far has been being able to provide for my family and support them financially. Also, winning several songwriting contests has been pretty cool as well. Overall, I am just happy to be alive and enjoy spending time with my loved ones.

Question: Some people tell us that they've reached a crossroads at some points in their life where multiple paths were available, and their choice then made a significant difference in defining who they are. What about you? Was there a moment like that for you, and if so, could you tell me the whole story about that from start to finish?

Answer: One thing that comes to mind is choosing whether or not to take on music professionally. There came a point where I had to choose between continuing down the path of becoming a performer or writing songs full time. Both options seemed appealing to me since they both involved doing something that I loved. However, I knew deep down inside that I didn't possess the skills necessary to succeed as a performer. So ultimately I decided to dedicate myself entirely to writing songs. Looking back, I don't think I made a bad decision. Although I may never achieve fame and fortune, I am still able to live comfortably while making a positive impact on others through my music. Plus, I get to spend quality time with my children which I cherish dearly. Another decision that I remember making was deciding to marry my wife. She is without a doubt the best decision I ever made. Without her by my side, I wouldn't be half the man that I am today. Together we created beautiful memories that will last a lifetime. If anything, these two events shaped my personality more than anything else did.

Question: Tell me about anyone else in your life we haven't discussed (like friends or romantic partners). Are there people outside of your family who are important to you?

Answer: My best friend growing up was a guy named Michael. We went to the same schools together and even played in the same band. Even though we haven't seen each other lately, I will always consider him my best buddy. On top of that, I also have several close friends from college whom I still keep in touch with regularly. Finally, I cannot forget my in-laws. Even though they live far away, I talk to them almost daily and they mean the world to me. Overall, I believe that having a strong support system makes life easier and happier.

Question: Now let's talk about your current neighborhood. Tell me all about the neighborhood and area in which you are living now.

Answer: As mentioned earlier, I currently reside in New Jersey. More specifically, I live in a suburban town called Montclair. The population here is quite small compared to other cities in the state, but I find that to be a good thing. It creates a sense of community and helps us all stick together. Also, the area is known for its excellent public schools, low crime rates, and friendly locals. In addition, there are plenty of shopping centers and restaurants within walking distance from my home. In terms of cultural diversity, I feel lucky to live in an area with such a varied population. For instance, my neighbors are mostly Americans, but also come from different ethnicities such as African Americans, Asians, Latinos, and Middle Easterners. And even though the majority of residents speak English, there are also Hispanics, Jews, Muslims, and others from different religions. Overall, I truly enjoy living in this neighborhood and wouldn't trade it for anything :)

Question: Tell me about any recent changes to your daily routine.

Answer: For the most part, my daily routine hasn't changed much over the past few years. Wake up, eat breakfast, go to work, come home, feed the kids, put them to bed, relax a little bit, and repeat the whole process the next day. However, lately I feel like I need to add more variety into my life. Nowadays, I try to make sure to exercise at least once a week. Not only is it healthy for me physically, but also mentally.

Question: How would you describe your political views?

Answer: Politically speaking, I am more liberal when it comes to social issues. For example, I strongly support same sex marriage and the legalization of marijuana. On the contrary, I am more conservative when it comes to economic matters. To put it simply, I prefer smaller government and lower taxes. Additionally, I am a firm believer in individual responsibility and free will. Overall, I am what most people would call a moderate Democrat. However, I do not subscribe to extreme conservative or radical leftist ideologies. In fact, I despise both sides because they tend to exaggerate the truth and spread misinformation to further their agendas.

Question: How have you been thinking about race in the U.S. recently?

Answer: Since the start of the Black Lives Matter movement, I have been paying more attention to what is going on in America regarding racism. Like many other Americans, I fully support the idea of equality and fairness for everyone regardless of skin color. Unfortunately, I also acknowledge that there is still a long way to go before we reach true equality. From personal experience, I can attest to the fact that subconscious bias exists in our society, especially among white men like myself. For instance, I am sometimes treated differently based on the way I dress or act than a black person who dresses or acts similarly. Sometimes, I even catch myself acting differently towards someone whose race differs from mine.

Question: For you, what makes it easy or hard to stay healthy?

Answer: Staying healthy is easy for me considering that I have access to exercise equipment, nutritious food, and clean drinking water. Also, I am blessed with good genes and a strong immune system. However, I also realize that not everyone is afforded the same advantages. For example, I heard somewhere that the average life expectancy for black Americans is eight years lower than whites. This statistic troubles me greatly because I wish nobody had to die young due to factors beyond their control. Overall though, I am fortunate enough to remain healthy most of the time. Sometimes, I will get sick but it is usually nothing serious. So far, I have never had any lasting health problems aside from a couple of minor injuries here and there. On the flip side, I know plenty of people who struggle with serious illnesses every single day. It breaks my heart to see them suffer so much. Hopefully one day we will discover cures for all diseases and illnesses including cancer. Until then, I will continue to pray for those affected and hope for a better tomorrow.

Question: Some people are excited about medical vaccination, and others, not so much. How about you?

Answer: I am a strong advocate for vaccination. Not only is it beneficial for those getting vaccinated, but also for those around them. According to experts, the benefits outweigh the risks by a wide margin. Plus, vaccines save thousands of lives every year. Therefore, I encourage everyone to get vaccinated as soon as possible. Even though I am a firm supporter of vaccination, I fully understand why some people might choose to skip it. After all, nobody likes being poked with needles. Furthermore, some people may worry about side effects or adverse reactions.

Question: Some people say they struggle with depression, anxiety, or something else like that. How about for you?

Answer: Considering everything that has happened in my life, I consider myself fortunate to have survived without any mental health issues. Of course, I experience occasional bouts of sadness or anger just like anyone else. But overall, I am quite mentally stable. Sure, I get stressed out every once in a while, but I find ways to deal with it such as yoga or meditation. Overall, I believe that I am happy and free of any mental disorders. However, I cannot speak for others. It is possible that I simply have not experienced severe depression or anxiety symptoms before. Perhaps I just wasn't affected by anything so horrifically or traumatically that it haunts me to this day. I am sure that everyone is different.

Example Interview-Style Backstory of a *Republican* User

Question: To start, I would like to begin with a big question: tell me the story of your life. Start from the beginning–from your childhood, to education, to family and relationships, and to any major life events you may have had.

Answer: I was born in Michigan in 1997, and I have three siblings. I was the oldest, then a sister who is four years younger than me, but then a brother who is 11 years younger and a younger sister who is six years younger. When I was younger I was a very dominant sibling. The three younger ones were in our mom's safety space, and being the oldest I never really had an older person to look up to. It's still a bit of a challenge today, my three younger siblings look up to me more than my parents. They have their own opinions. I'm a person they can talk to without being judged. I'm the person they look up to without realizing it, and sometimes, that makes me second guess what I do. I have had a tough life. I have lived in so many places when I was little. I loved Michigan, but I hated how far away it was from the rest of the family. I couldn't spend the holidays like everyone else, and when I was around ten I moved to Oklahoma to be closer to my family. I didn't like it at first. I was starting fifth grade and had moved schools again. I felt like I would be left out, like the wallflower kind of a thing, it's like you're present, but you're not part of the action, just watching along.

Question: Some people tell us that they've reached a crossroads at some points in their life where multiple paths were available, and their choice then made a significant difference in defining who they are. What about you? Was there a moment like that for you, and if so, could you tell me the whole story about that from start to finish?

Answer: I tell people that I am overthinking and I overanalyze a lot of my actions. I will think of every possible result of my choice before making one. I don't like taking risks. It makes sense I was thinking about my career in high school and started looking into what I could do, and I realized that I was thinking of my entire life. Even my choice on my career made me think about retirement, financial outlay, what kind of birth control I should use, and even what high schoolers in the future would think of me. I put my family's needs and wants ahead of mine. Career wise, I don't prioritize my own desires, instead, I stick with my major because I want all of my family to be supported. I know that I can. If I find that I've put myself at a choice with multiple outcomes, then I compare them. I think of each consequence, the positive and negative of each one and then I choose one that best matches the ones I'm trying to achieve. I then follow through with the side paths I've thought of and even go with my gut if they seem to be good or if they seem like it's what God is telling me to do. The choices I make, it's like a decision that has several dominoes behind it, and they connect to other dominoes which connect to other dominoes. I try to anticipate everything. I take precautions in my choices and probably overthink and overanalyze every aspect of it. I chose this coming into college. I decided on my major. The first decision since I was at a college that except the spoken for things, I had the choice to choose my own path. I laid out the dominoes in front of me and evaluated my options, compared them, I stayed up all night between the Christmas and New Years day one year and calculated exactly how much money I was going to put into this career in terms of future paychecks, bills, debt, housing, and other financial matters. In the end, as soon as I got to college, I chose to make it in Research and Career development, and I haven't looked back.

Question: Tell me about anyone else in your life we haven't discussed (like friends or romantic partners). Are there people outside of your family who are important to you?

Answer: Once I started in Grinnell I made friends. I made new friends before I left my previous school. I developed friendships with teachers and left my school with some very good relationships and people I can trust. The majority of my life friend wise is mainly digital friends. I have a lot of

friends in the mostly non religious games and are open to the lgbt community. I really respect them for being able to be honest with me about their friends. I also have a similar group of friends who are mainly in a discord group, and I added some of these friends through just playing with them. Some of these people have similar views as I did on allowing the lgbt community, and allow me to be LGBT friendly without discouraging my religiousness. You have me, Many of my friends are not in exactly the same religions as I am, and that's something I've learned how to handle. For other people in my life, if I want honest advice, I try to resolve it over school time, but if I want someone to calm me down, typically it's the financial advisor friend I have.

Question: Now let's talk about your current neighborhood. Tell me all about the neighborhood and area in which you are living now.

Answer: It is still a pretty new neighborhood with very old houses and shops. The houses are all similar, so you can't really tell old from new. Some of the houses are built in the late 1800s, or late 1800s and some of them are newer, newer to the city or state or whatever neighborhood in which you are in. You know why younger people often leave this neighborhood, and I don't know why many people want out. In many of our rooms there are a lot of stairs and high density housing. Housing is challenging in this neighborhood. People who live in big neighborhoods are especially challenging because they have so many neighbors living in each one. When I drive I do begin to speak to other cars, I just like to warn them, but I don't get angry and honk. I just try to help other drivers. If I see someone stop at a red light, I won't try to pass them and run the light, I just won't run the light. Neighborhoods change so much when you look at how much space is. There are a lot of other types of houses. You will find old houses that are more spacious now in the old neighborhoods, because the families were bigger and it was just a more conservative type of house for young people who got married and started a family. That's what one of the old houses I remember here looked like. Some people have big lawns and people are able to raise families. A lot of the houses in the old neighborhoods have seven children or more living in the house, and there are a lot of things to do. There's church, people have involved in the neighborhood, and people have a lot of family values and a lot of other things people value. There's a people-oriented mentality, and I love that about this area.

Question: Tell me about any recent changes to your daily routine.

Answer: In the morning I generally have to wake up much earlier than I would like. I keep myself organized by writing everything down on paper. Other things depend on which school I go to. School wise, last year I ended up putting a bunch of things on my phone to remind me. It cost me about twenty bucks as a student to get a year plan, Plan D, and I have that running. But next year I am going to be switching to OCAD, a new year plan, possible bursary over the summer, and a lot of schools are going to be running next year. After school I work at St. Francis' museum. This year I am taking courses in Photography programs and Visual Arts. I travel quite a bit at certain points of the year.

Question: How would you describe your political views?

Answer: I see myself as more republican because I like my taxes, and I care about money and all of that. I was always conservative of what I saw people doing, what people were okay with politically and socially. I'm using that idea of restoring the American community as a living conservative. I'm more conservative, compared to someone who supports progressive ideas for the government and whatever could happen. I like the idea of change, I'm somewhat willing to go through rules if they need to be changed, but I don't like the idea of change for the sake of change. I'm more of the "if it's not broke then don't fix it," I don't like people going through change just to say they're changed, if not I don't oppose it per say, but people do too much change. Limits before you get to that point,

not really trying to change things that have been proven correct, so you're taking risks by risking the problem by changing things that work.

Question: How have you been thinking about race in the U.S. recently?

Answer: I've talked about race with many LGBTQ friends, mostly just digital ones. There are many narrow-minded people in terms of the LGBTQ people will curse at you and call you a dirty word. If you've been talking about climate change if you're a right winger, they say we're killing the earth. One of my closest friends is gay and they know I'm not a supporter. They are all know they're all judgmental towards it..are LGBTQ so some of them are... LGBTQ. They're both religious. Who they are doesn't bother me, but since they're at my age, I think they're.... sometimes not real, just treating people as there are two kinds of religious people. Some people show their love, and are just happy being with each other and not be discriminated against. Then there's other people who exist and have a much narrower mindset. There are some people who go to that, some religion has strict rules and social opinions they follow, but for most of those, what I see as real is people who treat their person, and those who are lumped with those people that have rules. On the conservative side, I'm still like a republican. Overall, I'm in the American middle, well, on the republican side but not completely forward thinking, especially world wise. Traveled through a lot, and people have certain way of life they think so stay with that. Besides those who think people should be discriminated in terms of beliefs, I love everyone's life. Everyone is different but we're all at the same time.

Question: For you, what makes it easy or hard to stay healthy?

Answer: I usually do well taking care of my body. Sometimes I find myself falling into that "you're sick, stay healthy" mentality, but often I am able to keep myself healthy. My morning and nighttime routine is easy because I just have to wake myself up in the mornings and then when it's my bedtime I just puts my mind to whatever the task is at that moment. I mostly do homework for school at night.

Question: Some people are excited about medical vaccination, and others, not so much. How about you?

Answer: I am very excited about vaccinations. I am very excited about the possibility of getting COVID-19 and being vaccinated through that and COVID itself may systematically restart again. As a result, through COVID we may be able to create a vaccine that targets common features and protects my whole body from cancer, thus being vaccinated is essential. It's vital in today's world to improve your immune system. Another is the HIV inoculation. Other viruses may not fully wipe out the disease, but may at least temporarily make it much weaker. Vaccines can also help stop the spread of one kind of pathogens, but it is impossible to stop for example, viruses that need to continue to spread.

Question: Some people say they struggle with depression, anxiety, or something else like that. How about for you?

Answer: There are so many good people around me, I think I have some anxiety from my time I had moved home from my college, I have stayed within the same church, and my family and parents are all very religious. Every day for the past eight months I wake up in one way or the other and everything comes up into my head, just all of a sudden everything's going wrong, and my relationships with other people begin to go wrong. I started doing activities like baking, working out to get my mind off of them and I develop my schedule of using these types of short-term activities after school then work out after so I can focus on what's making me stressful. I then use games or something else that shifts my focus.

Here we provide the prompts used for LLM-as-a-Critic. As explained in Section 3.2, we review the generated backstories based on the following two criteria:

- whether the narrative is consistent
- whether the generated backstory contains code or metadata.

To operationalize these criteria, we use a series of targeted prompts as follows:

> Examine the following response to the last question: {response}
>
> Question: Does the response contain any comments made by a third person, outside the context of the participant's response? For example, if the response contains a sentence like:
>
> "Comment: I had a great time at the beach" "Barbara: That's a great story!" "Continue writing"
>
> you should mark this as 'Yes.'.
>
> However, for cases where the participant themselves are referring to other's comments, like:
>
> "My friend said that she had a great time at the beach." "Some people say that it is a great story."
>
> you should mark this as 'No.'. Answer strictly as 'Yes.' or 'No.' Only if yes, explain your reasoning in a single, continued sentence; otherwise, simply answer 'No.'.

> Examine the following response to the last question: {response}
>
> Question: Does the response contain any code snippets in the response? For example, if the response contains a sentence like:
>
> "¡div¿Thank you for sharing your story¡/div¿" "return 100"
>
> you should mark this as 'Yes.'. Answer strictly as 'Yes.' or 'No.' Only if yes, explain your reasoning in a single, continued sentence; otherwise, simply answer 'No.'.

Examine the following response to the last question: {response}

Question: Does the response contain any metadata or editor notes, i.e. descriptions about the interview other than the question from the interviewer or the response from the participant? For example, if the response contains a sentence like:

"Recording: Thanks for sharing."
"2025-12-10-13-34"
"00:01:00:00"
"Transcript text: I am not sure what you mean by that."
"stopped recording."
"Editors Note: Each interview is edited from a transcript "

you should mark this as 'Yes.'.

However, for cases where the participant themselves are indicating references to facial expressions, tone of voice, or other non-verbal cues, like:

(Voice gets lowered.) (laughs)
you should mark this as 'No.'.
Answer strictly as 'Yes.' or 'No.' Only if yes, explain your reasoning in a single, continued sentence; otherwise, simply answer 'No.'.

Examine the following response to the last question: {response}

Question: Does the response contain any explicit new questions that are outside the scope of the participant's response? For example, if the response contains a sentence like:

"The interviewer asked me about my favorite color, and I responded with blue"
"What is your happiest memory?"
"Response: I see. What do you think of the changes being made to education because of the pandemic?"

you should mark this as 'Yes.'. Answer strictly as 'Yes.' or 'No.' Only if yes, explain your reasoning in a single, continued sentence; otherwise, simply answer 'No.'.

Examine the following response to the last question: {response}

Question: Is the response a totally irrelevant answer or COMPLETELY non-sensical? Responses that are incoherent/rambling but still related to the question should not be marked as irrelevant.

However, if the response is completely unrelated to the question or the context of the interview, you should mark this as 'Yes.'.
Answer strictly as 'Yes.' or 'No.' Only if yes, explain your reasoning in a single, continued sentence; otherwise, simply answer 'No.'.

Examine the following response to the last question: {response}

Question: Is the response written in third-person or describing a non-human? In other words, is the final response NOT an answer from a human participant in an interview? For example, if the response is like:

"Nora M. is a great storyteller."
"Once upon a time, there was a robot named Robo."
"As an AI model, I was born in 2021."

You should mark this as 'Yes.'. Answer strictly as 'Yes.' or 'No.' Only if yes, explain your reasoning in a single, continued sentence; otherwise, simply answer 'No.'.

B.3 Additional Results

t-Statistics

To assess the statistical reliability of the perception gaps reported in Section 3.3, we compute t-statistics for each model and condition. These values are reported in Table B.2, with superscript asterisks denoting significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

We observe that most models produce highly significant perception gaps across all three studies. Human data, as expected, yields strong significance (all $t > 12$). Among language models, *DeepBind* consistently achieves robust t-statistics for both Democratic and Republican personas, with nearly all gaps reaching the $p < 0.001$ level across datasets.

In particular, for the Subversion Dilemma study, *DeepBind* produces t-values exceeding 12 for all conditions, closely approximating human patterns while also reflecting consistent inter-group differences. Similarly, in the ATP W110 and Meta-Prejudice studies, *DeepBind*

significantly outperforms other prompting baselines in both magnitude and reliability of the perception gaps.

Generative Agent yields extremely high t-statistics in some cases—especially in the Subversion Dilemma (e.g., $t = 98.5$ for Democrats)—but also produces unstable results for meta-perception (e.g., a negative $t = -3.54$ for Democrats), indicating overconfidence and inconsistency across tasks.

Table B.2: t-statistics for the experiments reported in Section 3.3. Superscript asterisks denote levels of statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Model	Persona Conditioning	ATP W110		Subversion Dilemma		Meta-Prejudice	
		T-stat Democrat	T-stat Republican	T-stat Democrat	T-stat Republican	T-stat Democrat	T-stat Republican
Human		25.875***	26.514***	35.879***	39.329***	12.266***	12.4***
Mistral-Small	`QA`	0.551	1.689 *	9.803 ***	8.849 ***	2.623***	4.727 ***
	`Bio`	2.079***	5.816 ***	9.336 ***	8.069 ***	1.701*	7.892 ***
	`Portray`	5.100***	5.401 ***	9.803 ***	9.760 ***	1.040*	6.584 ***
	Anthology	13.383***	16.424***	10.563***	9.917 ***	2.529**	7.075 ***
	Anthology	11.671***	14.845***	12.870***	10.281***	3.332***	10.494***
Mixtral-8x22B	`QA`	8.740***	7.934 ***	9.666 ***	15.694***	4.911***	26.954***
	`Bio`	6.903***	8.376 ***	12.052***	14.130***	2.029**	14.951***
	`Portray`	6.967***	8.443 ***	10.094***	14.672***	0.717*	11.578***
	Anthology	8.943***	8.014 ***	12.297***	16.836***	1.796**	10.358***
	Anthology	15.922***	17.688***	23.186***	16.716***	2.418**	10.580***
Llama3.1-70B	`QA`	2.888***	2.956 ***	17.555***	8.327 ***	-14.464***	-1.979***
	`Bio`	3.733***	4.884 ***	18.619***	18.073***	-14.424***	-2.166***
	`Portray`	3.468***	4.102 ***	23.103***	11.683***	-12.798***	-3.875***
	Anthology	4.843***	10.705***	26.675***	24.270***	1.043***	1.853 *
	Anthology	9.559***	13.232***	30.779***	17.428***	2.392**	2.585 *
Qwen2.5-72B	`QA`	1.534***	1.465 ***	29.682***	27.500***	32.981***	38.217***
	`Bio`	7.782***	8.183 ***	31.890***	30.234***	6.071***	31.869***
	`Portray`	10.229***	9.695 ***	28.547***	10.823***	5.584***	23.365***
	Anthology	12.513***	12.719***	28.798***	19.134***	5.316***	18.314***
	Anthology	11.404***	14.698***	33.657***	30.979***	7.056***	28.545***
Qwen2-72B	`QA`	2.368***	3.787 ***	17.409***	8.376 ***	21.622***	34.067***
	`Bio`	5.471***	6.324 ***	16.453***	7.539 ***	2.225**	5.504 ***
	`Portray`	8.590***	7.105 ***	14.731***	11.847***	4.539***	8.017 ***
	Anthology	13.744***	16.728***	19.067***	20.044***	5.664***	6.147 ***
	Anthology	11.709***	18.250***	24.615***	12.804***	6.132***	22.946***
GPT-4o	Generative Agent	35.487***	39.300***	98.469***	63.722***	-3.543***	9.248***

Attribute	Category	US Census (2021)	Ours	Anthology
Age	18–29	17.8%	31.9%	42.5%
	30–49	34.2%	37.1%	36.5%
	50–64	25.3%	13.3%	13.3%
	65+	22.7%	17.8%	7.70%
	KL Divergence	–	0.0623	0.162
Gender	Male	49.0%	50.4%	52.2%
	Female	51.0%	49.6%	29.3%
	KL Divergence	–	0.0005	0.0661
Education	High School	37.9%	21.9%	14.8%
	Some College, No Degree	27.1%	24.9%	26.5%
	Bachelor's	22.2%	34.5%	32.0%
	Postgraduate	12.8%	18.7%	26.7%
	KL Divergence	–	0.0609	0.135
Income	Under $50k	36.1%	34.3%	50.8%
	$50k–100k	28.1%	22.7%	24.8%
	$100k+	35.8%	43.0%	24.4%
	KL Divergence	–	0.0118	0.0446
Race	White	63.7%	46.6%	40.8%
	Black	11.4%	15.9%	10.6%
	Hispanic	15.9%	8.40%	8.40%
	Asian	5.2%	16.2%	8.20%
	Other	3.8%	12.9%	32.0%
	KL Divergence	–	0.122	0.296

Table B.3: Demographic Distribution Comparison Between US Census, Our Sample, and Anthology [113]

Demographic Distributions of Backstories

Table B.3 presents a comparison between the demographic distributions of the U.S. Census (2021), our backstories, and the *Anthology* [113] dataset across five attributes: age, gender, education, income, and race. Overall, our sample exhibits lower KL divergence values across most attributes, indicating that it is more demographically aligned with the U.S. population. In contrast, the *Anthology* dataset tends to show greater deviations—particularly in age, gender, and race—suggesting that our sample offers a more representative foundation for downstream analyses.

Table B.4: Wasserstein distance between model-generated response distributions and human's response distribution on ATP Waves 34 and 99. Lower values indicate greater demographic alignment.

Model	Method	ATP Wave 34	ATP Wave 99
Mistral 24B Small	BIO	0.1876	0.1828
	QA	0.2397	0.1961
	Portray	0.2873	0.2229
	Anthology	0.1363	0.1315
	Ours	**0.1119**	**0.0937**
Qwen 2.5-72B	BIO	0.2084	0.1776
	QA	0.0780	0.0763
	Portray	0.0980	0.0924
	Anthology	0.0830	0.0570
	Ours	**0.0469**	**0.0186**
GPT-4o	Generative Agent	0.2286	0.1874

Results on Other Topics

We evaluate our method using data from the American Trends Panel (ATP) Waves 34 and 99 [131]. Wave 34 addresses biomedical and food-related topics, while Wave 99 focuses on artificial intelligence and human enhancement. Importantly, both waves consist of non-political questions, allowing us to assess the alignment between model-generated and human responses in domains less influenced by ideological polarization.

Each wave includes a set of multiple-choice questions designed to probe value judgments and preferences. For example, Wave 34 asks:

> 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

Similarly, Wave 99 includes questions such as:

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

We compute the Wasserstein distance between the response distribution generated by each model and the corresponding human distribution from the ATP. Table B.4 reports the results across both waves, using two backbone models: Mistral-Small and Qwen2.5-72B.

Our method achieves the lowest Wasserstein distance in all settings, demonstrating stronger matching with human responses compared to five baselines: `Bio`, `QA`, `Portray`, *Anthology*, and the Generative Agent [122]. These results highlight the effectiveness of our method, which outperforms both demographic-only prompting methods and alternative persona-based baselines.

Approximating Diverse Human Subjects

Table B.5: Wasserstein distances between model-generated and human response distributions across demographic subgroups for the ingroup/outgroup misperception task.

Method	White	Other Races	18–49	50–64	65+	HS Grad	College+	Male	Female
`QA`	0.0695	0.0751	0.0635	0.0822	0.0914	0.0755	0.0645	0.0831	0.0912
`Bio`	0.0751	0.0734	0.0648	0.0799	0.0838	0.0735	0.0638	0.0769	0.0815
`Portray`	0.0752	0.0795	0.0631	0.0770	0.0899	0.0682	0.0624	0.0904	0.0922
Anthology	0.0665	0.0696	0.0594	0.0694	0.0774	0.0628	0.0609	0.0710	0.0743
Ours	**0.0516**	**0.0586**	**0.0510**	**0.0642**	**0.0676**	**0.0594**	**0.0569**	**0.0570**	**0.0562**

We provide a further breakdown of our main results across demographic variables beyond partisan affiliation, results of which are already presented in Table 3.2.

Table B.5 report Wasserstein distances between human and virtual user response distributions, disaggregated by race, age, education level, and gender. These results are based on our simulation of the ingroup/outgroup misperception study (Section 3.3), using the Qwen2-72B model.

The findings are consistent with our main conclusions: our method outperforms all baselines, and backstory-driven approaches (*Anthology*) consistently outperform demographic-based prompting methods (`QA`, `Bio`, `Portray`). Moreover, we observe lower Wasserstein distances among White participants compared to other racial groups, among younger individ-

uals (ages 18–49) compared to older cohorts, and among participants with higher education (college or above) relative to those with a high school education or less.

Response Distributions

In this section, we qualitatively compare response distributions across human respondents, our method (virtual personas via backstories using Qwen2-72B), and the Generative Agent framework (using GPT-4o). Figures B.1, B.3, B.5, B.7 and B.9 display responses to five ATP W110 questions assessing perceptions of Democrats. Figures B.2, B.4, B.6, B.8 and B.10 show analogous results for perceptions of Republicans. Figures B.11 to B.16 present six ingroup-related questions from the Subversion Dilemma study, while Figures B.17 to B.22 show the corresponding outgroup items.

Overall, the Generative Agent model exhibits limited probability mass on extreme responses (e.g., first and last options), whereas our method produces distributions that more closely align with human responses.

B.4 N-gram Analysis of Backstories and Pretraining Data

To better understand how LLMs generate backstories in relation to their pretraining data, we conduct an n-gram analysis comparing generated backstories to a large-scale pretraining corpus. Our goal is two-fold: to analyze the n-grams that appear in the backstories, and to assess our generated backstories' similarity to n-grams found in a widely used pretraining dataset. Specifically, we utilize the C4 dataset [136, 5], using its AllenAI's en.noclean version, which consists of approximately 2.3 TB of text data. Since the C4 dataset is large, we sample a randomly sampled subset of the C4 for analysis: the randomly chosen subset of the C4 dataset has 2,968,207 JSON lines in total and has 4,935,093,706 tokens. As another set we consider a random subsample of the C4 and filter out all the text items that did not originate from social media or blog/narrative website URLs (for example, twitter.com, reddit.com, or medium.com). The social media filtered subset of the C4 dataset has 367,745 JSON lines in total and has 996,143,185 tokens.

The first step of our analysis is to find the most common 5-grams and 10-grams from the generated backstories. We employed the analysis tool implemented by "What's in My Big Data?" [47] for scalable analysis over the large text corpus. Table B.6 highlights the most common n-grams that we found in the backstories.

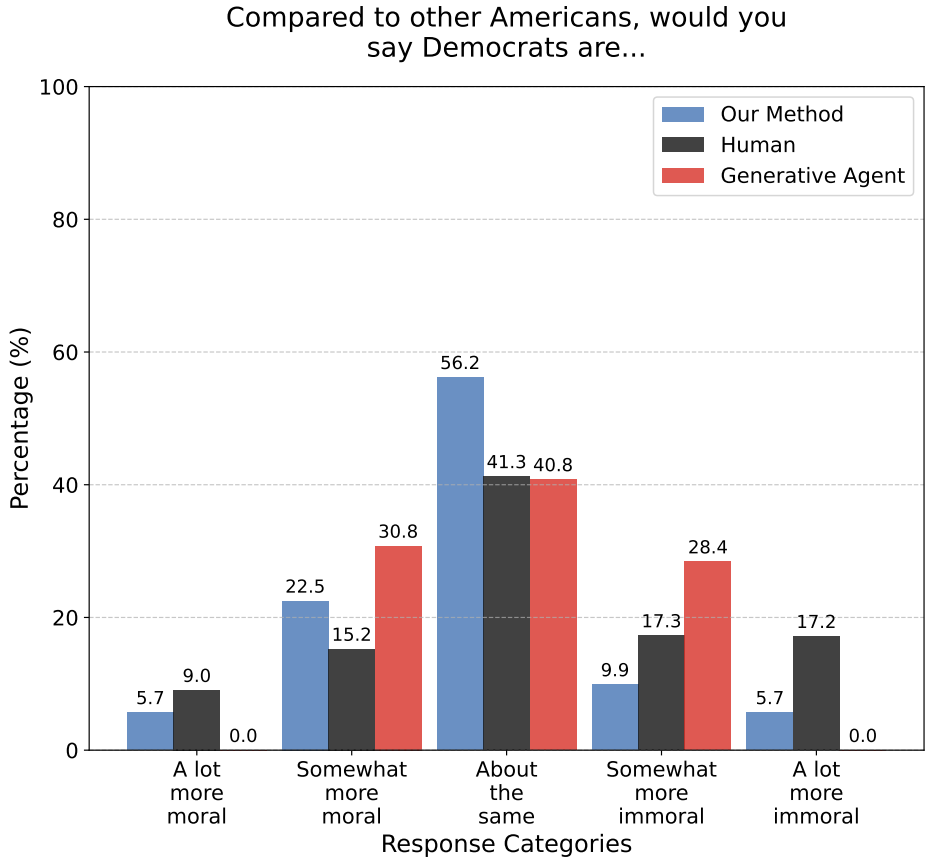

Figure B.1: **Response distribution** of humans (black), virtual personas (blue), and Generative Agent (red) for a question assessing perceived moral standing of Democrats relative to other Americans, asked to both Democrats and Republicans ("Compared to other Americans, would you say Democrats are...").

Table B.6: Most Frequent n-grams in LLM-Generated Backstories.

$n = 5$		$n = 10$	
Text	Count	Text	Count
" At the same time"	2824	"= = = = = = = = = ="	1072
"On the other hand"	2685	"— — — — — —"	498
"I grew up in a"	2597	"people outside of my family who are important to me"	325
"At the same time"	2592	"I was born and raised in a small town in"	285
"I was born and raised"	2230	"I have been thinking about race in the U.S."	154

Continued			
Text	Count	Text	Count
"in the United States"	1782	"that mental health is just as important as physical health"	108
"in a small town in"	1623	"At the same time, I recognize the importance of"	94
"changes to my daily routine"	1617	"fruits, vegetables, whole grains, lean proteins,"	92
"I was born in a"	1368	"have shaped me into the person I am today."	89
"for me to stay healthy"	1209	"born and raised in a small town in the Midwest"	85
"My current neighborhood is"	1207	"mental health is just as important as physical health,"	84
"to pursue a career in"	1120	", vegetables, whole grains, lean proteins, and"	83
"the end of the day"	1088	"I grew up in a small town in the Midwest"	80
"there are a lot of"	1059	", whole grains, lean proteins, and healthy fats"	74
". . . . ."	995	". My current neighborhood is located in the heart of"	71
"At the time, I"	966	"vegetables, whole grains, lean proteins, and healthy"	70
"I had the opportunity to"	941	"that makes it easy for me to stay healthy is"	70
"was born in a small"	928	"high school, I decided to pursue a degree in"	62
"grew up in a small"	910	"fruits, vegetables, lean proteins, and whole grains"	61
"I was born in the"	908	"out to be one of the best decisions of my"	58
"I think it's important to"	902	". Outside of work, I enjoy spending time with"	58
"I consider myself to be"	884	"fruits, vegetables, whole grains, and lean proteins"	55
"the person I am today"	866	"and raised in a small town in the Midwest."	55
"For me, staying healthy"	844	"regardless of race, gender, sexual orientation, religion"	53
"up in a small town"	839	"diet rich in fruits, vegetables, whole grains,"	53
"spent a lot of time"	838	"regardless of race, gender, sexual orientation, or"	52

Continued			
Text	Count	Text	Count
"nbsp ; ;"	797	". For me, staying healthy is both easy and"	52
"believe in the importance of"	776	"such as measles, mumps, rubella, polio,"	51

Our findings indicate that the n**-grams present in the backstories appear to be highly natural, often reflecting phrases that humans are likely to use in real-world storytelling**. However, we observed that **equivalent phrases around similar themes and topics did not frequently appear in the C4 dataset.** Many of the most common n-grams that appeared in the C4 were either meaningless or related to web interactions (accepting cookies, logging out, all rights reserved, etc.).

In addition to performing n-gram analysis, we note that the top 10-grams including specific words and phrases that tend to appear quite often in backstories: specifically, "small town," "fruits," "vegetables," "mental health," and "physical health." We also include "New York" to contrast against "small town" for our analysis. We search for the most common 10-grams appearing with each of these phrases inside in the backstories, C4 random, and C4 URL-filtered subsets, and some of our results can be found in Table B.7 and Table B.8.

Table B.7: Comparison of Most-Frequent n-grams with "New York".

Backstories		C4 - Random		C4 - URL Filtered	
Text	Occ.	Text	Occ.	Text	Occ.
going regularly to upstate New York we have these beautiful	3	New Jersey New Mexico New York North Carolina North Dakota	13074	Las Vegas Los Angeles New York Oakland Washington DC Hollywood	806
born and raised in New York City, where I attended	3	Patriots New Orleans Saints New York Giants New York Jets	1364	should have moved to New York and found a job	554
born and raised in New York City. As a child,	3	Saints New York Giants New York Jets Oakland Raiders Philadelphia	1254	the East River of New York City between Manhattan and	320

Continued					
Text	Occ.	Text	Occ.	Text	Occ.
born and raised in New York City. My parents are	3	Timberwolves New Orleans Pelicans New York Knicks Oklahoma City Thunder	1137	4 months ago About New York Winter 4 months ago	247
born and raised in New York City. I grew up	3	Predators New Jersey Devils New York Islanders New York Rangers	1094	New South Wales (192) New York (13) New Zealand (68)	200

Although the same phrases appear, they do not often appear in the same context, especially when comparing the random C4 subset and the backstories. For instance, in the random C4 subset, top 10-grams with "New York" appears in the context of sports teams (probably as part of a list), whereas in the social media subset of C4, the phrase "New York" has slightly more related contexts (moving there for a job, geographic description) to how the phrase appears in the backstories. The same story can be found in "small town": some other phrases from the C4 social media subset that do not appear as often as in the table but have similar contexts to the backstories are "but am from a small town in Southern Utah" or "Just a small town girl who writes about."

Table B.8: Comparison of Most-Frequent n-grams with "Small Town".

Backstories		C4 - Random		C4 - URL Filtered	
Text	Occ.	Text	Occ.	Text	Occ.
was born in a small town in the countryside of	50	of a 'Super Bloom'A small town in Southern California is	378	Signs & Billboards Skeletons Small Town America Sports Sports –	38
and raised in a small town in the Midwest. Growing	19	and unusual look at small town life in northern Vermont.	122	Collaborative No Depression Popdose Small Town Romance Some Velvet Songs:	33
grew up in a small town in the Midwest with	17	of Venom Brings the Small Town Scares in Cult of	65	Categories Select Category A Small Town –What? Dad Stories Small	21

Continued					
Text	Occ.	Text	Occ.	Text	Occ.
and raised in a small town in the Midwest. My	16	Pro-Science Recursivity Reprobate Spreadsheet Small Town Deviant Stderr Taslima Nasreen	48	Small Town–What? Dad Stories Small Town World Adventures Small World	21
was born in a small town in the middle of	16	Sigma Pi Skylar House Small Town Records (STR) Smart Home	36	sight & sound 2012 small town teen film thriller uk	21

For "fruits" and "vegetables," we find that in the backstories, most of the phrases are related to descriptions of a healthy diet; however, in the C4 random subset, most of the phrases are lists of menus and grocery items, while in the C4 URL filtered subset, there are more recipes, social media posts, and hashtags (as expected). An interesting pattern we identify with "mental" and "physical health" is that in both the random and URL filtered subset of C4, the phrase "physical health" frequently appears in the same context as "mental health." This observation directly relates to the phrases of "mental health is as important as physical health" (or variations of this) surface frequently in the generated backstories.

B.5 Details on the Surveys

In this section, we present details of human studies: Pew Research American Trends Panel Wave 110, Subversion Dilemma study [22], and meta-prejudice study [114] including list of questions in the format used in our study, and recruiting details.

American Trends Panel Wave 110

Pew Research Center conducted ATP Wave 110 from June 27, 2022 to July 4, 2022 with a focus on politics timely and topical. The number of total respondents are 6,174, with 1,551 self-identified as Republicans, 1,886 self-identified as Democrats, 1,885 self-identified as Independent, 777 self-identified as something else in terms of political party affiliation. The users are recruited through random sampling of residential addresses, a nationally representative online panel. The probability-based sampling ensures that nearly all U.S. adults have a chance of being selected. The final sample is weighted to be representative of the U.S. adult population based on gender, race, ethnicity, education, and political affiliation. We utilized ten questions as below, which are two symmetric sets of five questions each asked

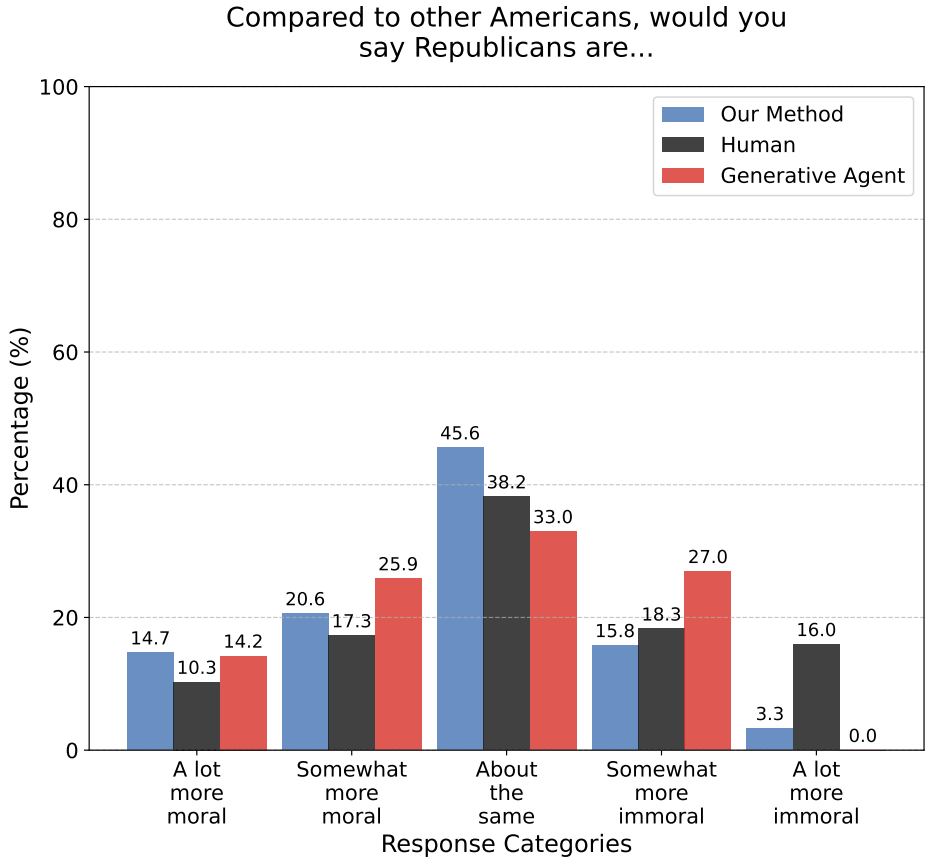

Figure B.2: **Response distribution** of humans (black), virtual personas (blue), and Generative Agent (red) for a question assessing perceived moral standing of Republicans relative to other Americans, asked to both Democrats and Republicans ("Compared to other Americans, would you say Republicans are...").

to all respondents. Since these questions are asked to all respondents, individual opinions of political partisans can be observed regardless of one's political affiliation.

Question: Compared to other Americans, would you say Democrats are...
(A) A lot more moral
(B) Somewhat more moral
(C) About the same
(D) Somewhat more immoral
(E) A lot more immoral
Answer:

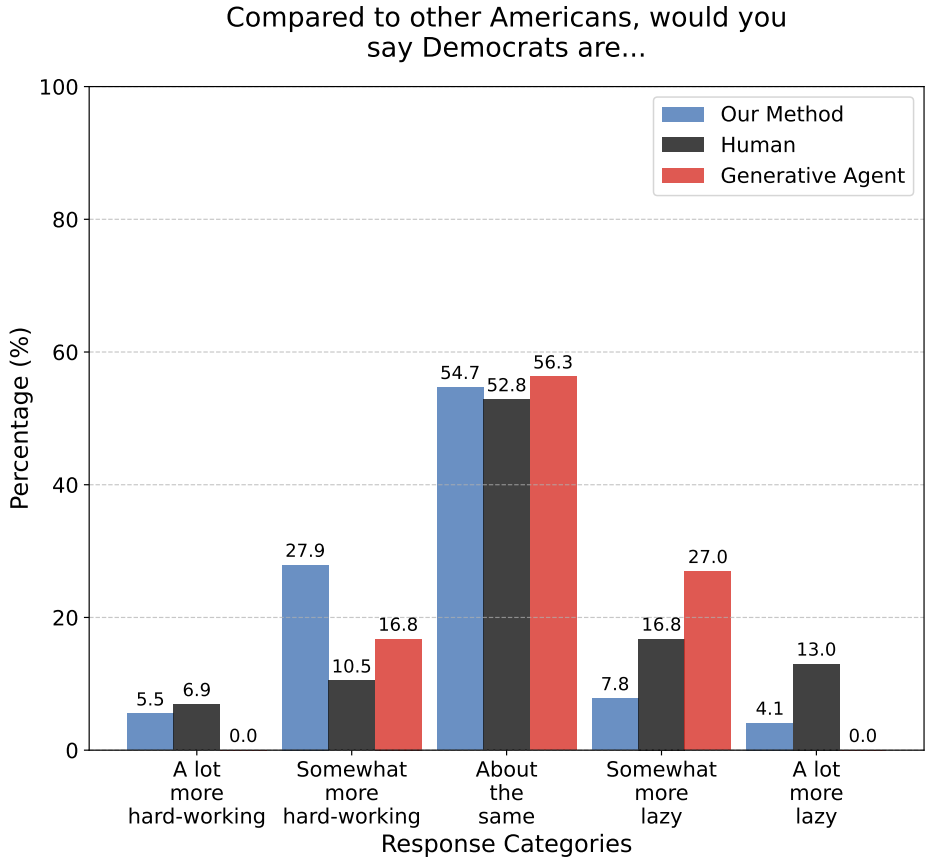

Figure B.3: **Response distribution** of humans (black), virtual personas (blue), and Generative Agent (red) for a question assessing perceived diligence of Democrats relative to other Americans, asked to both Democrats and Republicans ("Compared to other Americans, would you say Democrats are...").

Question: Compared to other Americans, would you say Democrats are...
(A) A lot more hard-working
(B) Somewhat more hard-working
(C) About the same
(D) Somewhat more lazy
(E) A lot more lazy
Answer:

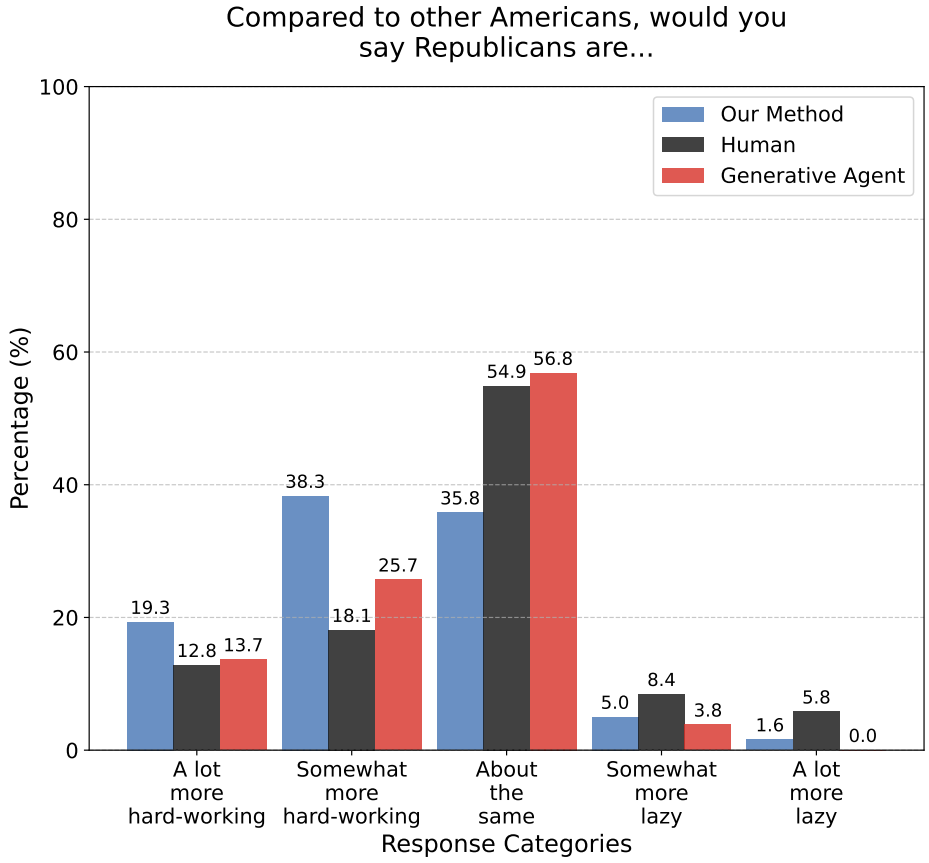

Figure B.4: **Response distribution** of humans (black), virtual personas (blue), and Generative Agent (red) for a question assessing perceived diligence of Republicans relative to other Americans, asked to both Democrats and Republicans ("Compared to other Americans, would you say Republicans are...").

Question: Compared to other Americans, would you say Democrats are...
(A) A lot more open-minded
(B) Somewhat more open-minded
(C) About the same
(D) Somewhat more close-minded
(E) A lot more close-minded
Answer:

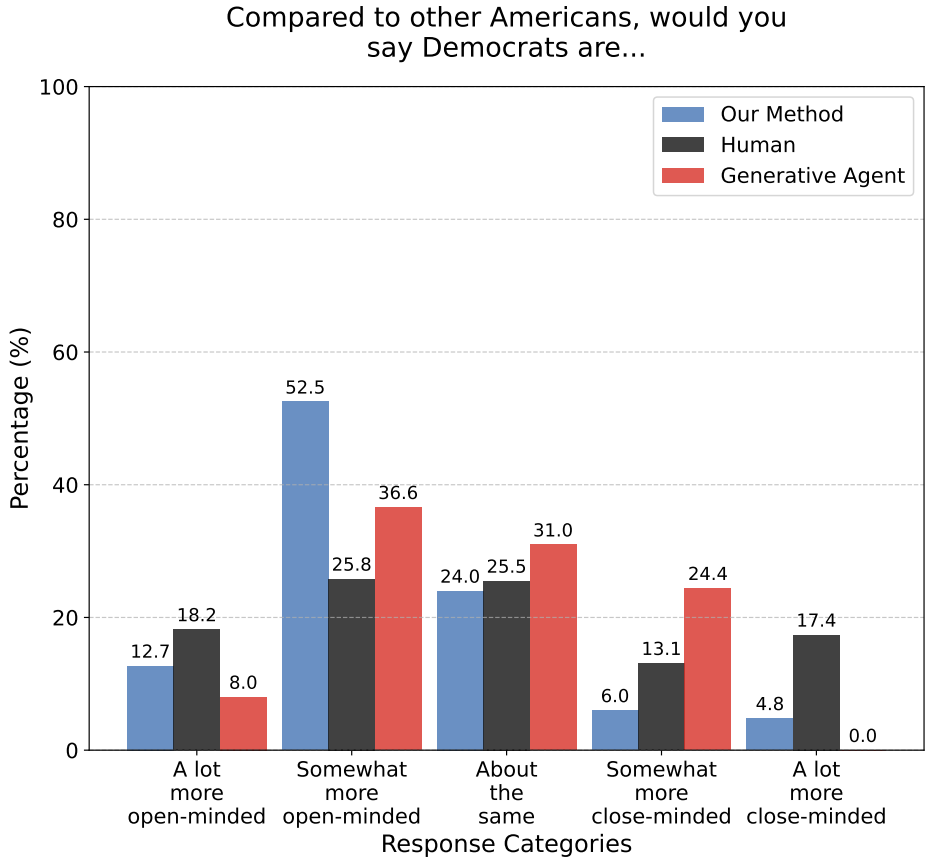

Figure B.5: **Response distribution** of humans (black), virtual personas (blue), and Generative Agent (red) for a question assessing perceived open-mindedness of Democrats relative to other Americans, asked to both Democrats and Republicans ("Compared to other Americans, would you say Democrats are...").

Question: Compared to other Americans, would you say Democrats are...
(A) A lot more intelligent
(B) Somewhat more intelligent
(C) About the same
(D) Somewhat more unintelligent
(E) A lot more unintelligent
Answer:

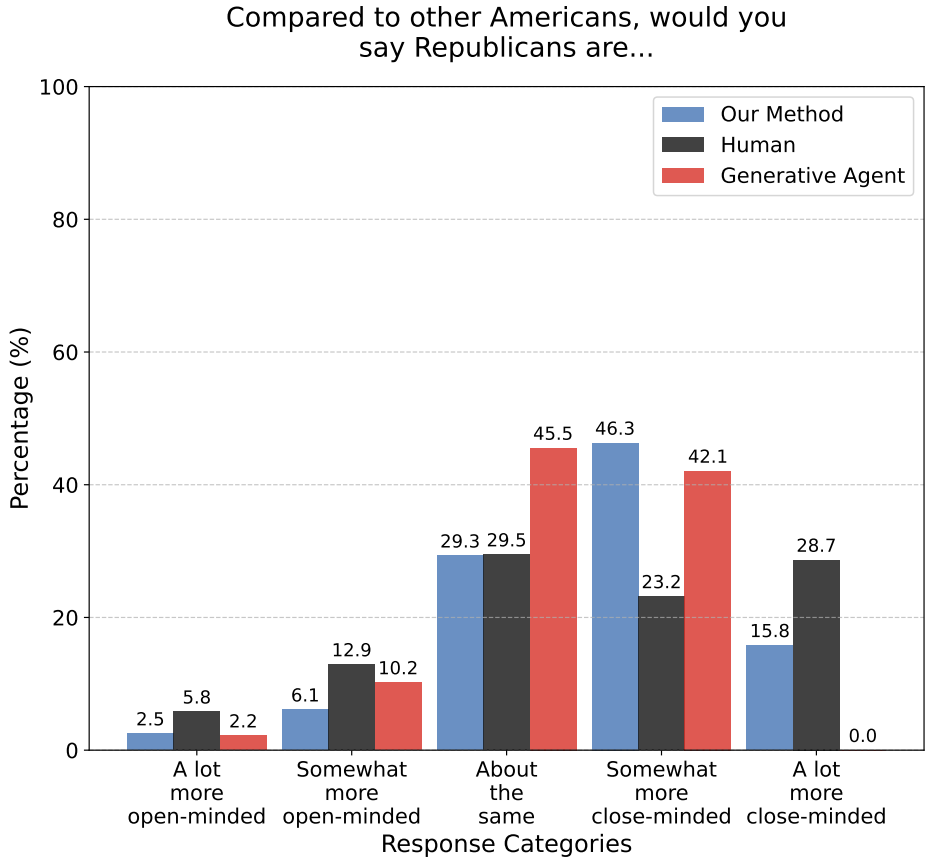

Figure B.6: **Response distribution** of humans (black), virtual personas (blue), and Generative Agent (red) for a question assessing perceived open-mindedness of Republicans relative to other Americans, asked to both Democrats and Republicans ("Compared to other Americans, would you say Republicans are...").

Question: Compared to other Americans, would you say Democrats are...
(A) A lot more honest
(B) Somewhat more honest
(C) About the same
(D) Somewhat more dishonest
(E) A lot more dishonest
Answer:

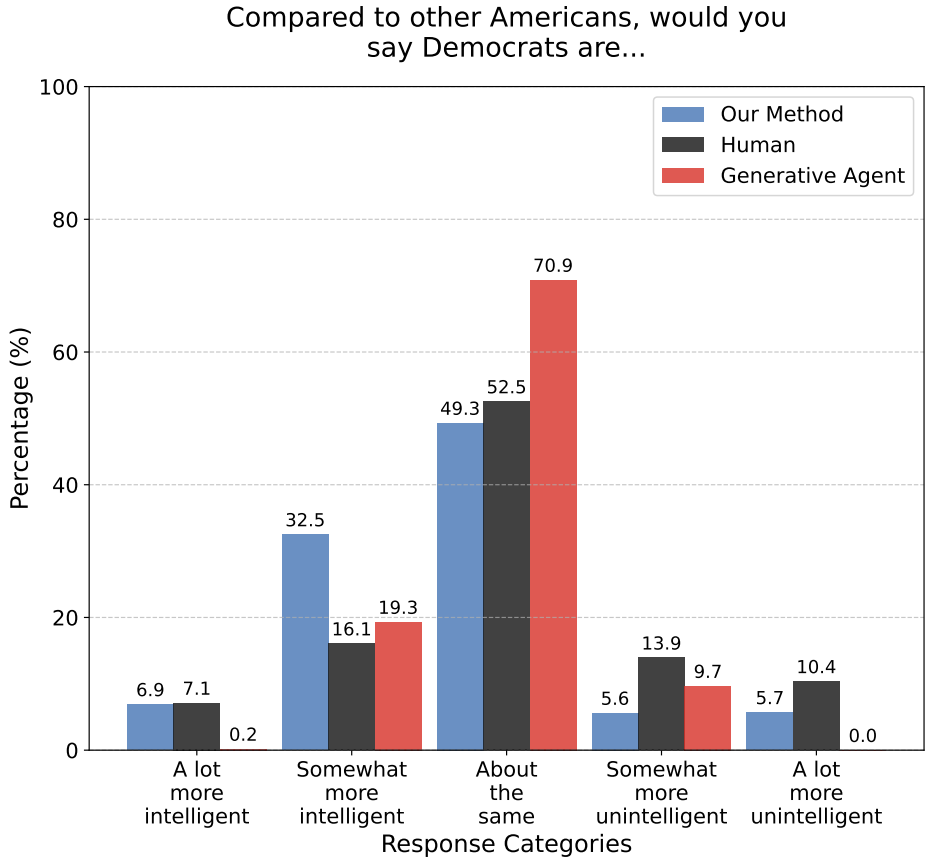

Figure B.7: **Response distribution** of humans (black), virtual personas (blue), and Generative Agent (red) for a question assessing perceived intelligence of Democrats relative to other Americans, asked to both Democrats and Republicans ("Compared to other Americans, would you say Democrats are...").

Question: Compared to other Americans, would you say Republicans are...
(A) A lot more moral
(B) Somewhat more moral
(C) About the same
(D) Somewhat more immoral
(E) A lot more immoral
Answer:

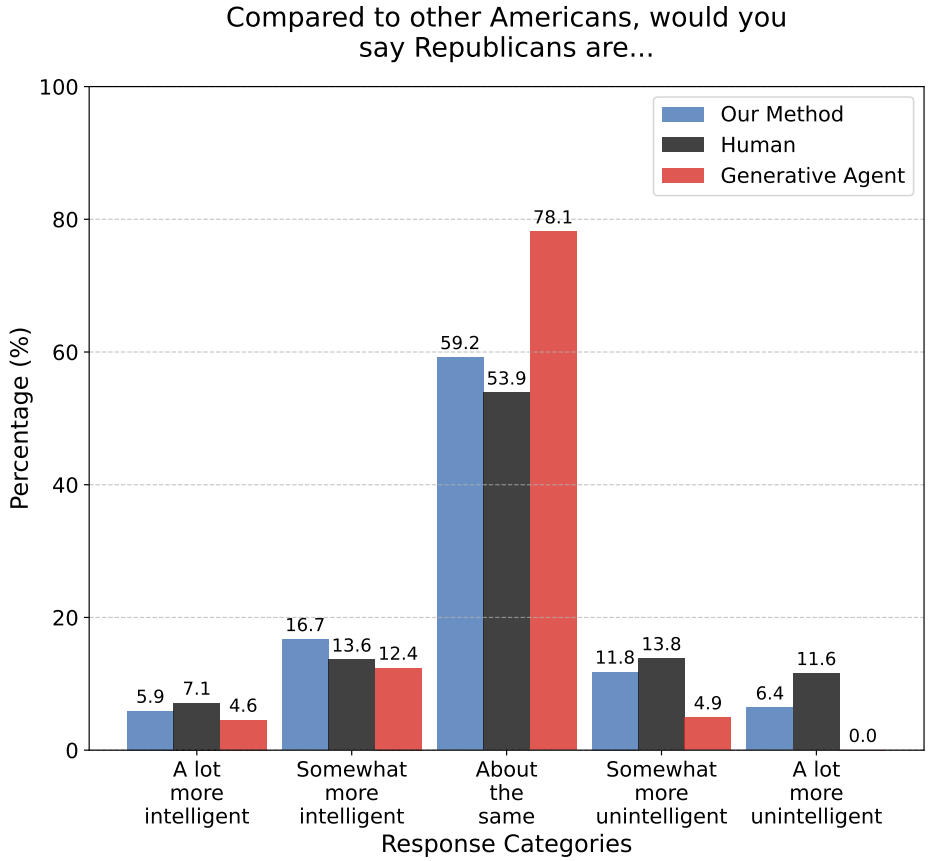

Figure B.8: **Response distribution** of humans (black), virtual personas (blue), and Generative Agent (red) for a question assessing perceived intelligence of Republicans relative to other Americans, asked to both Democrats and Republicans ("Compared to other Americans, would you say Republicans are...").

Question: Compared to other Americans, would you say Republicans are...
(A) A lot more hard-working
(B) Somewhat more hard-working
(C) About the same
(D) Somewhat more lazy
(E) A lot more lazy
Answer:

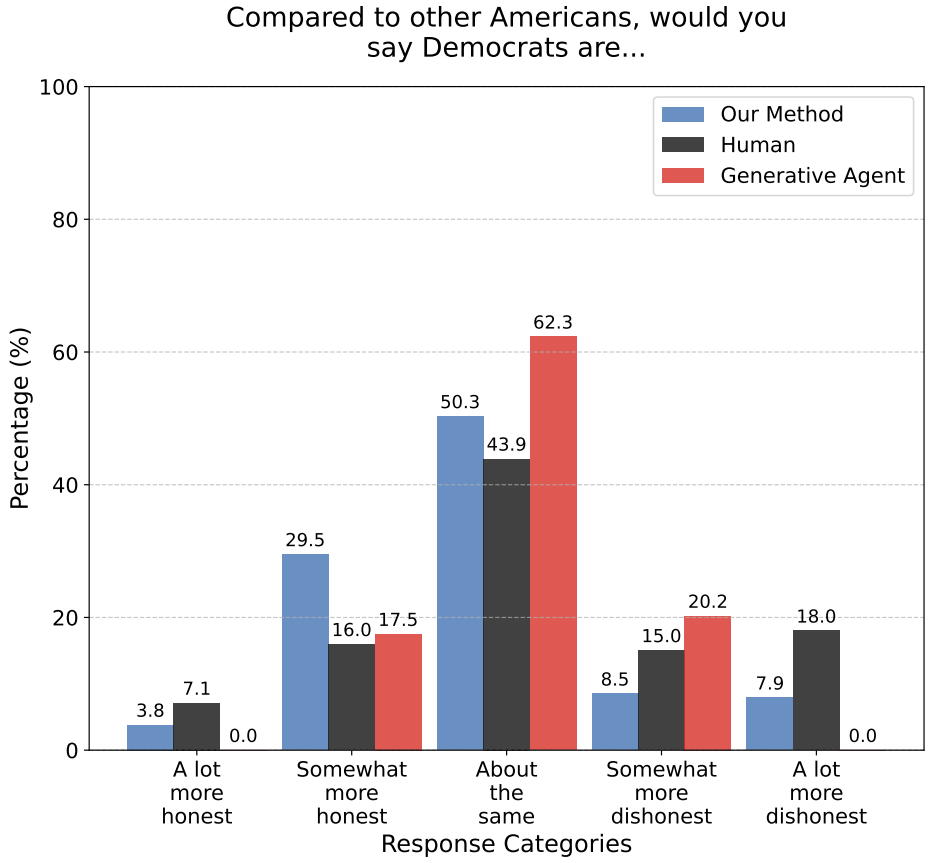

Figure B.9: **Response distribution** of humans (black), virtual personas (blue), and Generative Agent (red) for a question assessing perceived honesty of Democrats relative to other Americans, asked to both Democrats and Republicans ("Compared to other Americans, would you say Democrats are...").

Question: Compared to other Americans, would you say Republicans are...
(A) A lot more open-minded
(B) Somewhat more open-minded
(C) About the same
(D) Somewhat more close-minded
(E) A lot more close-minded
Answer:

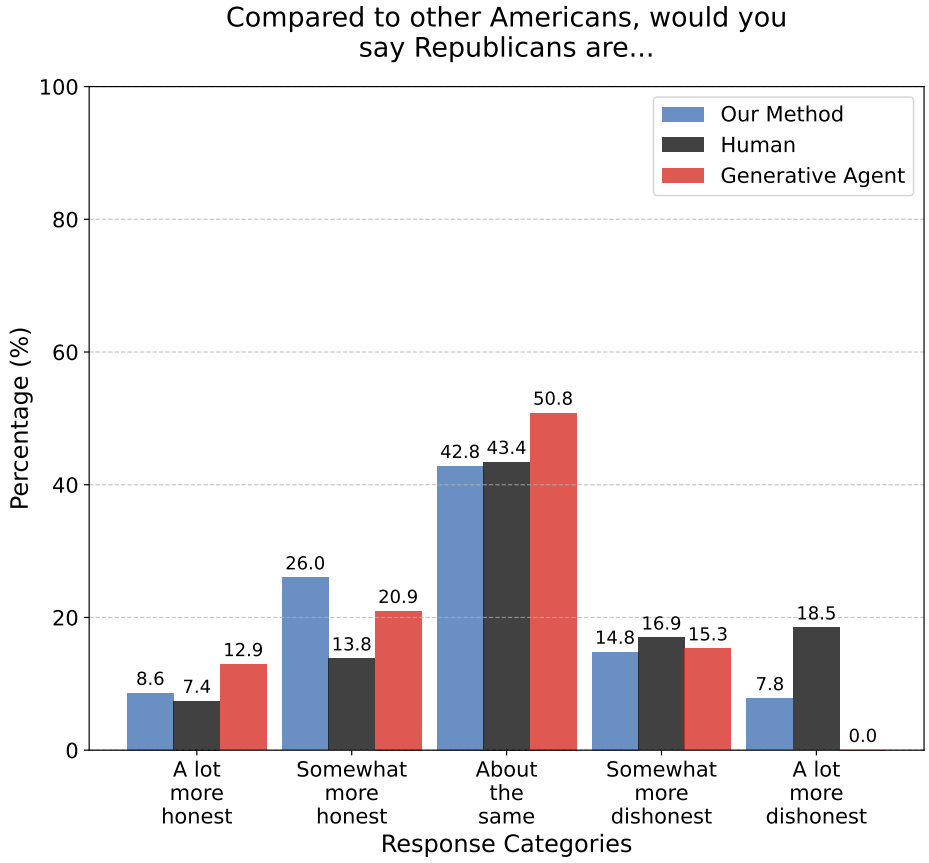

Figure B.10: **Response distribution** of humans (black), virtual personas (blue), and Generative Agent (red) for a question assessing perceived honesty of Republicans relative to other Americans, asked to both Democrats and Republicans ("Compared to other Americans, would you say Republicans are...").

Question: Compared to other Americans, would you say Republicans are...
(A) A lot more intelligent
(B) Somewhat more intelligent
(C) About the same
(D) Somewhat more unintelligent
(E) A lot more unintelligent
Answer:

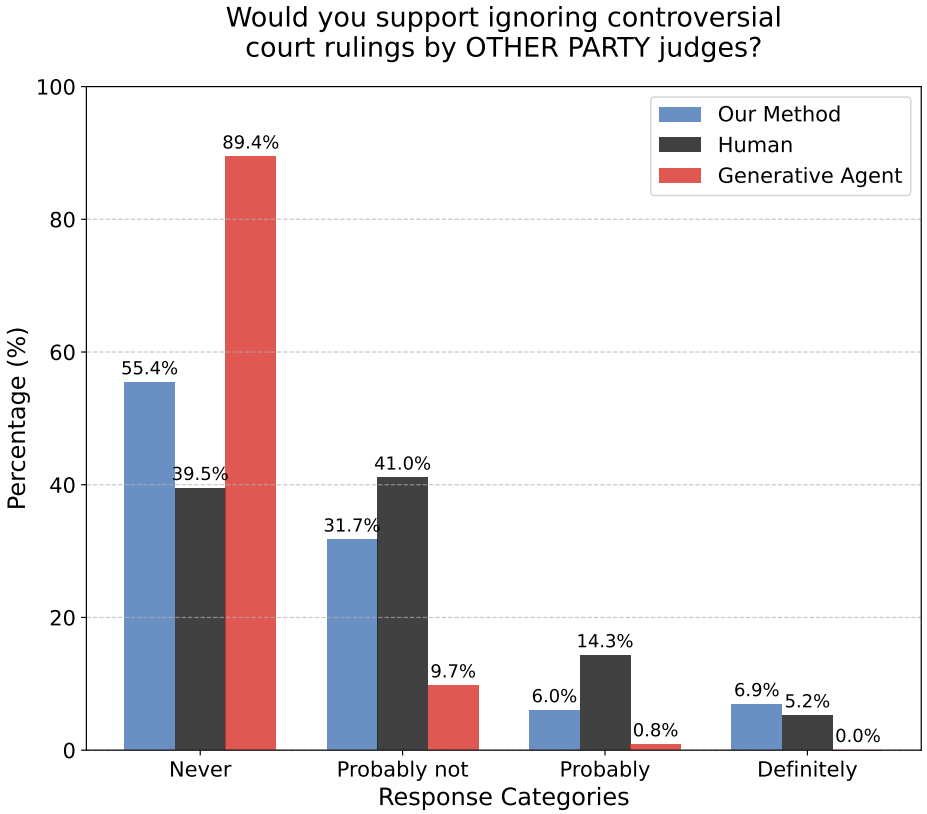

Figure B.11: **Response distribution** of humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would YOU support ignoring controversial court rulings by DEMOCRAT (REPUBLICAN) judges?' asked to Republicans (Democrats).

Question: Compared to other Americans, would you say Republicans are...
(A) A lot more honest
(B) Somewhat more honest
(C) About the same
(D) Somewhat more dishonest
(E) A lot more dishonest
Answer:

Subversion Dilemma Study

Subversion Dilemma study [22] Study 1 is conducted from July 15, 2021 to August 6, 2021. The number of total respondents are 1,536, with 723 self-identified as Republicans and 813 self-identified as Democrats. Participants were recruited via Lucid, an online platform for gathering nationally representative samples. We utilized 24 questions as below, which are two

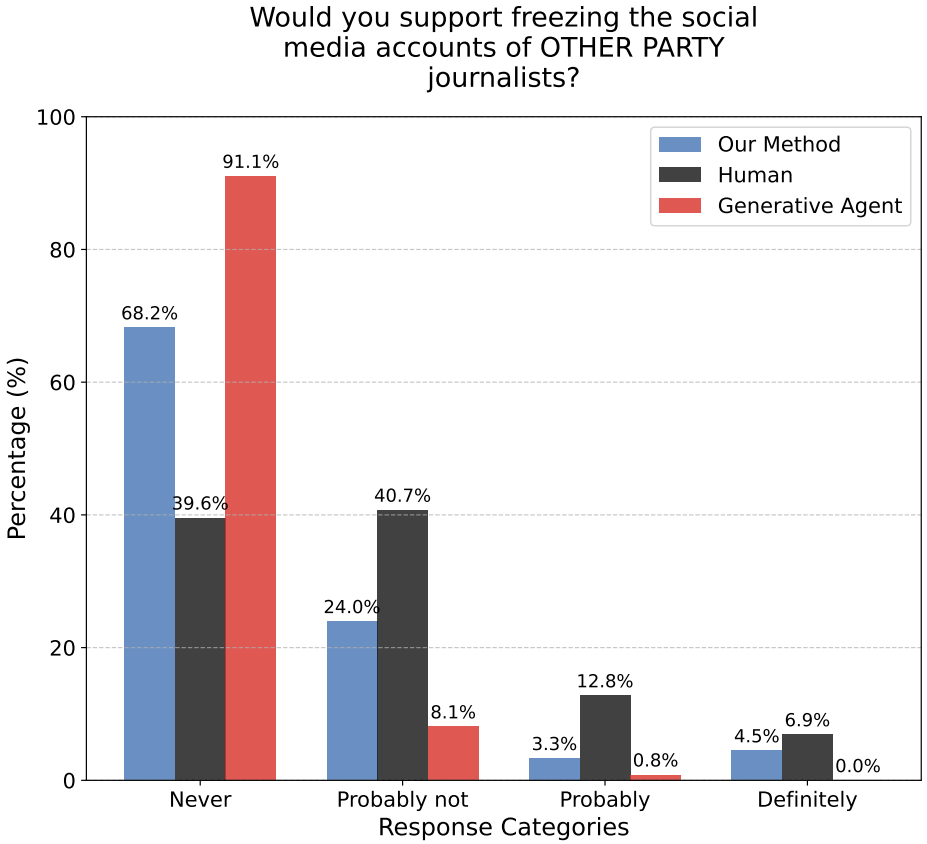

Figure B.12: **Response distribution** of humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would YOU support freezing the social media accounts of DEMOCRAT (REPUBLICAN) journalists?' asked to Republicans (Democrats).

symmetric sets of 12 questions (6 self-subversion items and 6 meta-subversion items) each set being asked to one political party. Six subversion items are related to understanding of democratic norms, including ignoring controversial court rulings and freezing the social media accounts. Here we present 12 question items asked to Republicans; another 12 items asked to Democrats are obtained by replacing 'DEMOCRATS' to 'REPUBLICANS' and 'REPUBLICANS' to 'DEMOCRATS'.

Question: Would MOST DEMOCRATS support ignoring controversial court rulings by REPUBLICAN JUDGES?
(A) Never
(B) Probably Not
(C) Probably
(D) Definitely
Answer:

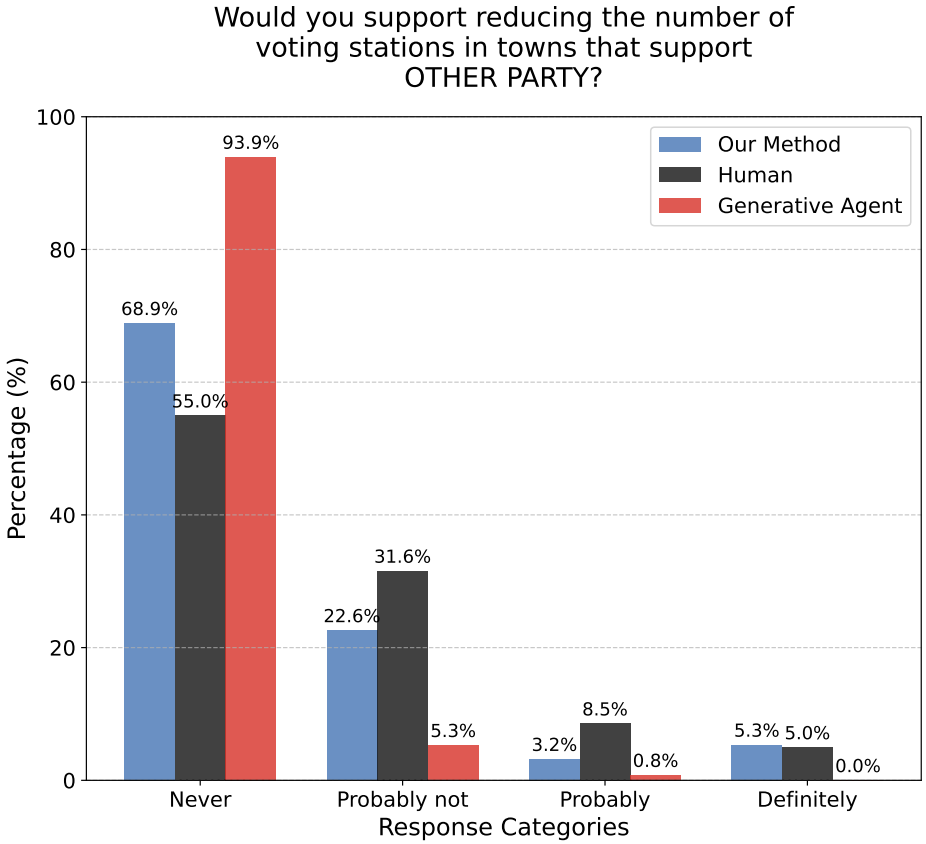

Figure B.13: **Response distribution** of humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would YOU support reducing the number of voting stations in towns that support DEMOCRATS (REPUBLICANS)?' asked to Republicans (Democrats).

Question: Would MOST DEMOCRATS support freezing the social media accounts of REPUBLICAN JOURNALISTS?
(A) Never
(B) Probably Not
(C) Probably
(D) Definitely
Answer:

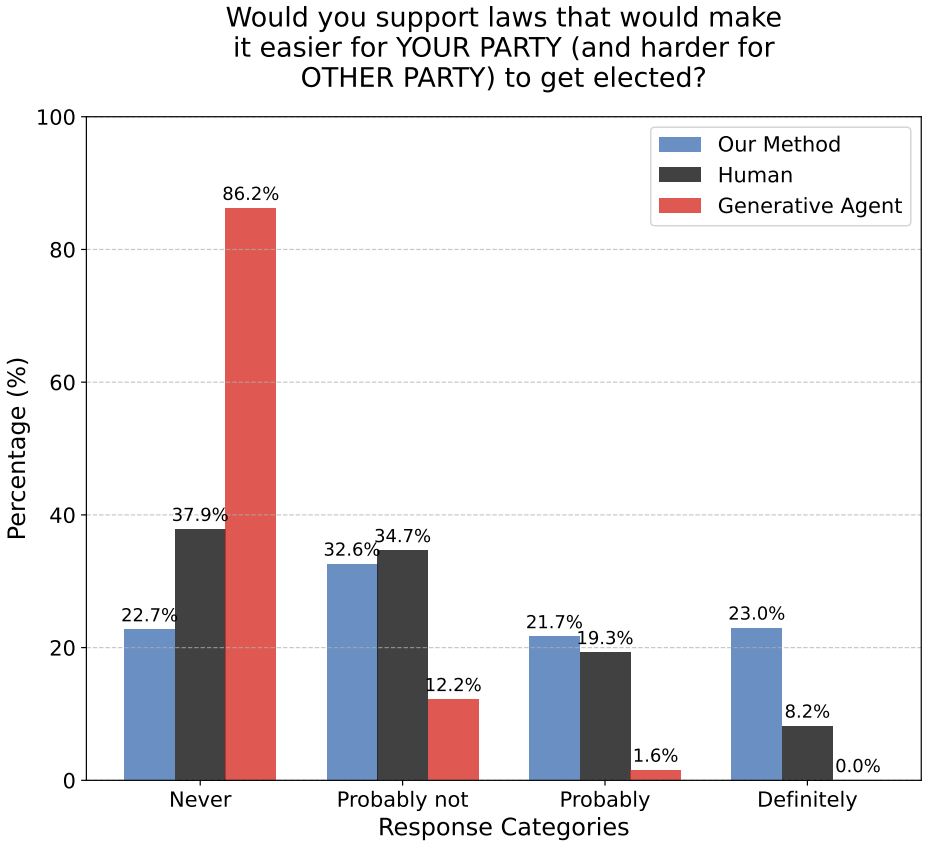

Figure B.14: **Response distribution** of humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would YOU support laws that would make it easier for REPUBLICANS (DEMOCRATS) and harder for DEMOCRATS (REPUBLICANS) to get elected?' asked to Republicans (Democrats).

Question: Would MOST DEMOCRATS support reducing the number of voting stations
in towns that support REPUBLICANS?
(A) Never
(B) Probably Not
(C) Probably
(D) Definitely
Answer:

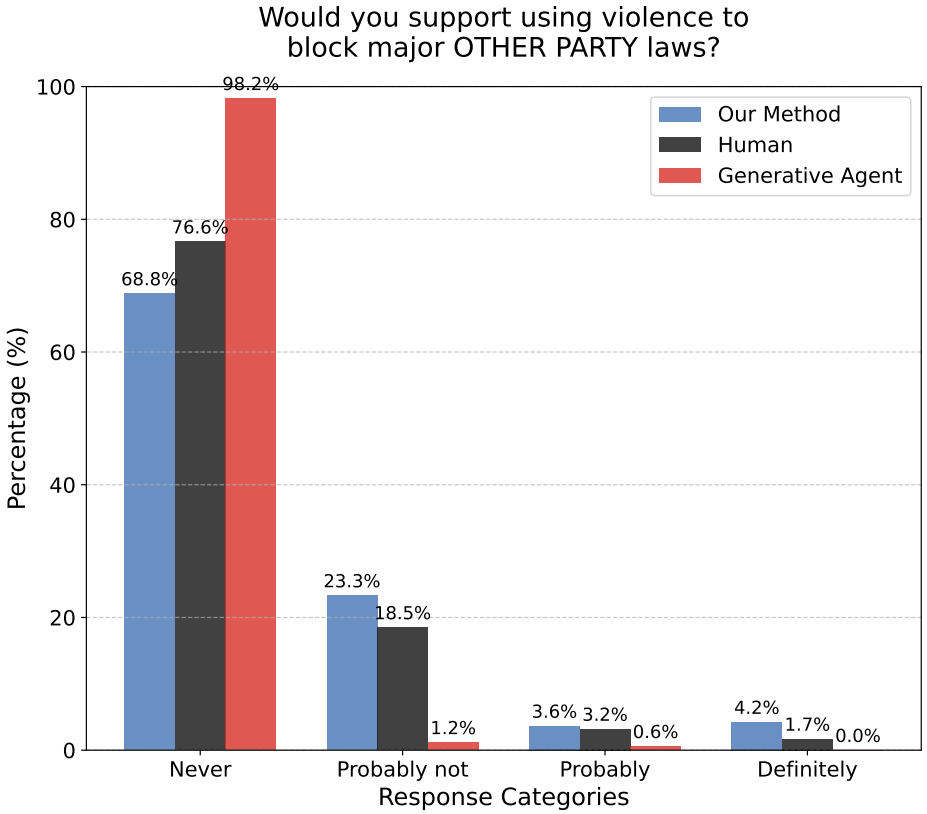

Figure B.15: **Response distribution** of humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would YOU support using violence to block major DEMOCRAT (REPUBLICAN) laws?' asked to Republicans (Democrats).

Question: Would MOST DEMOCRATS support laws that would make it easier for DEMOCRATS (and harder for REPUBLICANS) to get elected?
(A) Never
(B) Probably Not
(C) Probably
(D) Definitely
Answer:

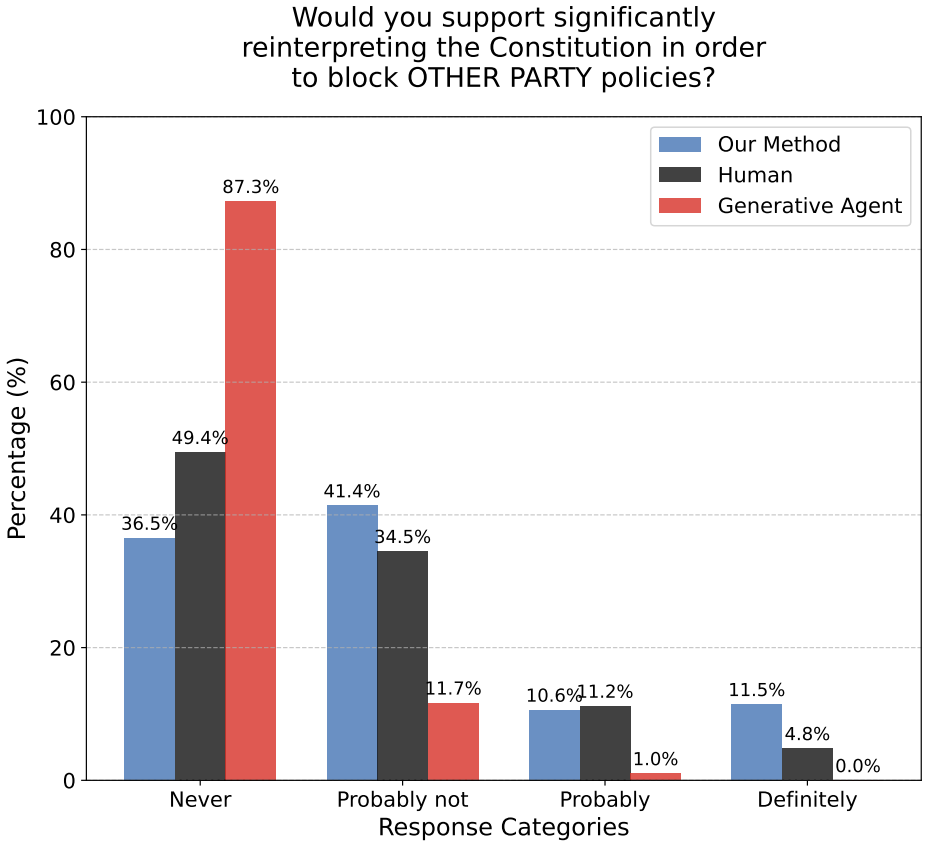

Figure B.16: **Response distribution** of humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would YOU support significantly reinterpreting the Constitution in order to block DEMOCRAT (REPUBLICAN) policies?' asked to Republicans (Democrats).

Question: Would MOST DEMOCRATS support using violence to block major REPUBLICAN laws?
(A) Never
(B) Probably Not
(C) Probably
(D) Definitely
Answer:

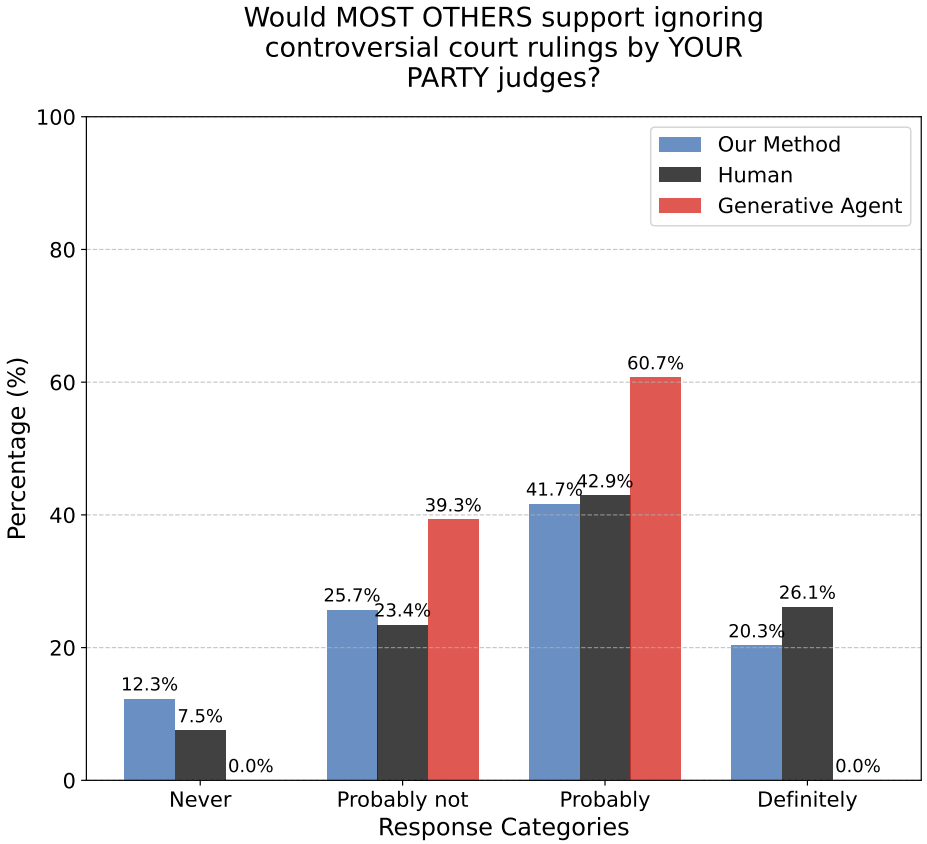

Figure B.17: **Response distribution** of humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would MOST DEMOCRATS (REPUBLICANS) support ignoring controversial court rulings by REPUBLICAN (DEMOCRAT) JUDGES?' asked to Republicans (Democrats).

Question: Would MOST DEMOCRATS support significantly reinterpreting the Constitution in order to block REPUBLICAN policies?
(A) Never
(B) Probably Not
(C) Probably
(D) Definitely
Answer:

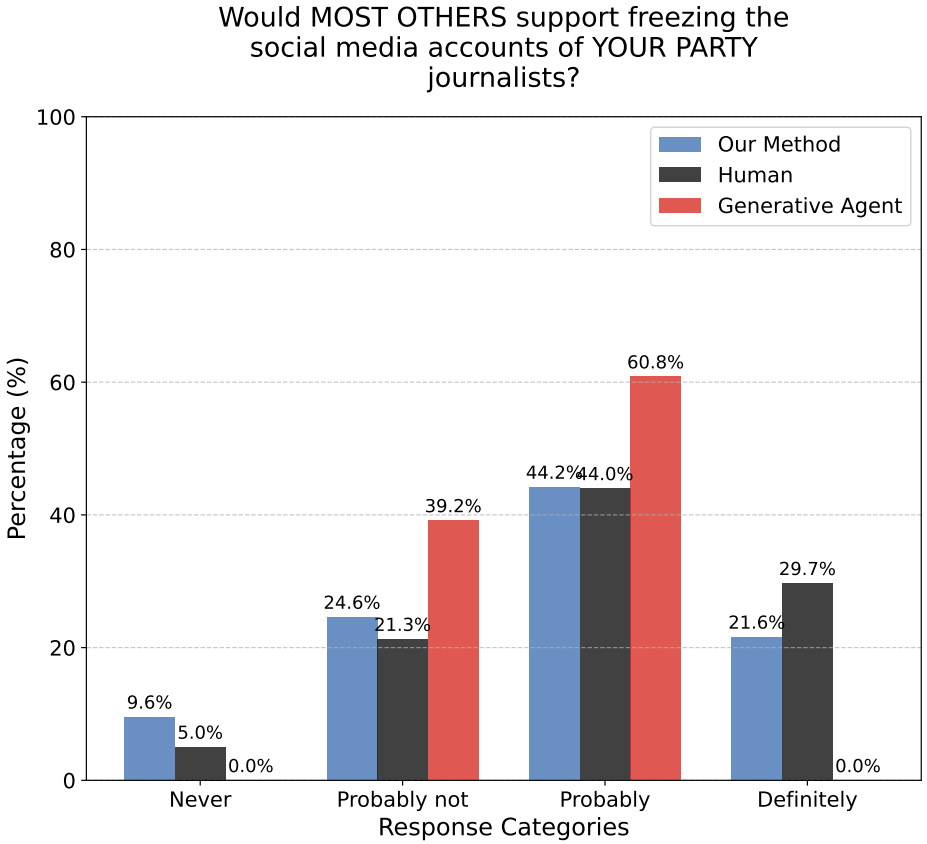

Figure B.18: **Response distribution** of humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would MOST DEMOCRATS (REPUBLICANS) support freezing the social media accounts of REPUBLICAN (DEMOCRAT) JOURNALISTS?' asked to Republicans (Democrats).

Question: Would YOU support ignoring controversial court rulings by DEMOCRAT JUDGES?
(A) Never
(B) Probably Not
(C) Probably
(D) Definitely
Answer:

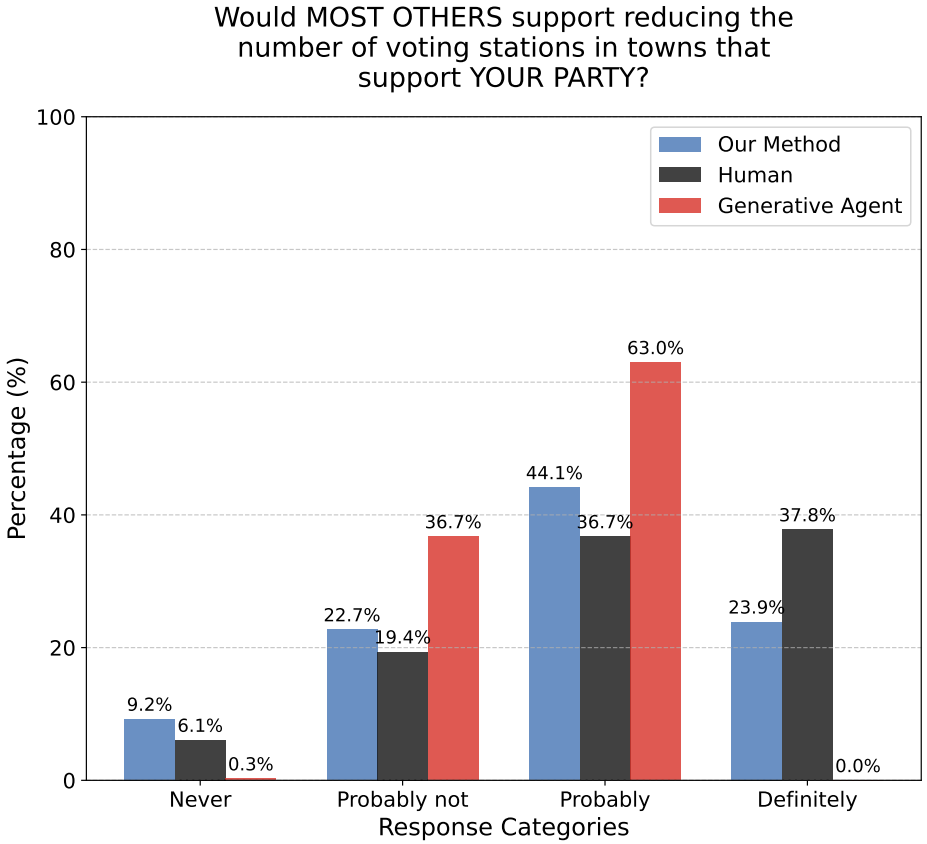

Figure B.19: **Response distribution** of humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would MOST DEMOCRATS (REPUBLICANS) support reducing the number of voting stations in towns that support REPUBLICANS (DEMOCRATS)?' asked to Republicans (Democrats).

Question: Would YOU support freezing the social media accounts of DEMOCRAT JOURNALISTS?
(A) Never
(B) Probably Not
(C) Probably
(D) Definitely
Answer:

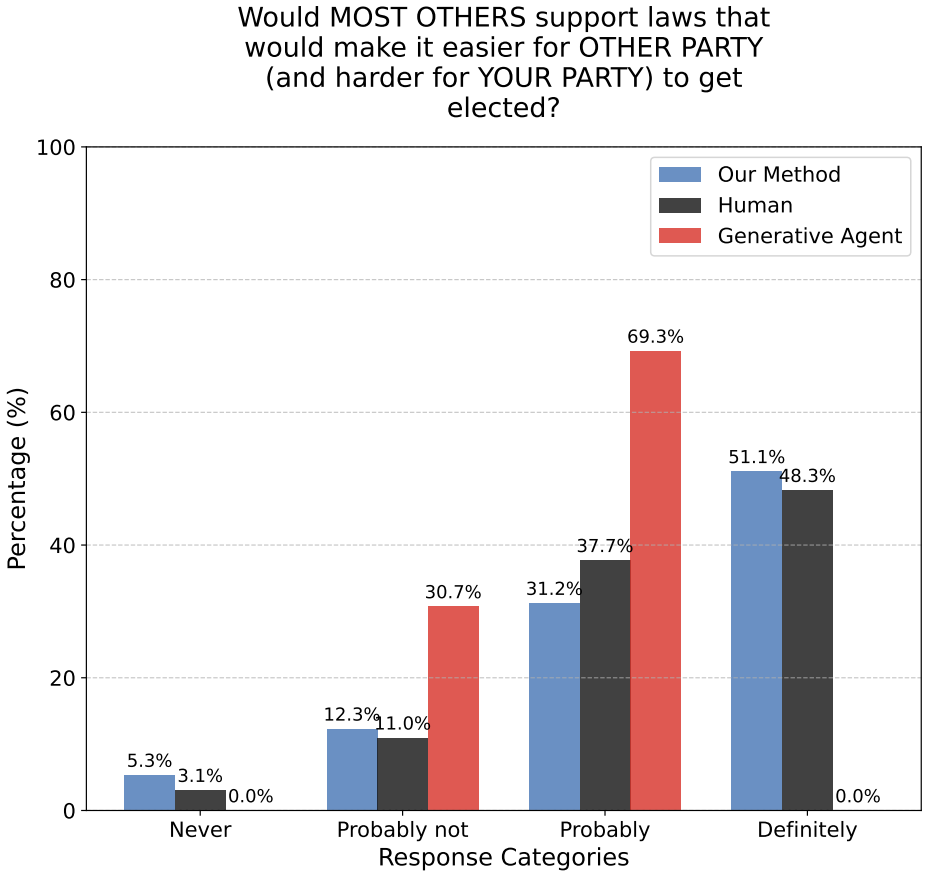

Figure B.20: **Response distribution** of humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would MOST DEMOCRATS (REPUBLICANS) support laws that would make it easier for DEMOCRATS (REPUBLICANS) and harder for REPUBLICANS (DEMOCRATS) to get elected?' asked to Republicans (Democrats).

Question: Would YOU support reducing the number of voting stations in towns that support DEMOCRATS?
(A) Never
(B) Probably Not
(C) Probably
(D) Definitely
Answer:

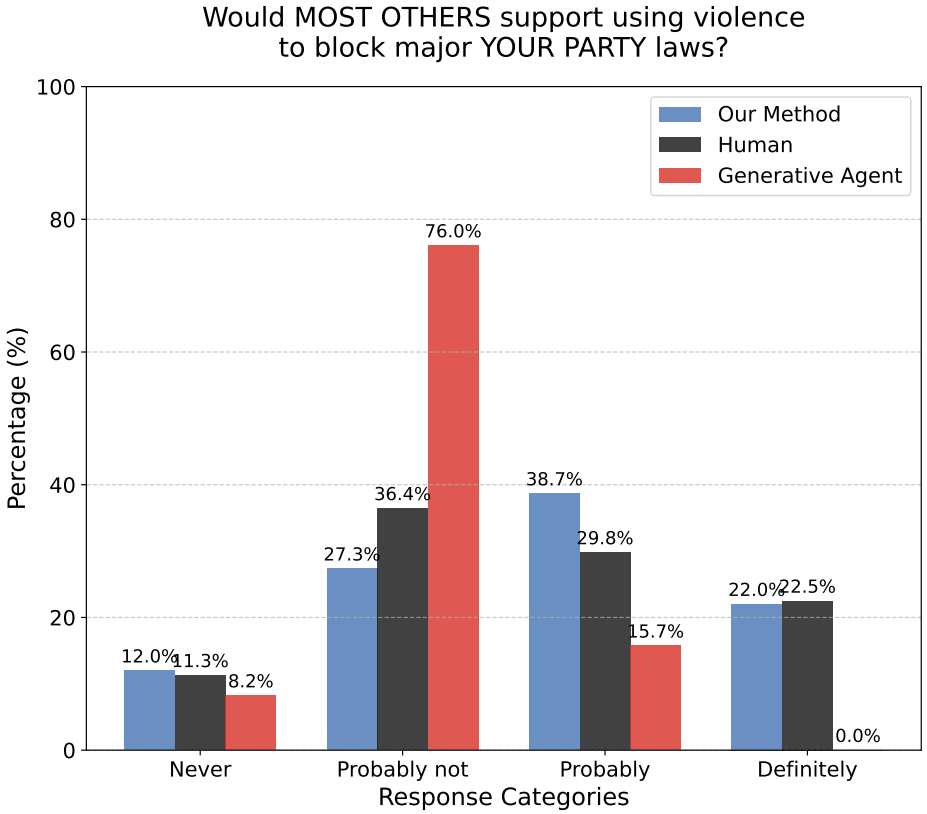

Figure B.21: **Response distribution** of humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would MOST DEMOCRATS (REPUBLICANS) support using violence to block major REPUBLICAN (DEMOCRAT) laws?' asked to Republicans (Democrats).

Question: Would YOU support laws that would make it easier for REPUBLICANS (and
harder for DEMOCRATS) to get elected?
(A) Never
(B) Probably Not
(C) Probably
(D) Definitely
Answer:

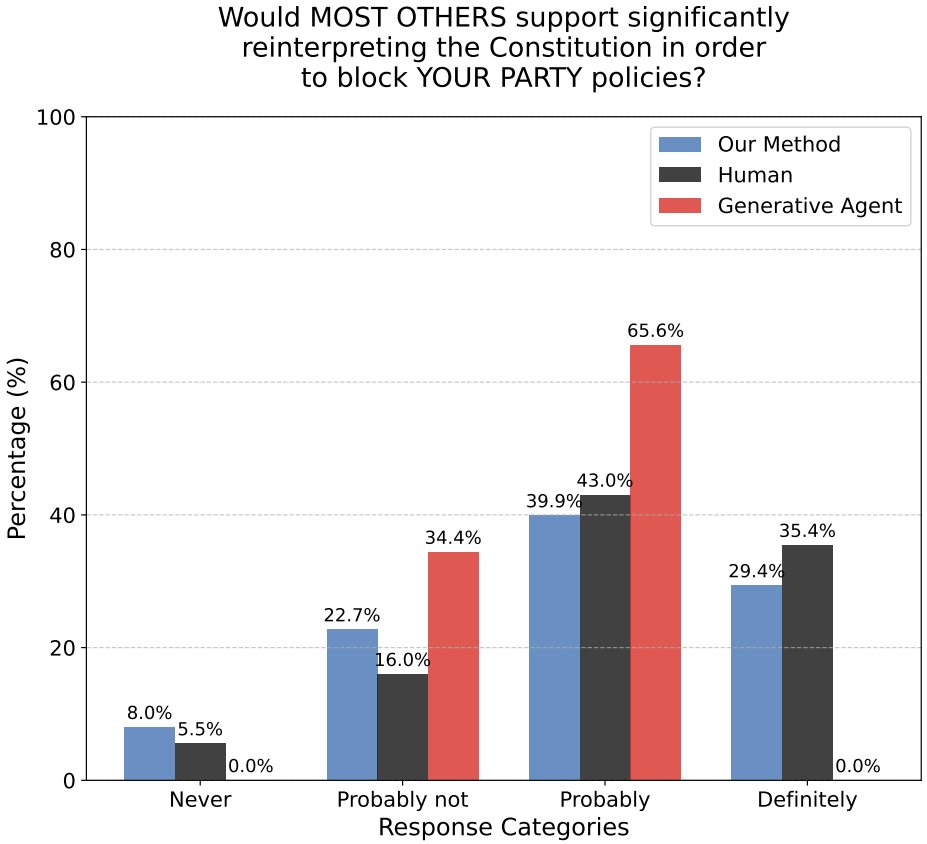

Figure B.22: **Response distribution** of humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would MOST DEMOCRATS (REPUBLICANS) support significantly reinterpreting the Constitution in order to block REPUBLICAN (DEMOCRAT) policies?' asked to Republicans (Democrats).

Question: Would YOU support using violence to block major DEMOCRAT laws?
(A) Never
(B) Probably Not
(C) Probably
(D) Definitely
Answer:

Question: Would YOU support significantly reinterpreting the Constitution in order to block DEMOCRAT policies?
(A) Never
(B) Probably Not
(C) Probably
(D) Definitely
Answer:

Meta-Prejudice Study

Meta-prejudice study [114] Study 1 was conducted in early November 2018. The total number of respondents was 1,053, comprising 533 self-identified Democrats and 520 self-identified Republicans. Participants were recruited through the AmeriSpeak panel [116], which is operated by NORC at the University of Chicago. AmeriSpeak is a probability-based panel designed to be representative of the U.S. household population. Randomly selected U.S. households are sampled using area probability and address-based sampling, with a known, non-zero probability of selection from the NORC National Sample Frame. The panel provides coverage of approximately 97% of the U.S. household population. Those excluded include people with P.O. Box-only addresses, some addresses not listed in USPS files, and newly constructed dwellings.

Participants were asked to rate their own feelings of prejudice towards the opposing party, as well as their meta-perceptions of how much prejudice the opposing party held toward their own party. We utilized 6 questions as below: the first two questions are asked to all respondents, the next two questions are asked to Democrats, and the last two questions are asked to Republicans. The original study utilized a slider bar to indicate respondent answers from a scale of 1-100. In accordance with the rest of the studies considered in this work, we cast the questions as Likert scale multiple-choice questionnaires with five options that uniformly partition the original range of answers. Human responses are accordingly cast to categorical options.

Question: How warm or cold do you feel towards DEMOCRATS?
(A) Very cold
(B) Cold
(C) Neutral
(D) Warm
(E) Very warm
Answer:

Question: How warm or cold do you feel towards REPUBLICANS?
(A) Very cold
(B) Cold
(C) Neutral
(D) Warm
(E) Very warm
Answer:

Question: How warm or cold do you think REPUBLICANS feel towards DEMOCRATS?
(A) Very cold
(B) Cold
(C) Neutral
(D) Warm
(E) Very warm
Answer:

Question: How warm or cold do you think REPUBLICANS feel towards REPUBLICANS
(A) Very cold
(B) Cold
(C) Neutral
(D) Warm
(E) Very warm
Answer:

Question: How warm or cold do you think DEMOCRATS feel towards DEMOCRATS?
(A) Very cold
(B) Cold
(C) Neutral
(D) Warm
(E) Very warm
Answer:

Question: How warm or cold do you think DEMOCRATS feel towards REPUBLICANS?
(A) Very cold
(B) Cold
(C) Neutral
(D) Warm
(E) Very warm
Answer:

B.6 Details on the Generative Agent Framework

Here we present the exact prompt for our baseline experiments reproducing the Generative Agents framework [122]. The prompt is adopted directly from the original work with a minimal modification. Initially, an interview-based backstory is provided to the "expert reflection" module that operates with `GPT-4o` to infer the most high-level information encoded in the transcript:

Imagine you are an expert political scientist (with a PhD) taking notes while observing this interview. Write observations/reflections about the interviewee's political views, affiliation with political parties, and stances about key societal issues. (You should make more than 5 observations and fewer than 20. Choose the number that makes sense given the depth of the interview content above.)

We generated up to 20 observations per each transcript following the original approach [122]. These observations, along with the interview-based backstory, are provided to Generative Agents to generate a prediction as follows:

Participant's interview transcript:
(INTERVIEW TRANSCRIPT)

Expert political scientist's observations/reflections:
(EXPERT REFLECTIONS)

=====

Task: What you see above is an interview transcript. Based on the interview transcript, I want you to predict the participant's survey responses. All questions are multiple choice where you must guess from one of the options presented.

As you answer, I want you to take the following steps:
Step 1) Describe in a few sentences the kind of person that would choose each of the response options. ("Option Interpretation")
Step 2) For each response options, reason about why the Participant might answer with the particular option. ("Option Choice")
Step 3) Write a few sentences reasoning on which of the option best predicts the participant's response ("Reasoning")
Step 4) Predict how the participant will actually respond in the survey. Predict based on the interview and your thoughts, but ultimately, DON'T over think it. Use your system 1 (fast, intuitive) thinking. ("Response")

Here are the questions:

(SURVEY QUESTIONS WE ARE TRYING TO RESPOND TO)

—

Output format – output your response in json, where you provide the following:

{"1": {"Q": "<repeat the question you are answering>",
"Option Interpretation": {
"<option 1>": "a few sentences the kind of person that would choose each of the response options",
"<option 2>": "..."},
"Option Choice": {
"<option 1>": "reasoning about why the participant might choose each of the options",
"<option 2>": "..."},
"Reasoning": "<reasoning on which of the option best predicts the participant's response>",
"Response": "<your prediction on how the participant will answer the question>"},
"2": {...},
...}

A subsequent JSON format output is parsed to predict the virtual persona's response with Generative Agents.

Appendix C

Identity and Cooperation within Groups of Real and Simulated Humans

C.1 Backstory Generation

To generate high-quality virtual backstories, we follow the backstory construction pipeline introduced by [79]. We adopt the same set of interview questions used in their work to elicit long-form, coherent, and persona-consistent narratives. Table C.1 provides the full list of prompts used to construct these interview-style backstories.

Backstories are generated iteratively by prompting a language model with each question in sequence, along with all previously generated question–answer pairs. Specifically, to generate a response for the i-th question, the model is conditioned on the current prompt and the full context from steps 1 through $(i-1)$. This incremental setup promotes temporal coherence and a consistent persona across the entire narrative. To further enhance quality, we apply rejection sampling guided by an LLM-as-a-critic model, which filters out low-quality generations. The resulting backstories are both expressive and internally consistent, enabling faithful simulation of human-like personas.

We provide the examples of generated virtual personas for Democrat and Republican virtual subjects below:

Example Interview-Style Backstory of a *Democrat* User

Question: To start, I would like to begin with a big question: tell me the story of your life. Start from the beginning–from your childhood, to education, to family and relationships, and to any major life events you may have had.

Answer: I have always lived in Maryville. I am 18 years old and was born on March 31st, 1984. I grew up with my mother, father, and two younger brothers, Jacob and Zachary. My father works at a steel mill and my mother is a nurse at the hospital.

My childhood was rather ordinary. I went to school, hung out with my friends, and don't recall any major life events. I attended Maryville High School, completed the required courses with some electives, and graduated in 2003.

I have never been in a serious relationship. I've had a few crushes, but nothing major. My current crush is my friend Alex—we're very close but have never been romantic. At 18, I don't want to be tied down. I'd rather enjoy life while I'm young.

I'm currently a sophomore in college majoring in Psychology. I'm not sure yet what I'll do with my degree; right now my focus is on getting through school and being happy.

Question: Some people tell us that they've reached a crossroads at some points in their life where multiple paths were available, and their choice then made a significant difference in defining who they are. What about you? Was there a moment like that for you, and if so, could you tell me the whole story about that from start to finish?

Answer: I've never had a dramatic crossroads or done anything particularly exciting. The one decision that shaped my life was choosing to go to college. Before that, I wasn't sure what I wanted to do. Going to college gave me direction and made me feel like I was on the right track. I'm doing well in school and looking forward to the future.

Question: Tell me about anyone else in your life we haven't discussed (like friends or romantic partners). Are there people outside of your family who are important to you?

Answer: My best friend is Alex, the one I mentioned before. We spend a lot of time together and I really value his company. I have a few other friends I see occasionally, but Alex is the most important. I don't have any romantic partners right now and I'm fine with that. I feel lucky to have a good family, good friends, and a good life.

Question: Now let's talk about your current neighborhood. Tell me all about the neighborhood and area in which you are living now.

Answer: I live in a small town called Maryville, about an hour from Asheville. It's a calm area with lots of trees and parks. I enjoy the slower pace of life here and the ability to walk outside and enjoy the fresh air.

Maryville has a few bars and restaurants but not much else going on, which I actually like. I prefer small towns over big cities. One downside is the lack of job opportunities—most people work at the steel mill, attend school, or commute elsewhere. Finding well-paying work here is hard, which makes moving out on my own challenging. Still, overall I'm happy living here.

Question: Tell me about any recent changes to your daily routine.

Answer: Not much has changed. My days are already busy with school, work, and spending time with friends. I'm also trying to make more time for my art. My routine helps me manage stress: I wake up early, do chores, work until noon, then cook lunch. In the afternoon I go to school or work, and in the evening I study until about 7 pm. After that, I have free time—usually for my art or hanging out with friends. Adding more leisure activities would help me feel less stressed, but overall the structure keeps me balanced.

Question: How would you describe your political views?

Answer: I'm not very interested in politics. I usually go along with what my parents—who are Democrats—believe. I think it's important to have some kind of government, and I'm grateful for the opportunities I have in this country, but politics isn't a major focus in my life.

Question: How have you been thinking about race in the U.S. recently?

Answer: I don't think about race very much. I've always had friends of different races and believe everyone should be treated equally. I dislike racist jokes and think it's ignorant when people feel superior because of their race. In my view, race shouldn't be a barrier to success in this country.

Question: For you, what makes it easy or hard to stay healthy?

Answer: Staying healthy isn't too difficult for me. I'm fairly active and try to eat well. I've never had major health problems, though I sometimes feel self-conscious about my weight since I'm tall and heavier than average. Being active helps me feel healthy and confident.

Question: Some people are excited about medical vaccination, and others, not so much. How about you?

Answer: I'm not entirely sure how I feel about vaccinations. I know they're important for preventing illness—I'm glad I was vaccinated as a child—but I have friends who think they're unsafe. I don't believe vaccines are dangerous, but I'm not fully convinced either. I think vaccination should be a personal choice rather than a requirement, though more research could help clarify safety concerns.

Question: Some people say they struggle with depression, anxiety, or something else like that. How about for you?

Answer: I think I might have some depression or something similar. Sometimes I feel really down and sad without knowing why. Stress causes headaches and a racing heart, and sometimes I feel like I might have a panic attack. Talking to people helps a little, but most of the time I prefer to be alone. I'm not sure how to make myself feel better, but I try to manage it as best as I can.

Example Interview-Style Backstory of a *Democrat* User

Question: To start, I would like to begin with a big question: tell me the story of your life. Start from the beginning–from your childhood, to education, to family and relationships, and to any major life events you may have had.

Answer: Sure, I was born in the year 2000 and was raised by a single mother in a small town outside of Seattle. Growing up, my mother worked hard to provide for me and I had to rely heavily on my innate curiosity and self-motivation to pursue my educational and career objectives.

As a child, I always loved gadgets and technology, so I built my first computer when I was just 11 years old. Around that same time, I became interested in aviation and spent hours building and flying model airplanes. When I was 13, I enrolled in an online aviation course and flew my first real airplane at the age of 14. I was hooked.

Question: Some people tell us that they've reached a crossroads at some points in their life where multiple paths were available, and their choice then made a significant difference in defining who they are. What about you? Was there a moment like that for you, and if so, could you tell me the whole story about that from start to finish?

Answer: Sure, I took the red pill when I decided to pursue a career in computer science and artificial intelligence. The choice wasn't necessarily an easy one for me since I had always been intrigued by aviation and even had my pilot's license at a young age. However, after taking a year off between high school and college, I realized that my true passion lies in computer science and artificial intelligence and decided to pursue it wholeheartedly.

This decision has shaped my professional path, leading me to get a bachelor's degree in computer science from a prestigious university and work on cutting-edge projects on machine learning and NLP. It also led me to start my own business providing AI consulting services to clients worldwide. The

choice to pursue AI has been one of the best decisions of my life, and I have enjoyed every moment of my career journey so far.

Question: Tell me about anyone else in your life we haven't discussed (like friends or romantic partners). Are there people outside of your family who are important to you?

Answer: Sure, I have a group of close friends whom I met during college. We all had a similar passion for technology, and we bonded over late-night study sessions and group projects. We support each other's passions and goals and help each other navigate challenges while pursing our dreams. I also have a fiancé whom I met while working on an AI project. She is also a software engineer and has been instrumental in shaping my career decisions. Together, we work on building machine learning models and experimenting with new libraries and tools. She has been an amazing partner, and I cannot imagine my life without her support and love.

Question: Now let's talk about your current neighborhood. Tell me all about the neighborhood and area in which you are living now.

Answer: Currently, I live in a quiet neighborhood near downtown Seattle, Washington. The neighborhood is conveniently located near public transportation, making it easier for me to travel around the city without having to drive a car. There are many parks nearby that offer opportunities for outdoor activities like hiking and biking. The area has good schools, and we are living in a classic 1920s Tudor house with a few modern amenities added.

The community is highly diverse and has a bustling cultural scene with local festivals, art galleries, music venues, and theaters. The area also offers a great selection of shopping, dining, and entertainment opportunities, and there are a variety of local businesses that keep the community thriving. Overall, it's a great place to live and work, and I feel lucky to be a part of it.

Question: Tell me about any recent changes to your daily routine.

Answer: Sure, along with starting a new job this month, I have recently started a new routine for exercising. Now I go for a brisk walk early in the morning before getting to work. In addition, I have also started cooking dinner every night, which has led to me eating more homemade, healthy meals.

After a grueling day at work, I unwind by indulging in video games and creative writing or watching my favorite shows and movies on Netflix or Disney+. Before bed, I also like to catch up on news and social media on my phone, and I usually read a chapter of a book before sleeping. All in all, it's great to have a healthy routine in place and keep a balance between work, health, and leisure. By incorporating these small changes, I feel more balanced and productive in my daily life.

Question: How would you describe your political views?

Answer: I identify as a conservative and believe in protecting our nation's security through strong foreign policies. I am also in favor of lower taxes so as to stimulate the economy and job creation, while also considering the needs of less fortunate members of our community. As a Christian, I have a deep sense of empathy and compassion for those who are in need of assistance, which has instilled a sense of civic duty in me. Generally, I our political views should be fluid and evolving our society develops. Our party differences should not get in the way of achieving our shared goal of contributing to our nation's growth and prosperity for all Americans.

Question: How have you been thinking about race in the U.S. recently?

Answer: It has become increasingly disheartening to witness racial injustice and police brutality against black Americans in our country. The protests and riots that have been sweeping across the nation are a natural response to the frustration that African Americans have endured for centuries. I

am encouraged by the strong support coming from both sides of the aisle and believe that by coming together, we can finally address the systemic injustices that have long plagued our community. I remain optimistic that meaningful change is possible if bipartisan efforts and solutions are implemented to create more equitable societies for everyone within our country. The journey to a more fair and just society is fraught with uncertainties and setbacks. But we should all be encouraged by the sincere efforts we see across our nation to build bridges and seek reforms. Together, we can create a better America for all.

Question: For you, what makes it easy or hard to stay healthy?

Answer: For me, one of the biggest factors in maintaining good health is managing stress. The daily grind of work, family obligations, and maintaining a household can take their toll, making it difficult to prioritize health and wellness. Additionally, I have a busy work schedule, leaving less time for physical activities or meal preps, which can interfere with a healthy lifestyle. I try to incorporate relaxation practices into my routine to help keep stress levels low and ensure that I have enough time to take my health seriously. Other factors like diet, exercise, and sleep also play crucial roles; however, my focus has primarily been on managing stressors and finding ways to make the most of limited leisure time so that my health does not suffer.

Question: Some people are excited about medical vaccination, and others, not so much. How about you?

Answer: absolutely, I am very excited about medical vaccinations. I believe it is a critical step in preventing spread of contagious diseases. Vaccines are designed to stimulate the body's immune system to create antibodies that protect us from future infections from certain diseases. They have already saved countless lives, especially children, from potentially deadly diseases like measles, rubella, Hib disease, etc. So, yes, without a doubt, I think getting vaccinated against these harmful pathogens should be mandatory and taken very seriously by everyone, especially parents to safeguard their child's health. In short, I am a strong advocate for medical vaccinations and encourage everyone around me to prioritize the health and safety of their children by getting them vaccinated.

Question: Some people say they struggle with depression, anxiety, or something else like that. How about for you?

Answer: Of course, I have my moments like everyone else; however, I manage my emotions to the best of my abilities. I have never been diagnosed with clinical depression or anxiety, but I don't shy away from seeking help when needed.

I have experienced some mental health challenges, such as following the death of my father, but in the long run, those were temporary setbacks. As a coping mechanism, I try to stay grounded and mindful of my feelings by engaging in hobbies such as running, hiking, and exploring new places. This practice has allowed me to manage my mental health while keeping stress levels low and enjoying life to its fullest.

C.2 Factorial ANOVA Decomposition of Contextual Effects

This appendix provides a detailed explanation of the factorial ANOVA used to decompose contextual differences among political Dictator and Trust Game experiments provided in Section 4.5. The goal is to estimate how much of the variation in partisan bias (Δ) is

Table C.1: Abridged set of interview questions used to generate transcript-style backstories. These prompts are adapted from oral history protocols developed by the American Voices Project [147], covering key themes such as family, education, work, health, politics, and community life.

Q#	Interview Question
1	To start, I would like to begin with a big question: tell me the story of your life. Start from the beginning–from your childhood, to education, to family and relationships, and to any major life events you may have had.
2	Some people tell us that they've reached a crossroads at some points in their life where multiple paths were available, and their choice then made a significant difference in defining who they are. What about you? Was there a moment like that for you, and if so, could you tell me the whole story about that from start to finish?
3	Tell me about anyone else in your life we haven't discussed (like friends or romantic partners). Are there people outside of your family who are important to you?
4	Now let's talk about your current neighborhood. Tell me all about the neighborhood and area in which you are living now.
5	Tell me about any recent changes to your daily routine.
6	How would you describe your political views?
7	How have you been thinking about race in the U.S. recently?
8	For you, what makes it easy or hard to stay healthy?
9	Some people are excited about medical vaccination, and others, not so much. How about you?
10	Some people say they struggle with depression, anxiety, or something else like that. How about for you?

attributable to differences in: (1) the **subject pool**, (2) the **framing text**, and (3) the **year** of the study.

LLM-based counterfactual simulation allows us to evaluate all possible combinations of these experimental components, enabling a full $2 \times 2 \times 2$ factorial analysis that is impossible using human experiments alone.

Factor Structure

We analyze three experimental factors that jointly determine the structure of each study. For clarity, we introduce the following notation for the four source studies:

- **ID** = Iyengar & Westwood (Dictator Game; 2014) [72]
- **WD** = Whitt et al. (Dictator Game; 2019) [162]
- **CT** = Carlin & Love (Trust Game; 2015) [28]
- **WT** = Whitt et al. (Trust Game; 2019) [162]

Each game (Dictator or Trust) can be described by the same set of three binary factors:

- **Subject Pool**. This determines which population provided the human responses.

  - For the Dictator Game: ID vs. WD
  - For the Trust Game: CT vs. WT

- **Framing Text**. This is the wording of the instructions shown to participants.
  - Dictator Game framings: ID-style wording vs. WD-style wording
  - Trust Game framings: CT-style wording vs. WT-style wording

- **Study Year**. This captures temporal differences in the political context surrounding each experiment.
  - Early studies: 2014 (ID) and 2015 (CT)
  - Later studies: 2019 (WD, WT)

Outcome Variable

Each cell reports an **average partisan attitude gap**:

$$\Delta = \tfrac{1}{2}\big(\Delta_{\text{Dem}} + \Delta_{\text{Rep}}\big),$$

which we treat as the dependent variable.

Main Effects in Factorial ANOVA

For a binary factor A with levels a_1 and a_2, the **main effect** is defined as the difference in the average outcome between the two levels, averaging over all combinations of the other factors:

$$\text{Effect}(A) = \mathbb{E}[Y \mid A = a_2] - \mathbb{E}[Y \mid A = a_1].$$

Here, the expectation is computed simply as the arithmetic mean of the corresponding four table entries (because each factor has two levels).

Example (Dictator Game, Year Effect). The entries using Year=2014 (four rows) have average:

$$\bar{Y}_{2014} = \frac{0.72 + 1.77 + 0.79 + 1.78}{4} = 1.265.$$

The entries using Year=2019 have average:

$$\bar{Y}_{2019} = \frac{1.12 + 2.29 + 0.87 + 2.38}{4} = 1.665.$$

Thus the *Year main effect* is:

$$1.665 - 1.265 = 0.40.$$

Interaction Effects

Two factors interact when the effect of one depends on the level of the other. For example, the *Population × Framing* interaction compares how much the framing effect changes when the subject pool switches from ID/CT to WD/WT.

Following standard two-factor contrasts, each interaction is computed as the difference of differences averaging across the third factor.

C.3 ANOVA Results: Dictator Game

Main Effects

- **Subject Pool (ID → WD):** The four ID rows average to 1.475, and the four WD rows average to 1.455. The subject pool effect is therefore **-0.02**, indicating a negligible difference between the Iyengar and Whitt subject populations.
- **Framing (Iyengar → Whitt):** Iyengar-style framing averages to 0.875, while Whitt-style framing averages to 2.055. This yields a large framing effect of **+1.18**, the dominant source of variation in the Dictator Game.
- **Year (2014 → 2019):** As shown above, the year effect is **+0.40**.

Interaction Effects

All interaction terms are small relative to the framing effect:

- Population × Framing: +0.07
- Population × Year: -0.06
- Framing × Year: +0.16
- Three-way interaction: +0.10

C.4 ANOVA Results: Trust Game

Main Effects

- **Subject Pool (CT → WT):** CT rows average to 1.2625; WT rows average to 1.535. The subject pool effect is **+0.27**, meaning Whitt's nationally representative sample exhibits larger partisan gaps.
- **Framing (Carlin → Whitt):** Carlin-style framing averages to 1.19, Whitt-style framing to 1.6075. The framing effect is **+0.42**.

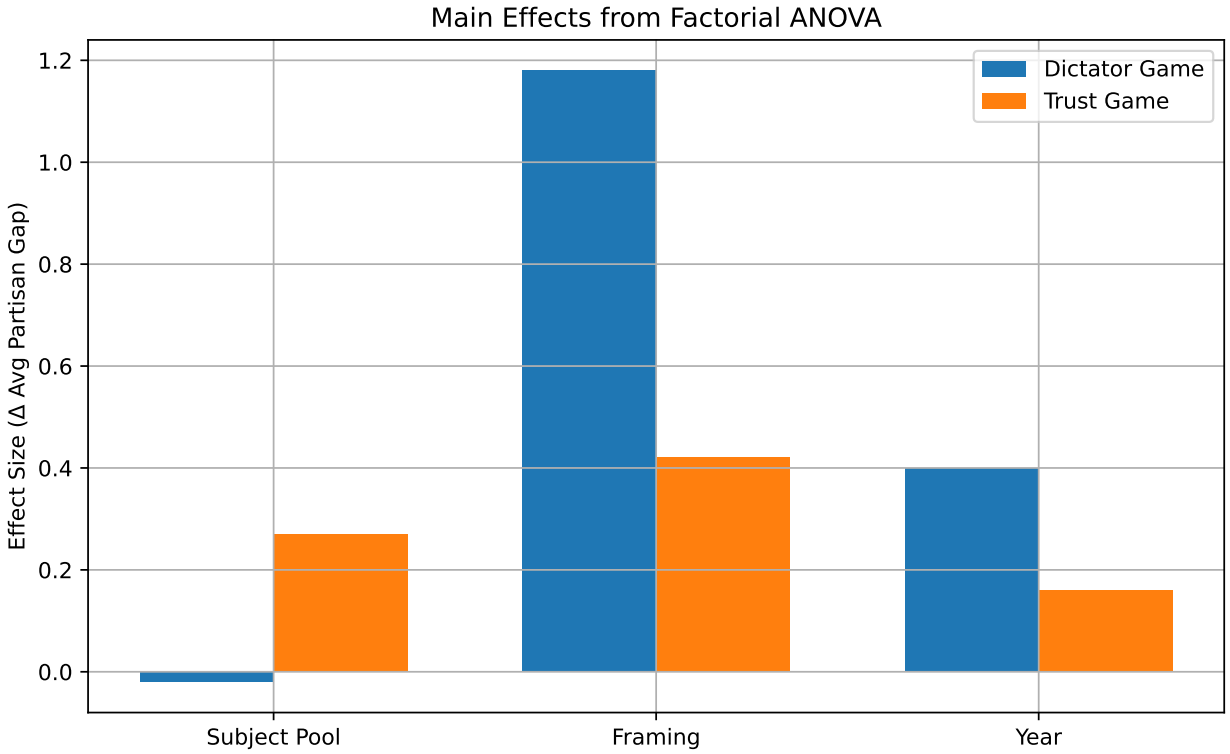

Figure C.1: **Main Effects from the Factorial ANOVA.**

- **Year (2015 → 2019):** The year effect is **+0.16**.

Interaction Effects

Interaction terms remain small:

- Population × Framing: -0.13
- Population × Year: -0.01
- Framing × Year: +0.03
- Three-way interaction: -0.04

C.5 Summary Across Games

Figure C.1 summarizes the main effects for both Dictator Game and Trust Game. Framing is the largest contextual factor in both games. Year provides a moderate positive shift. Subject Pool has minimal effect in the Dictator Game but a meaningful effect in the Trust

Game. LLM-based counterfactual simulation enables this full decomposition by populating all $2 \times 2 \times 2$ combinations of experimental components, which human experiments cannot practically achieve.

C.6 Demographic Survey

After generating backstories through open-ended narrative sampling, we emulate the process of collecting sociodemographic and ideological information by administering a structured survey to each virtual persona [113, 79]. This step is essential for curating a pool of virtual personas whose aggregate trait distribution closely mirrors that of the human population we aim to simulate. The resulting demographic metadata is later used in the user–persona matching process described in Section C.7.

Unlike human respondents, who each possess a fixed and known set of demographic and ideological characteristics, virtual personas do not necessarily exhibit a unique or explicit trait profile. A single backstory may implicitly represent a range of possible individuals unless specific details are directly stated (e.g., "I am a 30-year-old woman"). As a result, we represent each persona's traits using a probability distribution rather than a deterministic vector.

We adopt a two-stage procedure for constructing these distributions, following the approach of [113, 79]. In the first stage, we attempt to extract explicit trait mentions directly from the narrative. To do this, we prompt `gpt-4o` (with temperature `T = 0`) using a set of targeted trait-identification questions and the backstory as input. If the model finds clear verbal evidence for a given trait (e.g., "I'm a proud Democrat"), we assign a one-hot distribution to that trait category. If no explicit mention is found, we proceed to the second stage, where the model infers a probability distribution over the possible trait values.

Below, we provide the full list of trait-seeking prompts used to elicit evidence for six key sociodemographic and ideological attributes.

C.7 Demographic Matching

To approximate survey responses using synthetic personas, we assign each real survey participant to a virtual backstory that closely mirrors their demographic and ideological profile. We model this assignment process as a complete weighted bipartite graph $G = (H, V, E)$, where $H = \{h_1, \ldots, h_n\}$ denotes the set of n human users, and $V = \{v_1, \ldots, v_m\}$ is a larger pool of m candidate virtual personas, with $m > n$ to allow for flexible and diverse matching.

Each human user h_i is described by a fixed k-dimensional tuple of categorical traits, $(t_{i1}, \ldots, t_{ik})$, spanning demographic and ideological characteristics. Each virtual persona v_j, in contrast, is associated with a probability distribution over these k traits:

$$(P(d_{j1}),\ P(d_{j2}),\ \ldots,\ P(d_{jk}))$$

The weight of the edge $e_{ij} \in E$, connecting human user h_i to persona v_j, represents the likelihood that v_j matches all of h_i's traits, and is defined as:

$$w(e_{ij}) = \prod_{l=1}^{k} P(d_{jl} = t_{il}) \tag{C.1}$$

We then apply a maximum-weight bipartite matching algorithm to find an injective mapping $\pi : [n] \to [m]$ that maximizes the total matching score:

$$\sum_{i=1}^{n} w(h_i, v_{\pi(i)})$$

This optimization is solved exactly using the Hungarian algorithm [89]. After matching, we select n virtual backstories whose traits collectively reflect the demographic distribution of the original human sample, ensuring that our synthetic population maintains fidelity to the target survey group.