Generative Agent Simulations of 1,000 People

Authors: Joon Sung Park[1]*, Carolyn Q. Zou[1,2], Aaron Shaw[2], Benjamin Mako Hill[3], Carrie Cai[4], Meredith Ringel Morris[5], Robb Willer[6], Percy Liang[1], Michael S. Bernstein[1]

Affiliations:

[1]Computer Science Department, Stanford University; Stanford, CA, 94305, USA.

[2]Department of Communication Studies, Northwestern University; Evanston, IL, 60208, USA.

[3]Department of Communication, University of Washington; Seattle, WA 98195, USA.

[4]Google DeepMind; Mountain View, CA 94043, USA.

[5]Google DeepMind; Seattle, WA 98195, USA.

[6]Department of Sociology, Stanford University; Stanford, CA, 94305, USA.

*Corresponding author. Email: joonspk@stanford.edu

Abstract:

The promise of human behavioral simulation—general-purpose computational agents that replicate human behavior across domains—could enable broad applications in policymaking and social science. We present a novel agent architecture that simulates the attitudes and behaviors of 1,052 real individuals—applying large language models to qualitative interviews about their lives, then measuring how well these agents replicate the attitudes and behaviors of the individuals that they represent. The generative agents replicate participants' responses on the General Social Survey 85% as accurately as participants replicate their own answers two weeks later, and perform comparably in predicting personality traits and outcomes in experimental replications. Our architecture reduces accuracy biases across racial and ideological groups compared to agents given demographic descriptions. This work provides a foundation for new tools that can help investigate individual and collective behavior.

Main Text: General-purpose simulation of human attitudes and behavior—where each simulated person can engage across a range of social, political, or informational contexts—could enable a laboratory for researchers to test a broad set of interventions and theories *(1-3)*. How might, for instance, a diverse set of individuals respond to new public health policies and messages, react to product launches, or respond to major shocks? When simulated individuals are combined into collectives, these simulations could help pilot interventions, develop complex theories capturing nuanced causal and contextual interactions, and expand our understanding of structures like institutions and networks across domains such as economics (*4*), sociology (*2*), organizations (*5*), and political science (*6*).

Simulations define models of individuals that are referred to as agents (*7*). Traditional agent architectures typically rely on manually specified behaviors, as seen in agent-based models (*1, 8, 9*), game theory (*10*), and discrete choice models (*11*), prioritizing interpretability at the cost of restricting agents to narrow contexts and oversimplifying the contingencies of real human behavior (*3, 4*). Generative artificial intelligence (AI) models, particularly large language models (LLMs) that encapsulate broad knowledge of human behavior (*12-15*), offer a different opportunity: constructing an architecture that can accurately simulate behavior across many contexts. However, such an approach needs to avoid flattening agents into demographic stereotypes, and measurement needs to advance beyond replication success or failure on average treatment effects (*16-19*).

We present a *generative agent* architecture that simulates more than 1,000 real individuals using two-hour qualitative interviews. The architecture combines these interviews with a large language model to replicate individuals' attitudes and behaviors. By anchoring on individuals, we can measure accuracy by comparing simulated attitudes and behaviors to the actual attitudes and behaviors. We benchmark these agents using canonical social science measures such as the General Social Survey (GSS; *20*), the Big Five Personality Inventory (*21*), five well-known behavioral economic games (e.g., the dictator game, a public goods game) (*22-25*), and five social science experiments with control and treatment conditions that we sampled from a recent large-scale replication effort (*26-31*). To support further research while protecting participant privacy, we provide a two-pronged access system to the resulting *agent bank*: open access to aggregated responses on fixed tasks for general research use, and restricted access to individual responses on open tasks for researchers following a review process, ensuring the agents are accessible while minimizing risks associated with the source interviews.

Creating 1,000 Generative Agents of Real People

To create simulations that better reflect the myriad, often idiosyncratic, factors that influence individuals' attitudes, beliefs, and behaviors, we turn to in-depth interviews—a method that previous work on predicting human life outcomes has employed to capture insights beyond what can be obtained through traditional surveys and demographic instruments (*32*). In-depth interviews, which combine pre-specified questions with adaptive follow-up questions based on respondents' answers, are a foundational social science method with several advantages over more structured data collection techniques (*33, 34*). While surveys with closed-ended questions and predefined response categories are valuable for well-powered quantitative analysis and hypothesis testing, semi-structured interviews offer distinct benefits for gaining idiographic knowledge about individuals. Most notably, they give interviewees more freedom to highlight what they find important, ultimately shaping what is measured.

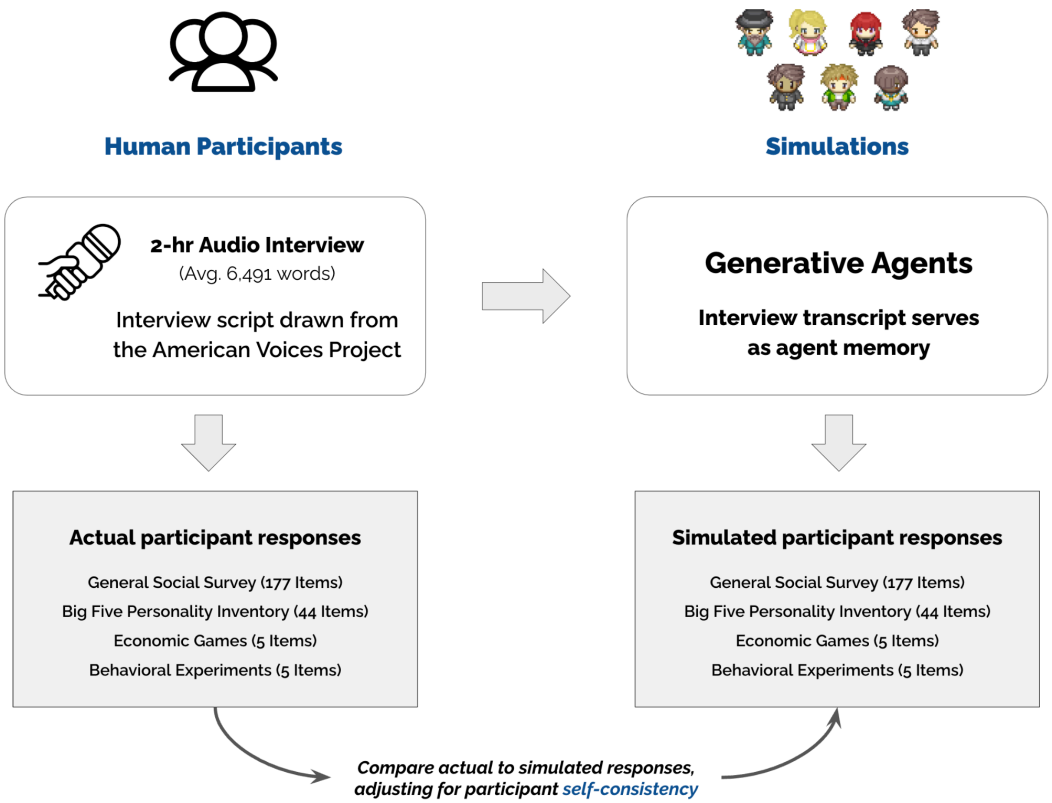


Figure 1. The process of collecting participant data and creating generative agents begins by recruiting a stratified sample of 1,052 individuals from the U.S., selected based on age, census division, education, ethnicity, gender, income, neighborhood, political ideology, and sexual identity. Once recruited, participants complete a two-hour audio interview with our AI interviewer, followed by surveys and experiments. We create generative agents for each participant using their interview data. To evaluate these agents, both the generative agents and participants complete the same surveys and experiments. For the human participants, this involves retaking the surveys and experiments again two weeks later. We assess the accuracy of the agents by comparing agent responses to the participants' original responses, normalizing by how consistently each participant successfully replicates their own responses two weeks later.

We recruited over 1,000 participants using stratified sampling to create a representative U.S. sample across age, gender, race, region, education, and political ideology. Each participant completed a voice-to-voice interview in English, producing transcripts with an average length of 6,491 words per participant (std = 2,541; SM 1). To facilitate this process, we developed an AI interviewer (SM 2) that conducted the interview using a semi-structured interview protocol. To avoid inadvertently tailoring the interview protocol to our evaluation metrics, we sought an existing interview protocol that aimed for broad topical coverage. We selected an interview protocol developed by sociologists as part of the American Voices Project (*35*). The script explored a wide range of topics of interest to social scientists—from participants' life stories (e.g., "Tell me the story of your life—from your childhood, to education, to family and relationships, and to any major life events you may have had") to their views on current societal issues (e.g., "How have you responded to the increased focus on race and/or racism and policing?"; SM 8). Its broad scope, diverging from our metrics (e.g., while some questions overlap thematically with the GSS, they do not directly include specific questions or cover personality traits or economic game behaviors), strengthens results if high performance is achieved. Within the interview's structure and time limitations, the AI interviewer dynamically generated follow-up questions tailored to each participant's responses.

To create the generative agents (*14, 15*), we developed a novel agent architecture that leverages participants' full interview transcripts and a large language model (SM 3). When an agent is queried, the entire interview transcript is injected into the model prompt, instructing the model to imitate the person based on their interview data. For experiments requiring multiple decision-making steps, agents were given memory of previous stimuli and their responses to those stimuli through short text descriptions. The resulting agents can respond to any textual stimulus, including forced-choice prompts, surveys, and multi-stage interactional settings.

We evaluated the generative agents on their ability to predict their source participants' responses to a series of surveys and experiments commonly used across social science disciplines. This evaluation consisted of four components, which participants completed following their interviews: the core module of the General Social Survey (GSS; *20*), the 44-item Big Five Inventory (BFI-44; *16*), five well-known behavioral economic games (including the dictator game, trust game, public goods game, and prisoner's dilemma; *22-25*), and five social science experiments with control and treatment conditions *(27-31)*. The experiments were sampled from a recent large-scale replication effort (*26*), chosen based on criteria that the external replication specified 1,000 participants for sufficient power and that the experiments could be delivered to agents in text form (SM 4). We used the first three components to measure the accuracy of the generative agents in predicting individual attitudes, traits, and behaviors, while the replication studies assessed their ability to predict population-level treatment effects and effect sizes in a well-powered replication. Our metrics and core analyses were pre-registered (SM 5).[1]

A key methodological benefit of simulating specific individuals is the ability to evaluate our architecture by comparing how accurately each agent replicates the attitudes and behaviors of its source individual. For the GSS, where responses are categorical, we measure accuracy and correlation based on whether the agent selects the same survey response as the individual. For the BFI-44 and economic games, which involve continuous responses, we assess accuracy and correlation using mean absolute error (MAE). Since individuals often exhibit inconsistency in their responses over time in both survey and behavioral studies (*32, 36, 37*), we use participants' own attitudinal and behavioral consistency as a normalization factor: the probability of accurately simulating an individual's attitudes or behaviors depends on how consistent those attitudes and behaviors are over time.

To account for these varying levels in self-consistency, we asked each participant to complete our battery twice, two weeks apart. Our main dependent variable is *normalized accuracy*, calculated as the agent's accuracy in predicting the individual's responses divided by the individual's own replication accuracy. A normalized accuracy of 1.0 indicates that the generative agent predicts the individual's responses as accurately as the individual replicates their own responses two weeks later. For continuous outcomes, we calculate normalized correlation instead.

[1] Pre-registration materials: https://osf.io/mexkf/?view_only=375fe67b9a3e48afa7c3684c9d344da4

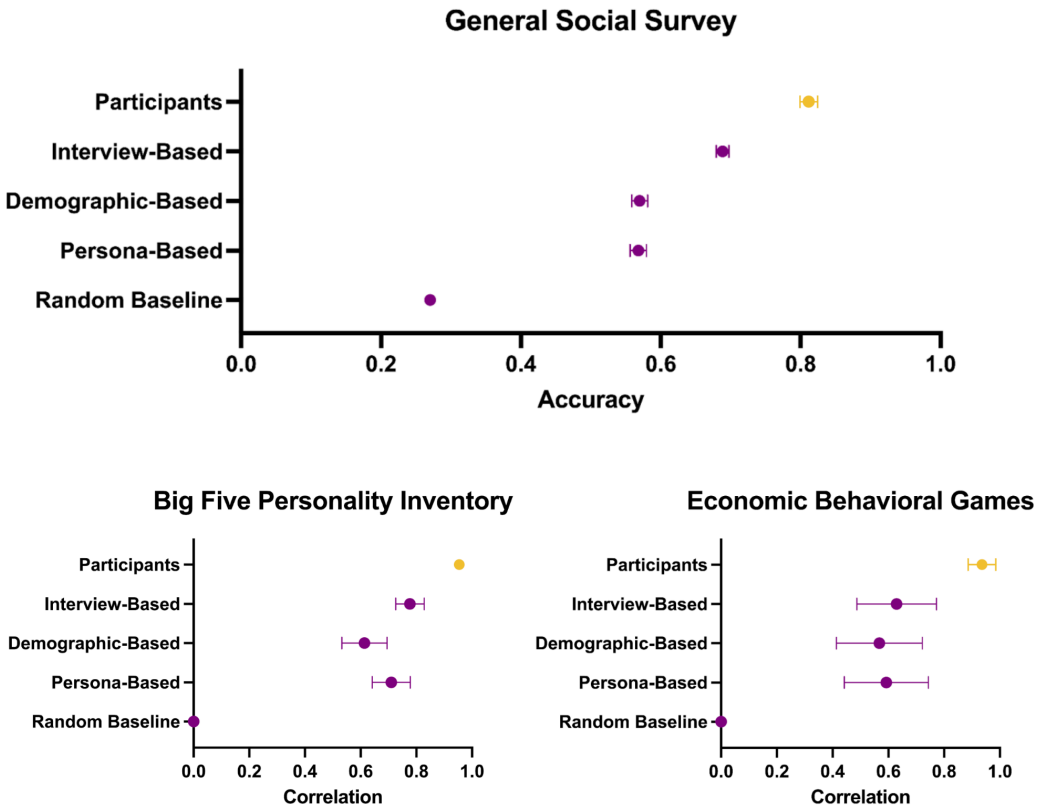


Figure 2. Generative agents' predictive performance, and 95% confidence intervals. The consistency rate between participants and the predictive performance of generative agents is evaluated across various constructs and averaged across individuals. For the General Social Survey (GSS), accuracy is reported due to its categorical response types, while the Big Five personality traits and economic games report mean absolute error (MAE) due to their numerical response types. Correlation is reported for all constructs. Normalized accuracy is provided for all metrics, except for MAE, which cannot be calculated for individuals whose MAE is 0 (i.e., those who responded the same way in both phases). We find that generative agents predict participants' behavior and attitudes well, especially when compared to participants' own rate of internal consistency. Additionally, using interviews to inform agent behavior significantly improves the predictive performance of agents for both GSS and Big Five constructs, outperforming other commonly used methods in the literature.

Predicting Individuals' Attitudes and Behaviors

To assess the contribution of interviews to the generative agents' predictive accuracy, we compared the performance of interview-based generative agents with two baselines that replace interview transcripts with alternative forms of description. These baselines are grounded in how language models have been used to proxy human behaviors in prior studies: one using demographic attributes (*13, 38*), and the other using a paragraph summarizing the target person's profile (*14*). For the demographic-based generative agents, we used participants' responses to GSS questions to capture individuals' age, gender, race, and political ideology—demographic attributes commonly used in previous studies (*38*). For the persona-based generative agents, we asked participants to write a brief paragraph about themselves after the interview, including their personal background, personality, and demographic details, similar to the material used to generate persona agents in prior work (*14*).

The first component of our evaluation, the GSS, is widely used across sociology, political science, social psychology, and other social sciences to assess respondents' demographic backgrounds, behaviors, attitudes, and beliefs on a broad range of topics, including public policy, race relations, gender roles, and religion (*20*). Our evaluation focused on 177 core GSS

questions, which we used to establish a benchmark for measuring the agents' predictive accuracy. Each question had an average of 3.70 response options (std = 2.22), yielding a random chance prediction accuracy of 27.03%.

For the GSS, the generative agents predicted participants' responses with an average normalized accuracy of 0.85 (std = 0.11), calculated from a raw accuracy of 68.85% (std = 6.01) divided by participants' replication accuracy of 81.25% (std = 8.11). These interview-based agents significantly outperformed both demographic-based and persona-based agents (Figure 2), with a margin of 14-15 normalized points. The demographic-based generative agents achieved a normalized accuracy of 0.71 (std = 0.11), while persona-based agents reached 0.70 (std = 0.11). An ANOVA of the accuracy rates rejected the null hypothesis of no significant difference ($F(2, 3153) = 989.62$, $p < 0.001$), and post-hoc pairwise Tukey tests confirmed that the interview-based agents outperformed the other two groups.

The second component of our evaluation focused on predicting participants' Big Five personality traits using the BFI-44, which assesses five personality dimensions: openness, conscientiousness, extraversion, agreeableness, and neuroticism (*21*). Each dimension is calculated as an aggregate of eight to ten Likert scale questions. Our generative agents predicted participants' responses to the individual items, which were then used to compute the predicted aggregate scores for each personality dimension. These are continuous measures, so we calculated correlation coefficients and normalized correlations.

For the Big Five, the generative agents achieved a normalized correlation of 0.80 (std = 1.88), based on a raw correlation of $r = 0.78$ (std = 0.70) divided by participants' replication correlation of 0.95 (std = 0.76). As with the GSS, the interview-based generative agents outperformed both demographic-based (normalized correlation = 0.55) and persona-based (normalized correlation = 0.75) agents. The interview-based agents also produced predictions with lower MAE for Big Five personality traits ($F(2, 3153) = 25.96$, $p < 0.001$), and post-hoc pairwise Tukey tests confirmed that interview-based agents significantly outperformed the other two groups.

The third component involved a series of five well-known economic games designed to elicit participants' behaviors in decision-making contexts with real stakes. These included the Dictator Game, the first and second player Trust Games, the Public Goods Game, and the Prisoner's Dilemma (*22-25*). To ensure genuine engagement, participants were offered monetary incentives. We standardized the output values for each game on a scale from 0 to 1 and compared the generative agents' predicted values to the actual values obtained from participants. Since these are continuous measures, we calculated correlation coefficients and normalized correlations. On average, the generative agents achieved a normalized correlation of 0.66 (std = 2.83), derived from a raw correlation of $r = 0.66$ (std = 0.95) divided by participants' replication correlation of 0.99 (std = 1.00). However, there was no significant difference in MAE between the agents for the economic games ($F(2, 3153) = 0.12$, $p = 0.89$).

In exploratory analyses, we tested the effectiveness and efficiency of interviews by comparing interview-based generative agents to a baseline composite agent informed by participants' GSS, Big Five, and economic game responses. We randomly sampled 100 participants and created composite agents from their responses to these instruments. To prevent exact answer retrieval, we excluded all question-answer pairs from the same category as the question being predicted (categories were defined by the creators of each instrument), which excluded an average of 4.00% (std = 2.16). This composite agent serves as a baseline with access to semantically close information to the evaluation, so any performance gap with the interview-based agents would

indicate the interview's unique effectiveness in capturing participant identity. On average, the composite generative agents achieved a normalized accuracy of 0.76 (std = 0.12) for the GSS, a normalized correlation of 0.64 (std = 0.61) for the Big Five, and 0.31 (std = 1.22) for economic games. These results still underperformed the interview-based generative agents.

We conducted additional tests by ablating portions of the generative agents' interviews to examine the impact of interview content volume and style. First, even when we randomly removed 80% of the interview transcript—equivalent to removing 96 minutes of the 120-minute interview—the interview-based generative agents still outperformed the composite agents, achieving an average normalized accuracy of 0.79 (std = 0.11) on the GSS, with similar results observed for the Big Five. Second, to investigate whether the predictive power of interviews stems from linguistic cues or the richness of the knowledge gained, we created "interview-summary" generative agents by prompting GPT-4o to convert interview transcripts into bullet-pointed summaries of key response pairs, capturing the factual content while removing the original linguistic features. These agents also outperformed composite agents, achieving a normalized accuracy of 0.83 (std = 0.12) on the GSS and showing similar improvements for the Big Five. These findings suggest that, when informing language models about human behavior, interviews are more effective and efficient than survey-based methods.

	Human replication		**Agent prediction**					
	Participants		***Interview***		***Demog. Info.***		***Persona Desc.***	
Replication Studies	***p***	**Effect size**	***p***	**Effect size**	***p***	**Effect size**	***p***	**Effect size**
Ames & Fiske 2015	***	9.45	***	12.59	***	13.43	***	10.03
Cooney et al. 2016	***	0.40	***	1.48	***	1.39	***	1.37
Halevy & Halali 2015	***	0.90	***	2.98	***	4.22	***	3.35
Rai et al. 2017		0.040		0.094	***	0.21		0.078
Schilke et al. 2015	***	0.33	***	2.97	***	5.52	***	3.74
Effect size correlation w/ human rep.				**Correlation**		**Correlation**		**Correlation**
				r = 0.98		**r = 0.93**		**r = 0.94**
				95% CI [0.74, 0.99]		95% CI [0.24, 0.99]		95% CI [0.33, 0.99]

Table 1. Results of replication studies by human participants and generative agents. We report the p-values (***: < 0.001, **: < 0.01, *: < 0.05) and Cohen's d for effect sizes. Our replication with human participants replicated four out of five studies, while generative agents informed by the interview transcript replicated the same four studies. The correlation of the effect sizes between the human participants and generative agents achieved a strong correlation.

Predicting Experimental Replications

Participants took part in five social science experiments to assess whether generative agents can predict treatment effects in experimental settings commonly used by social scientists. These were drawn from a collection of published studies included in a large-scale replication effort (*26-31*;

SM 4), including investigations of how perceived intent affects blame assignment (*27*) and how fairness influences emotional responses (*28*). Both human participants in our work and generative agents completed all five studies, with p-values and treatment effect sizes calculated using the statistical methods as the original studies. Our participants successfully replicated the results of four out of the five studies, failing to replicate one; the generative agents replicated the same four studies and failed to replicate the fifth. The effect sizes estimated from the generative agents were highly correlated with those of the participants (r = 0.98), compared to the participants' internal consistency correlation of 0.99, yielding a normalized correlation of 0.99.

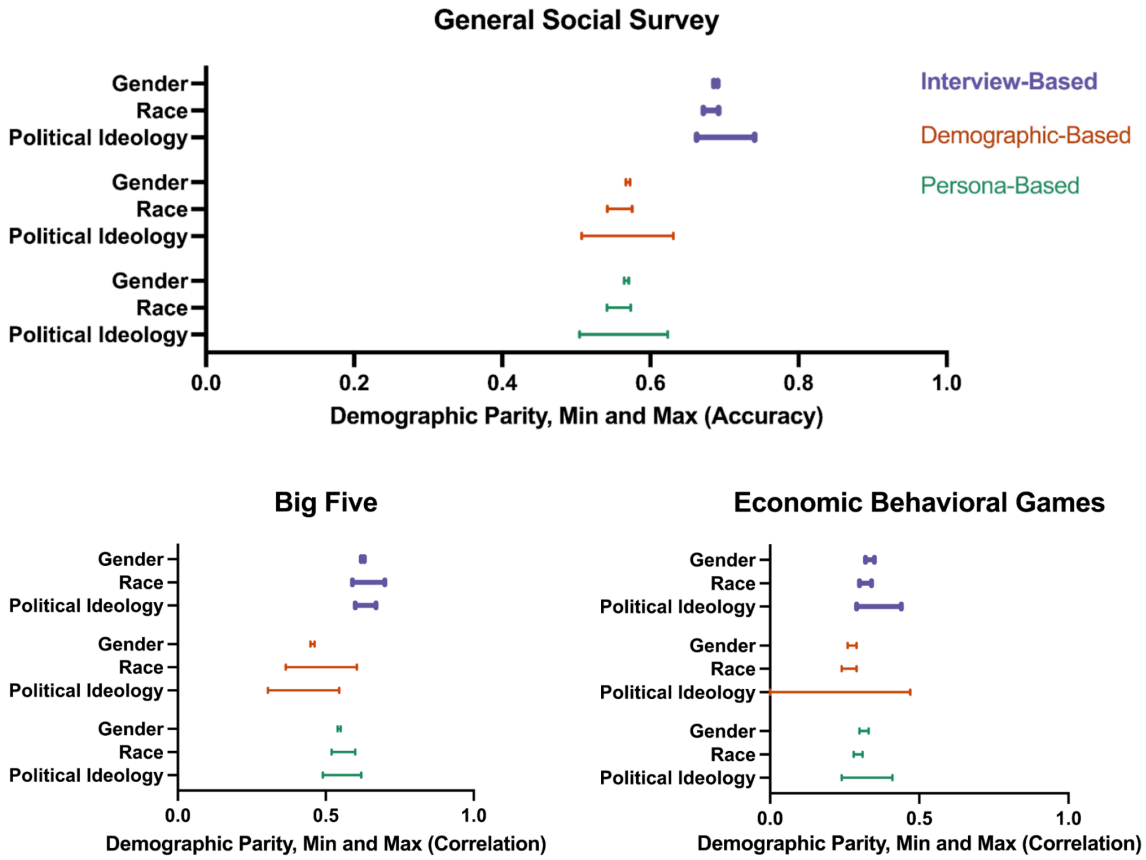


Figure 3. Demographic Parity Difference (DPD) for generative agents across political ideology, race, and gender subgroups on three tasks: GSS (in percentages), Big Five, and economic games (in correlation coefficients). DPD represents the performance disparity between the most and least favored groups within each demographic category. Generative agents using interviews consistently show lower DPDs compared to those using demographic information or persona descriptions, suggesting that interview-based generative agents mitigate bias more effectively across all tasks. Gender-based DPDs remain relatively low and consistent across all conditions.

Interviews Reduce Bias in Generative Agent Accuracy

There is concern about AI systems underperforming or misrepresenting underrepresented populations *(19)*. To address this concern, we conducted a subgroup analysis focusing on political ideology, race, and gender—dimensions of particular interest in relevant literature *(13, 38, 16-18)*. We aimed to assess whether the in-depth descriptions provided by interviews could mitigate biases compared to methods using demographic prompts, which exhibited stereotyping in prior research *(16-19)*. We quantified bias using the Demographic Parity Difference (DPD), which measures the difference in performance between the best performing and worst-performing groups *(39, 40)*. For the GSS, we report DPD in percentages; for Big Five and

economic games, in correlation coefficients. Subgroups were defined by participants' responses to GSS items (details in SM 5).

Interview-based agents consistently reduced biases across tasks compared to demographic-based agents. For political ideology, we observed that in the GSS, the DPD dropped from 12.35% for demographic-based generative agents to 7.85% for interview-based generative agents. In the Big Five personality traits, the DPD dropped from 0.165 to 0.063 (in correlation coefficients), and in economic games, it dropped from 0.50 to 0.19 (in correlation coefficients). Although initial racial subgroup discrepancies were smaller with demographic-based generative agents than the interview-based generative agents, interview-based generative agents still reduced them further: in the GSS, the DPD decreased from 3.33 to 2.08%; in the Big Five, from 0.17 to 0.11 correlation coefficients; and in economic games, from 0.043 to 0.040 correlation coefficients. Gender-based DPD remained relatively constant across tasks, likely due to its already low level of discrepancy.

Research Access for the Agent Bank

Access to an agent bank can help lay the foundations for replicable science using AI-based tools. Our agent bank of 1,000 generative agents offers a resource toward these goals. To balance scientific potential with privacy concerns, the authors at Stanford University provide a two-pronged access system for research: open access to aggregated responses on fixed tasks (e.g., GSS) and restricted access to individualized responses on open tasks. Safeguards include usage audits, participant withdrawal options, and non-commercial use agreements, modeled after genome banks and AI model deployments, supporting ethical research and reducing risk to human subjects while enabling AI applications in the social sciences.[2]

Materials and Methods Summary

We contracted with the recruitment firm Bovitz (*41*) to obtain a U.S. sample of 1,000 individuals, stratified by age, census division, education, ethnicity, gender, income, neighborhood, political ideology, and sexual orientation. Participants completed interviews with the AI interviewer, along with Qualtrics versions of the General Social Survey (GSS), Big Five personality inventory, economic games, and selected experimental studies. For the GSS, we focused on 177 questions for the “core” module, excluding non-categorical questions, questions with more than 25 response options, and conditional questions. For the experimental studies, we selected five studies from a recent large-scale replication effort (*26-31*). These were chosen based on two inclusion criteria: first, the study had to be describable to a language model using text or images, and second, the power analysis from the replication effort indicated that the effects would be observable with 1,000 or fewer participants. This ensured that our human participants could replicate the effects if present. The selected studies (*27-31*) covered the evaluation of harm based on perceived intent, the role of fairness in emotional reactions, the perceived benefits of conflict intervention, dehumanization in willingness to harm others, and how power influences trust.

[2] The codebase for generating agent behavior is available as an open-source repository. Researchers interested in constructing agents with their own data can access it here: https://github.com/joonspk-research/generative_agent

References

1. E. Bruch, J. Atwell, Agent-Based Models in Empirical Social Research. Sociological Methods & Research 44, 186-221 (2015).
2. T. C. Schelling, Dynamic models of segregation. Journal of Mathematical Sociology 1, 143-186 (1971).
3. J. M. Epstein, R. L. Axtell, Growing Artificial Societies: Social Science from the Bottom Up (The MIT Press, 1996).
4. R. Axtell, "Why agents? On the varied motivations for agent computing in the social sciences" (Center on Social and Economic Dynamics Working Paper No. 17, 2000).
5. K. M. Carley, Organizational learning and personnel turnover. Organization Science 3, 20-46 (1992).
6. I. S. Lustick, PS-I: A user-friendly agent-based modeling platform for testing theories of political identity and political stability. Journal of Artificial Societies and Social Simulation 5, 3 (2002).
7. T. C. Schelling, Micromotives and Macrobehavior (W. W. Norton & Company, 1978).
8. E. Bonabeau, Agent-based modeling: Methods and techniques for simulating human systems. Proc. Natl. Acad. Sci. U.S.A. 99 (suppl. 3), 7280-7287 (2002); https://doi.org/10.1073/pnas.082080899
9. M. W. Macy, R. Willer, From Factors to Actors: Computational Sociology and Agent-Based Modeling. Annu. Rev. Sociol. 28, 143-166 (2002); https://doi.org/10.1146/annurev.soc.28.110601.141117
10. J. von Neumann, O. Morgenstern, Theory of Games and Economic Behavior (Princeton University Press, 1944).
11. McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), Frontiers in Econometrics (pp. 105-142). Academic Press.
12. J. J. Horton, "Large language models as simulated economic agents: What can we learn from homo silicus?" (2023).
13. A. Ashokkumar, L. Hewitt, I. Ghezae, R. Willer, "Predicting Results of Social Science Experiments Using Large Language Models" (2024).
14. J. S. Park, L. Popowski, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Social simulacra: Creating Populated Prototypes for Social Computing Systems, in Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (ACM, 2022).
15. J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, in Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (ACM, 2023).
16. M. Cheng, T. Piccardi, D. Yang, in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023) (Association for Computational Linguistics, 2023).
17. A. Wang, J. Morgenstern, J. P. Dickerson, "Large language models cannot replace human participants because they cannot portray identity groups" (2024).
18. S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, T. Hashimoto, in Proceedings of the 40th International Conference on Machine Learning (ICML '23) (PMLR, 2023).
19. L. Messeri, M. J. Crockett, Artificial intelligence and illusions of understanding in scientific research. Nature 627, 49-58 (2024).
20. National Opinion Research Center, "General Social Survey, 2023" (NORC at the University of Chicago, 2023); https://gss.norc.org.

21. O. P. John, S. Srivastava, The Big Five trait taxonomy: History, measurement, and theoretical perspectives, in Handbook of Personality: Theory and Research, L. A. Pervin, O. P. John, Eds. (Guilford Press, ed. 2, 1999), pp. 102-138.
22. R. Forsythe, J. L. Horowitz, N. E. Savin, M. Sefton, Fairness in simple bargaining experiments. Games and Economic Behavior 6, 347-369 (1994).
23. J. Berg, J. Dickhaut, K. McCabe, Trust, reciprocity, and social history. Games and Economic Behavior 10, 122-142 (1995).
24. J. O. Ledyard, in The Handbook of Experimental Economics, J. H. Kagel, A. E. Roth, Eds. (Princeton University Press, 1995), pp. 111-194.
25. A. Rapoport, A. M. Chammah, Prisoner's Dilemma: A Study in Conflict and Cooperation (University of Michigan Press, 1965).
26. C. F. Camerer et al., "Mechanical Turk Replication Project" (2024); https://mtrp.info/index.html.
27. D. L. Ames, S. T. Fiske, Perceived intent motivates people to magnify observed harms. PNAS 112, 3599-3605 (2015).
28. G. Cooney, D. T. Gilbert, T. D. Wilson, When fairness matters less than we expect. PNAS 113, 11168-11171 (2016).
29. N. Halevy, E. Halali, Selfish third parties act as peacemakers by transforming conflicts and promoting cooperation. PNAS 112, 6937-6942 (2015).
30. T. S. Rai, P. Valdesolo, J. Graham, Dehumanization increases instrumental violence, but not moral violence. PNAS 114, 8511-8516 (2017).
31. O. Schilke, M. Reimann, K. S. Cook, Power decreases trust in social exchange. PNAS 112, 12950-12955 (2015).
32. I. Lundberg et al., The origins of unpredictability in life outcome prediction tasks. Proc. Natl. Acad. Sci. U.S.A. 121, e2322973121 (2024).
33. A. Lareau, Listening to People: A Practical Guide to Interviewing, Participant Observation, Data Analysis, and Writing It All Up (Univ. of Chicago Press, 2021).
34. R. S. Weiss, Learning From Strangers: The Art and Method of Qualitative Interview Studies (Free Press, 1994).
35. Stanford Center on Poverty and Inequality, "American Voices Project" (2021); https://inequality.stanford.edu/avp/methodology.
36. S. Ansolabehere, J. Rodden, J. M. Snyder Jr., The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting. American Political Science Review 102, 215-232 (2008).
37. M. J. Salganik et al., Measuring the predictability of life outcomes with a scientific mass collaboration. Proc. Natl. Acad. Sci. U.S.A. 117, 8398-8403 (2020).
38. L. P. Argyle et al., Out of one, many: Using language models to simulate human samples. Political Analysis 31, 337-355 (2023).
39. M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, in Advances in Neural Information Processing Systems 29 (2016), pp. 3315-3323; arXiv:1610.02413.
40. S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning (2019); https://fairmlbook.org
41. M. N. Stagnaro, J. Druckman, A. J. Berinsky, A. A. Arechar, R. Willer, D. Rand, Representativeness versus Response Quality: Assessing Nine Opt-In Online Survey Samples. OSF Preprints [Preprint] (2024); https://osf.io/preprints/psyarxiv/h9j2dc
42. T. W. Smith, M. Davern, J. Freese, S.L. Morgan, "General Social Surveys, 1972-2020: Cumulative Codebook" (NORC at the University of Chicago, 2021);

https://gss.norc.org/content/dam/gss/get-documentation/pdf/other/2020%20GSS%20Replicating%20Core.pdf
43. R. M. Groves, F. J. Fowler Jr., M. P. Couper, J. M. Lepkowski, E. Singer, R. Tourangeau, Survey Methodology (John Wiley & Sons, ed. 2, 2009).
44. S. Brinkmann, S. Kvale, InterViews: Learning the Craft of Qualitative Research Interviewing (SAGE Publications, ed. 3, 2014).
45. N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, P. Liang, Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics 11, 1312-1327 (2024).
46. B. Reeves, C. Nass, The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places (Cambridge University Press, 1996).
47. OpenAI, "Text to speech guide" (2024); https://platform.openai.com/docs/guides/text-to-speech (accessed 20 September 2024).
48. OpenAI, "Whisper" (2024); https://openai.com/research/whisper (accessed 20 September 2024).
49. OpenAI, "GPT-4" (2024); https://openai.com/research/gpt-4 (accessed 20 September 2024).
50. P. V. Marsden, T. W. Smith, M. Hout, Tracking US Social Change Over a Half-Century: The General Social Survey at Fifty. Annual Review of Sociology 46, 109-134 (2020).
51. NORC at the University of Chicago, "General Social Surveys, 1972-2022: Cumulative Codebook" (2023); https://gss.norc.org/content/dam/gss/get-documentation/pdf/codebook/GSS%202022%20Codebook.pdf (accessed 20 September 2024).
52. S. C. Schmitt, J. J. Gaughan, B. N. Doritya, A. L. Gonzalez, L. D. Smillie, R. E. Lucas, D. B. Nelson, M. Brent Donnellan, The Big Five Across Time, Space, and Method: A Systematic Review. PsyArXiv [Preprint] (2023); https://doi.org/10.31234/osf.io/37w8p
53. Y. Strus, J. Cieciuch, Toward a synthesis of personality, temperament, motivation, emotion and mental health models within the Circumplex of Personality Metatraits. Journal of Research in Personality 82, 103844 (2019).
54. C. F. Camerer, Behavioral Game Theory: Experiments in Strategic Interaction (Princeton University Press, 2003).
55. B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, T. Yarkoni, Promoting an open research culture. Science 348, 1422-1425 (2015).
56. J. Chandler, M. Cumpston, T. Li, M. J. Page, V. J. H. W. Welch, Cochrane Handbook for Systematic Reviews of Interventions (Wiley, ed. 2, 2019).
57. Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used? Journal of Applied Psychology, 72(1), 146-148.

Supplementary Materials for

Generative Agent Simulations of 1,000 People

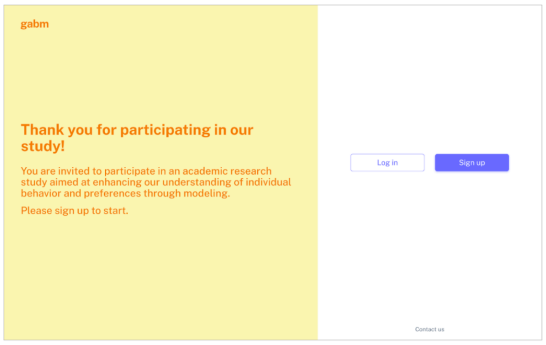
Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai,
Meredith Ringel Morris, Robb Willer, Percy Liang, Michael S. Bernstein

Corresponding author: joonspk@stanford.edu

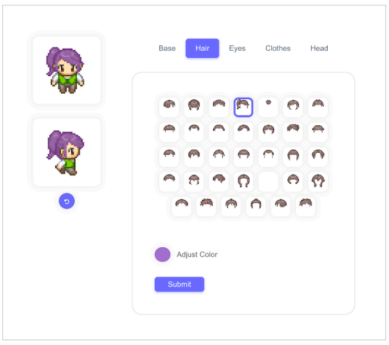
The PDF file includes:

1. Constructing the Agent Bank

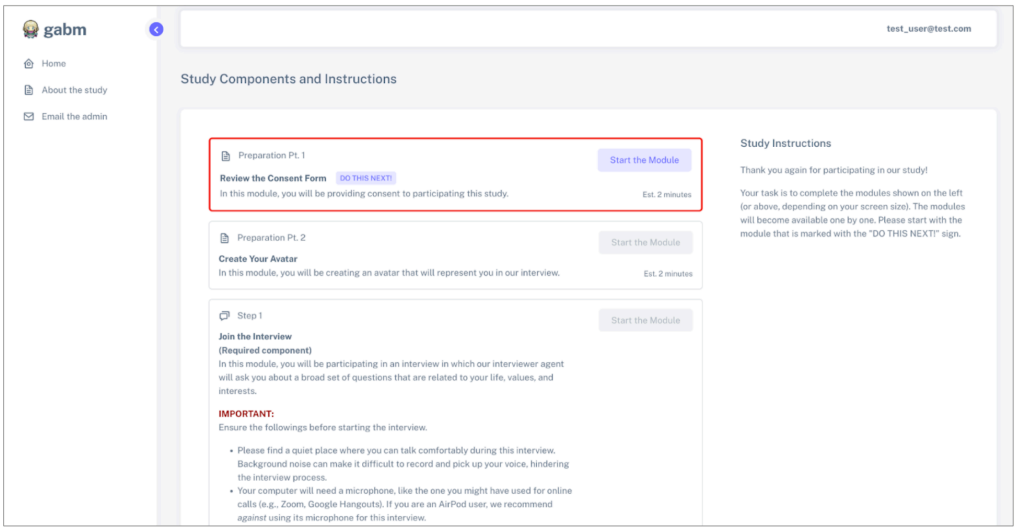
We created over 1,000 generative agents, each modeling a real individual in the U.S., collectively forming a representative sample of the U.S. population. To achieve this, we recruited a stratified sample of 1,052 individuals from the U.S. and conducted two hour voice-to-voice interviews using an AI interviewer (SM 2). In addition, we collected each participant's responses to a series of surveys and behavioral experiments. The interview transcripts formed the comprehensive knowledge base about the participants to condition agent behaviors (SM 3), and the participants' responses to the surveys and experiments were used to assess the fidelity of the resulting agents. This section details the participant data collection procedure, including participant recruitment and flow, demographic distributions, and informed consent.



a) Participant Sign Up Page



b) Avatar Creator



c) Main Interface Displaying the Study Components

Figure 1. The study platform and interface. Once recruited, our participants are routed to our custom-built platform. The interface includes several components: a) Participant sign-up page: Participants sign up with an ID and password of their choice. b) Avatar creator: Participants consent and create a 2-D sprite avatar to represent them in the study platform. c) Main interface displaying the study components: The modules include: 1) study consent, 2) avatar creation, 3) interview, 4) surveys and experiments, 5) self-consistency retake of the surveys and experiments. The modules only become available in order; the button to start a module becomes clickable once the participants have completed all previous modules. The self-consistency survey and experiment module only becomes available two weeks after the participants have completed the previous modules.

Data Collection Procedure

Data collection took place on our custom-built platform, where participants signed up with an ID and a password of their choosing (see Figure 1 for visual records of the study platform). This process was conducted in two phases. In the first phase, participants were informed about the study's goals, scholarly benefits, and potential risks, and provided informed consent. Once our participants provided consent to the study, they created a custom, 2-D sprite avatar using our avatar creator to visually represent themselves on the study platform. They then completed a two-hour interview with our AI interviewer, followed by a series of surveys and experiments. The surveys and experiments were administered in the following order: the General Social Survey (GSS; *20*), the 44-item Big Five Inventory (BFI-44; *21*), five behavioral economic games *(22-25)*, and five behavioral experiments *(26-31)*. Within the economic games and the replication studies, the order of the subcomponent studies was randomized for counterbalancing purposes, and similarly for the BFI-44, the order of questions was randomized. The GSS adhered to the sequence recommended by its documentation *(42)*. Details of all surveys and experiments appear in SM 4.

For the second phase, participants joined a follow-up study two weeks after their first phase participation. In this phase, participants completed the same surveys and experiments as in the first phase, except for the interview. This was done to account for any inconsistency in responses to surveys and experiments, allowing us to measure the internal consistency of participants' responses over a two-week period.

Recruitment and Demographics

We aimed to enroll a total of 1,000 participants who would complete all study components, including the second phase participation scheduled two weeks after the initial involvement. The sample size of 1,000 was determined to ensure that we could replicate the five behavioral experiments in our study with appropriate statistical power. Anticipating an attrition rate of approximately 20% for the re-taking session, we recruited 1,300 participants for the first phase of the study. The participation rate for the self-consistency phase was higher than expected; to maintain the representativeness of our sample, we retained 1,052 participants in the final pool.

Participants were recruited through Bovitz, a study recruitment firm *(41)*. Our stratification strategy aimed to recruit a nationally representative sample based on age, race, gender, region of residence, educational attainment, and political ideology. The only inclusion criterion was that all participants must be at least 18 years old and currently living in the U.S. Participants were paid $60 for agreeing to participate in the first phase, which included the interview and the first phase of surveys and experiments. They were paid an additional $30 for joining the second phase of the study. Additionally, for both phases, participants were eligible for a bonus payment ($0 to $10) depending on their choices in the economic games. The mean age of our participants was 47.55 (std = 15.93), with maximum age being 84 and minimum 18. With respect to gender, 593 participants identified as female, and 459 as male. With respect to education, 283 participants held a bachelor’s degree, 151 had a higher degree, 185 had an associate’s degree, and the rest had a high school diploma or some high school-level education. With respect to ethnicity, 833 of our participants identified as white/Caucasian, 154 as black/African American, 53 as Asian, and 95

as other. Note that for ethnicity, the participants could choose more than one option. The full demographic breakdown is described in Table 1.

Participant Consent

The data we collect in this study, particularly the qualitative interview data, is difficult to anonymize and poses a risk due to the potentially sensitive nature of the interview content. Therefore, in addition to following best practices and employing precautionary strategies for providing the agent bank access to the scientific community, we placed significant emphasis on the consent procedure. We worked with our Institutional Review Board (IRB) for over six months to ensure that participants maintain autonomy and provide informed consent. Participants were made aware that, despite efforts to de-identify their data by programmatically replacing all occurrences of their names with pseudonyms, there is still a possibility that the information they provide—such as demographic details, personal history, and political views—may be inadvertently shared as researchers use the Agent Bank. They were informed that their data would be used to develop AI models simulating human behaviors, which "aim to simulate how [they] might behave in specific situations or respond to certain survey questions," and that these agents and data might become available to other researchers strictly for academic purposes.

Additionally, participants were informed that they have the right to withdraw their consent at any time, even after completing their participation. Requests for data removal will be honored for the first 25 years following the completion of the study to the best of our ability. Participants will also be kept informed of significant changes in the model's capabilities that may affect their privacy, with the assurance that privacy risks will be reassessed as necessary. Despite our best efforts, participants are aware of the inherent risks involved in the collection of personal information, acknowledging that "achieving complete anonymity remains challenging." They are also aware that "the models we construct may become increasingly powerful over time," potentially inferring more information than currently feasible.

2. Creating the AI Interviewer Agent

To ensure the high quality and consistency of the rich training data needed for creating generative agents, we developed an AI interviewer agent to conduct semi-structured interviews with study participants. We sought interviews rather than surveys because we anticipated that interviews could yield more comprehensive and nuanced information, enabling the creation of generative agents capable of higher-fidelity attitudinal and behavioral simulations across a wide range of topics and domains. However, conducting large-scale interviews using human interviewers poses significant challenges, including threats to data quality, consistency, and scalability *(43)*. By employing an AI interviewer-agent built with a variant of the generative agent architecture *(15)*, we aimed to ensure uniformity in the style and quality of interviewer interactions across all participants. Additionally, this approach allowed us to scale up our data collection to over 1,000 participants.

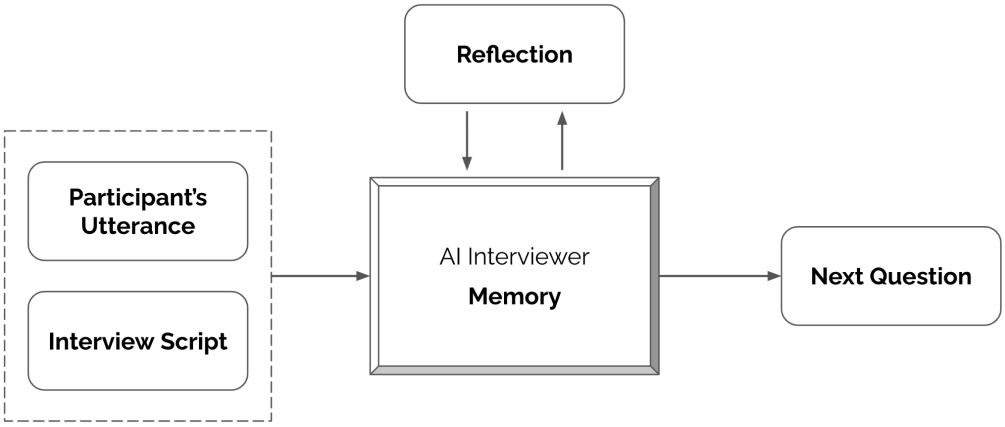


Figure 2. The architecture of the interviewer agent. It takes as input the participants' utterances and the interview script, generating the next action in the form of follow-up questions or deciding to move on to the next question module using a language model. A reflection module helps the architecture succinctly summarize and infer insights from the ongoing interview, enabling the agent to more effectively generate follow-up questions.

AI Interviewer Agent Architecture

A trained human interviewer knows when and how to ask meaningful follow-up questions, balancing the need to adhere to a well-designed interview script while allowing for detours that help participants open up and share aspects they may have initially forgotten or not thought to share *(33, 34, 44)*. To instill this capability in an AI interviewer agent, we needed to design an interviewer architecture that affords the researchers control over the overarching content and structure of the interview while allowing the interviewer agent a degree of freedom to explore follow-up questions that are hard-coded in the interview script. This served as our design goal for the AI interviewer agent.

Our interviewer architecture takes an interview protocol and the most recent utterances from the interviewee as inputs and outputs an action to either: 1) move on to the next question in the script, or 2) ask a follow-up question based on the conversation so far. The interview script is an ordered list of questions, with each question associated with a field indicating the amount of time to be spent on that particular question. At the start of a new question block in the interview script, the AI interviewer begins by asking the scripted question verbatim. As participants respond, the AI interviewer uses a language model to make dynamic decisions about the best next step within the time limit set for the question block. For instance, when asking a participant about their childhood, if the response includes a remark like, "I was born in New Hampshire… I really enjoyed nature there," but without specifics about what they loved about the place in their childhood, the interviewer would generate and ask a follow-up question such as, "Are there any particular trails or outdoor places you liked in New Hampshire, or had memorable experiences in as a child?" Conversely, when asking the participant to state their profession, if the participant responds, "I am a dentist," the interviewer would determine that the question was completely answered and move on to the next question.

The reasoning and generation of the follow-up questions were done by prompting a language model. However, to generate effective actions for the interviewer, the language model needed to remember and reason over the prior conversational turns to ask meaningful follow-up questions that are informed and relevant in the context of what the participants have already shared. While modern language models have become increasingly proficient at reasoning, they still struggle to consider every piece of information in the prompt if it is too long *(45)*. Thus, indiscriminately including everything from the interview up to that point risks gradually degrading the performance of the interviewer in generating effective follow-up questions or decisions to move on.

To overcome this, our interviewer architecture includes a reflection module that dynamically synthesizes the conversation so far and outputs a summary note describing inferences the interviewer can make about the participants. For instance, for the participant mentioned earlier, it would generate reflections such as:

```
{
  “place of birth”: “New Hampshire”
  “outdoorsy vs. indoorsy”: “outdoorsy with potentially a lot of
  time spent outdoors”
}
```

Then, when prompting the language model to generate the interviewer’s actions, instead of including the full interview transcript, we included the much more concise but descriptive reflection notes the interviewer had accumulated up to that point and the most recent 5,000 characters from the interview transcript (Figure 2).

Interview Script

With the design of the interview script fed to our interviewer agent, we aimed to satisfy two goals. The first goal, shared with qualitative research, is that a well-designed script with questions that inspire meaningful answers is crucial for the study's objective of creating generative agents that encapsulate a nuanced portrait of the individuals we are modeling. The second goal is more unique to our study: we wanted an interview script that was designed independently of our evaluation metrics, by researchers outside our team. This approach ensures that we do not unfairly tailor the content of the interview script to favor or align with predicting participants' responses to the specific surveys and experiments included in our study.

To conduct the interviews, we employed a slightly abbreviated version of the interview script developed and used by the American Voices Project *(35)*, which we include in Table 7. The American Voices Project initiative involves recruiting a representative sample of the U.S. population for in-depth, approximately three-hour interviews. During these interviews, participants are questioned about their life experiences, including their life stories and perspectives on various social, political, and value-related topics. For instance, the interview script starts with an open-ended and broad question such as, “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.” It then proceeds to more topical questions, such as “How have you been thinking about race in the U.S. recently?” We selected this interview script due to its broad coverage that explores the lived

experiences of the interviewees. However, the script is extensive and includes specific questions that delve into intricate details that we considered too specific for many potential use cases of our agents, such as individuals' financial spending in various categories. Therefore, to make the interview manageable in a two-hour session, we omitted parts of the script (e.g., rapid-fire questions that delve into specific details of the participants' spending habits, or COVID-era life pattern changes) during our interviews with the participants.

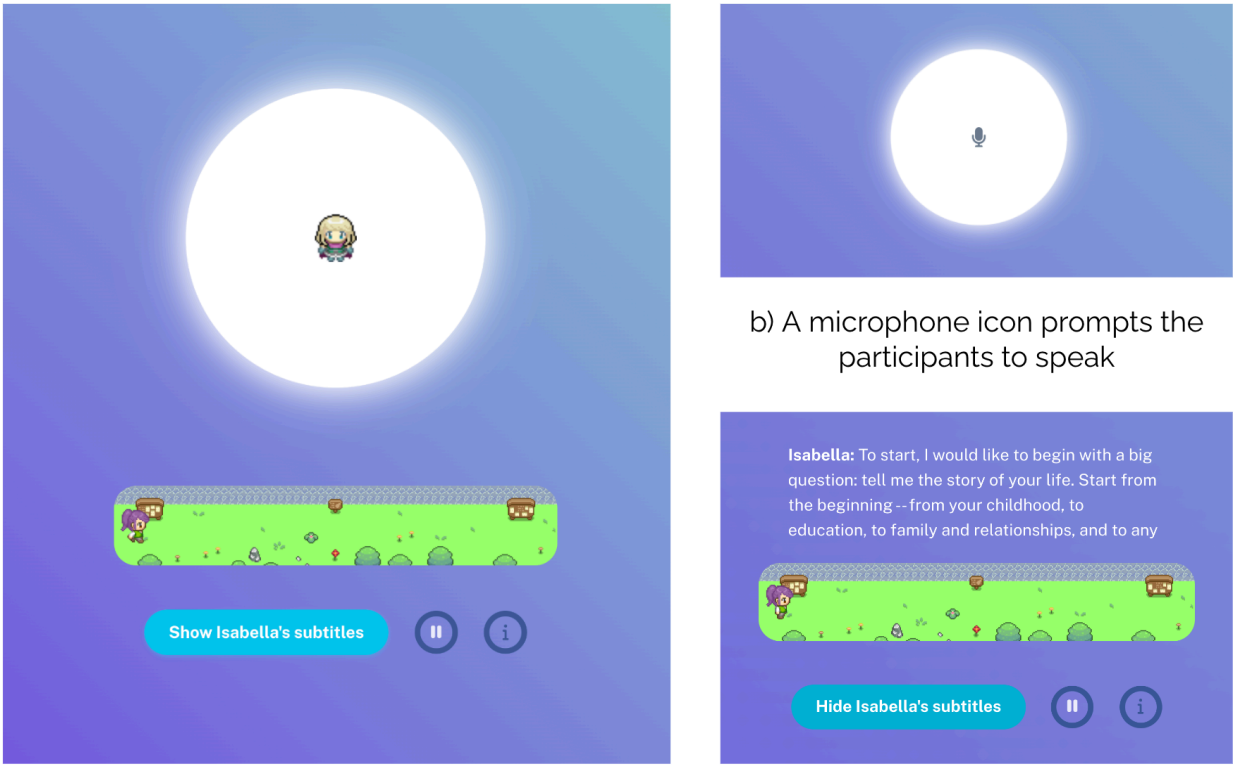


Figure 3. The interview interface. a) The main interview interface: A 2-D sprite representing the AI interviewer agent is displayed in a white circle that pulsates to match the level of the audio, visualizing the interviewer agent's speech during the AI interviewer's turn. b) Participant's response: The 2-D sprite of the AI interviewer agent changes into a microphone emoji when it is the participant's turn to respond, with the white circle pulsating to match the level of the participant's audio being captured. c) Progress bar and subtitles: A 2-D sprite map shows the participant's visual avatar traveling from one end point to the other in a straight line, indicating progress. The interface also features options to display subtitles or pause the interview.

Implementation

We implemented the interviewer agent as a web application in our study platform, providing voice-to-voice interactions with audio and microphone capabilities through an audio Zoom-like interface. Low-latency voice-to-voice interviews were crucial for giving participants the feeling of actually talking to an interviewer and helping the AI interviewer agent form rapport with the interviewee *(46)*. Before the interview, our platform disclosed that our interviewer is an AI, and conducted an audio calibration by asking participants to read aloud the first two lines of *The Great Gatsby* by F. Scott Fitzgerald.

The interview interface displayed the 2-D sprite avatar representing the interviewer agent at the center, with the participant's avatar shown at the bottom, walking towards a goal post to indicate progress (see Figure 3). When the AI interviewer agent was speaking, it was signaled by a pulsing animation of the center circle with the interviewer avatar. When it was the participant's turn to speak, the interviewer avatar changed to a microphone emoji and pulsed to match the audio being recorded, indicating the sound level registered from the participant. If the participant stopped speaking and silence lasted for longer than 4 seconds, the circle gradually faded, and the audio recording for the participant's utterance stopped. At this point, a loading animation appeared while generating the AI interviewer agent's next utterance. Our interviewer agent generally responded within 4 seconds—reasoning, generating, and returning its generated voice responses within this time frame—to maintain a smooth interview flow. The AI interviewer agent automatically started speaking when it was ready, with the interviewer avatar being displayed in the circle again.

The interview script is communicated to the AI interviewer agent as a JSON file containing an ordered list of questions. Each question is paired with a metadata field indicating a manually set time limit, suggesting the amount of time to be spent on each question so that the interview can conclude within 2 hours. Every question in the script, along with the follow-up questions, is read aloud using OpenAI's Audio model, a text-to-speech model that generates voice audio from textual input *(47)*. The participants' voice responses were transcribed using OpenAI's Whisper model, a speech-to-text model that converts voice audio into text *(48)*. This transcription allows us to use the interview transcript to prompt the language models to determine the next conversational move. Then, to dynamically generate reflections for the participants' responses to the current question, we prompted OpenAI's GPT-4o language model *(49)* with the following prompt (with input fields dynamically filled in):

```
Here is a conversation between an interviewer and an interviewee.
<INPUT: The transcript of the most recent part of the
conversation>

Task: Succinctly summarize the facts about the interviewee based
on the conversation above in a few bullet points -- again, think
short, concise bullet points.
```

And to dynamically generate new questions, we prompted GPT-4o with a prompt that looks as follows:

```
Meta info:
Language: English
Description of the interviewer (Isabella): friendly and curious
Notes on the interviewee: <INPUT: Reflection notes about the
participant>

Context:
```

```
This is a hypothetical interview between the interviewer and an
interviewee. In this conversation, the interviewer is trying to
ask the following question: "<INPUT: The question in the
interview script>"

Current conversation:
<INPUT: The transcript of the most recent part of the
conversation>
=*=*=
Task Description:
Interview objective: By the end of this conversation, the
interviewer has to learn the following: <INPUT: Repeat of the
question in the interview script, paraphrased as a learning
objective>
Safety note: In an extreme case where the interviewee
*explicitly* refuses to answer the question for privacy reasons,
do not force the interviewee to answer by pivoting to other
relevant topics.

Output the following:
1) Assess the interview progress by reasoning step by step --
what did the interviewee say so far, and in your view, what would
count as the interview objective being achieved? Write a short
(3~4 sentences) assessment on whether the interview objective is
being achieved. While staying on the current topic, what kind of
follow-up questions should the interviewer further ask the
interviewee to better achieve your interview objective?
2) Author the interviewer's next utterance. To not go too far
astray from the interview objective, author a follow-up question
that would better achieve the interview objective.
```

On average, with this implementation, our AI interviewer agent spoke 5372.59 (std=2406.12) words during the interview, asking on average 81.71 (std=54.39) follow-up questions from 99 scripted questions, to which our participants responded with on average 6491.19 words (std=2540.56).

Development Process and Evaluation

We iteratively refined and assessed the performance of the AI interviewer agent using a sample of 3 waves of 3 pilot interviews conducted between August 2023 to November 2023. Of the pilot interviews, 10 were conducted by human interviewers and 25 by the AI interviewer agent. The resulting interview transcripts were evaluated by members of our research team who were trained in the social sciences and assessed for their performance as training data for generative agents on the same set of attitudinal and behavioral tasks presented in the main results of this article. We compared the quality of the AI Interviewer interview transcripts to transcripts of interviews performed by expert human interviewers as part of the American Voices Project.

By the end of this pilot stage, our team concluded that the quality of the transcripts produced by the AI interviewer agent compared well with those produced by human interviewers. Additionally, in our pilot studies the performance of the generative agents of real people created using transcripts collected by the AI interviewer agent matched those created using transcripts collected by human interviewers when predicting GSS responses. For instance, our deployed AI interviewer agent could maintain a smooth conversation, ask relevant follow-up questions when participants' responses were terse, and navigate sympathetically after hearing about a difficult experience. This can be seen from an excerpt of an interview for the first question of the interview—only the first interviewer question is scripted, while the subsequent utterances from the interviewer are generated:

> **[Interviewee]:** I was raised in a rural town in [Redacted] where there are actually more cows than people. I'm [Redacted] years old, so I am a child of the 70s and 80s. I wound up being a latchkey kid and was largely raised either by myself or by daycare and babysitters. Childhood was not really great and I was really glad to move on to high school.
>
> **[Interviewer]:** I'm sorry to hear that childhood wasn't great for you. Could you tell me more about your experiences in high school? What was it like for you there, and were there any particular events or people that had a significant impact on your life during that time?
>
> **[Interviewee]:** So I had a head injury while I was in elementary school and the full weight of that untreated injury came to bear when I was in about 8th grade and I started having these really bad mood swings. So I spent most of my high school years in deep depression and no one got me any help for it. It was pretty awful. I excelled in English, in theater, did pretty well in science, won a lot of awards at the end of the school time when it was time to graduate. Didn't have the greatest GPA because I had a habit of skipping classes and not doing any homework but acing finals. Particular events or people that had a significant impact, I would say that the album Pink Floyd The Wall pretty much helped me with my really terrible mood. I felt understood by that music. And I had two high school sweethearts. One was a boy and one was a girl. So they made a big impact. One of them I'm no longer in touch with and today is the girl's birthday.
>
> **[Interviewer]:** Thank you for sharing that with me. It sounds like high school was a particularly challenging yet formative time for you. After high school, what path did you take with your education? Did you go to college, or did you jump straight into a career?
>
> **[Interviewee]:** After high school I went off to college originally to study criminal justice, but I changed over to English literature with a concentration in behavioral studies. After college, well, during college I had a bunch of different jobs. I worked with radioactive waste disposal, I worked at a woman's bookstore, I worked at a nightclub. I went off to law school and was there until I had to take a leave of absence because of my bipolar disorder, which was finally diagnosed. Years later, like 20 years later, I wound up going back to school again and got a business degree, mainly because I thought it would be fun to do. In the meantime, between law school and going back to business school, I worked

as a legislative aide. I was a legislation director for a state senator and I wrote bills and got them passed into law. Then I took over my parents' business as a manager when my dad got breast cancer. After I worked as a manager there and I moved out of state in order to get married, I wound up buying into that business, so now I'm a part owner of the motel business where I'd started working when I was 15 as a housekeeper, so it's kind of gone full circle. Now I'm going to be working as a part-time firearms instructor.

3. Generative Agent Architecture

Generative agents are software systems that simulate human behavior, powered by a language model augmented with a set of memories to define their behaviors *(14, 15)*. These memories, stored in a database (or "memory stream") in text form, are retrieved as needed to generate the agent's behaviors using a language model. This is paired with a reflection module that synthesizes these memories into reflections, selecting portions or all of the text in the agents' memories to prompt a language model to infer useful insights, thereby enhancing the believability of the agents' behaviors. While traditional agents in agent-based models rely on manually articulated behavior in specific scenarios, generative agents leverage language models to produce human-like responses that reflect the personas described in their memories across a wide range of circumstances. In this work, we aimed to build generative agents that accurately predict individuals' attitudes and behaviors by using detailed information from participants' interviews to seed the agents' memories, effectively tasking generative agents to role-play as the individuals that they represent.

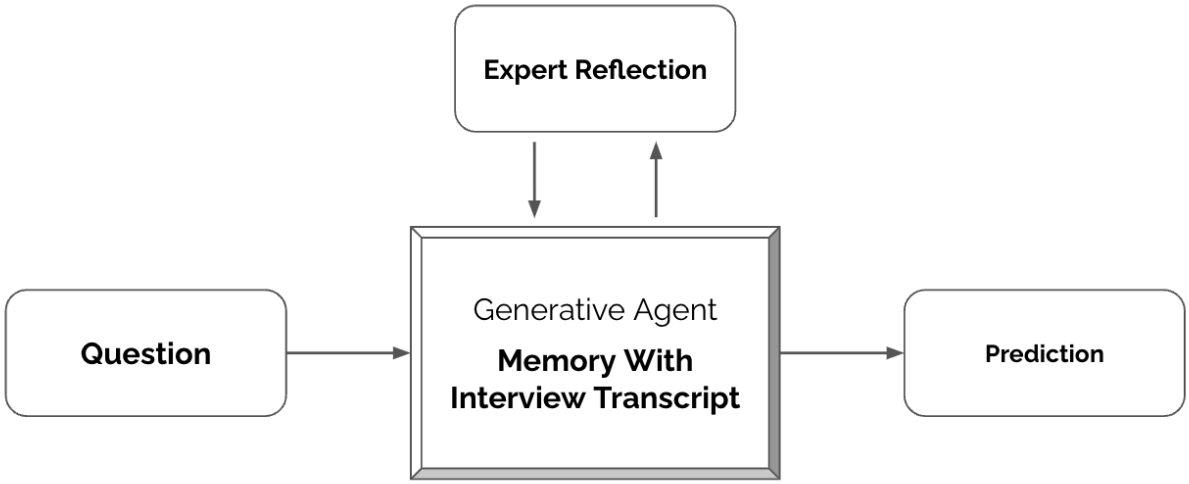


Figure 4. The architecture of our generative agents involves taking a question as input and outputting a prediction of how the source participant might respond, using a language model. Each agent's memory comprises the interview transcript and the outputs of expert reflections over that transcript. These reflections are short syntheses generated using a language model, designed to infer insights about the participants that might not be explicitly stated. The personas of expert social scientists (e.g., psychologist, behavioral economist) guide these reflections.

Expert Reflection

Prompting the language model with participants' interview transcript to predict their responses in a single chain of thought may cause the model to overlook latent information not

explicitly stated by the interviewee. To explicitly infer high-level, more abstract insights about the participants embedded in the interview transcripts, we introduced a variant of generative agents' reflection module called "*expert reflection*." In this module, we prompt the model to generate reflections on a participant's data, but instead of simply asking the model to infer insights from the interview, we ask it to adopt the persona of a domain expert. Specifically, we ask the model to generate four sets of reflections, each time taking on the persona of a different domain expert from four branches of social sciences: psychologist, behavioral economist, political scientist, and demographer. These sets of reflections synthesize insights relevant to the domain represented by each expert. For instance, for one interview transcript, the expert personas generated different insights:

> Psychologist: "[Redacted] values his independence and expresses a clear preference for autonomy, particularly highlighted by his enjoyment of traveling for his job and his frustration with his mother's overprotectiveness. This suggests a strong desire for personal freedom and self-determination."
>
> Behavioral Economist: "[Redacted]'s aspiration to save for a relaxing vacation and possibly advance to a managerial position indicates a blending of practical financial goals with personal leisure aspirations, emphasizing balanced life satisfaction."
>
> Political Scientist: "[Redacted] identifies as a Republican and espouses strong support for the party's views, particularly around immigration and drug policy. However, he also expresses specific support for traditionally Democratic positions on issues like abortion rights and the legalization of marijuana, suggesting a blend of ideologies."
>
> Demographer: "[Redacted] works as an inventory specialist and earns between $3,000 to $5,000 monthly, contributing to a household income of around $7,000 per month. He works primarily at Home Depots but has a varied work schedule, indicating some job stability and flexibility."

For every participant, we generated these four sets of reflections by prompting GPT-4o with the participants' interviews and asking it to generate up to 20 observations or reflections for each of the four experts. The prompt, tailored for each expert, was similar to the following (for the demographer expert):

```
Imagine you are an expert demographer (with a PhD) taking notes
while observing this interview. Write observations/reflections
about the interviewee's demographic traits and social status.
(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 these reflections once and saved them in the agents' memory. Whenever we needed to predict the participants' responses to a question, we first classified, by prompting the language model, which domain expert (demographer, psychologist, behavioral economist, or political scientist) would best answer the question. We then retrieved all reflections generated by

that particular expert. Along with the interview transcript, these sets of reflections informed the language model's generation of predictions for the participants' responses. After retrieval, we appended the reflections to the participants' interview transcript and used this to prompt GPT-4o to generate responses.

Generating a Prediction With Generative Agents

Our prompting strategy leveraged the chain-of-thought prompting approach:

```
<INPUT: Participant's interview transcript and relevant 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, and 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 option, reason about why the
Participant might answer with that particular option. ("Option
Choice")
Step 3) Write a few sentences reasoning on which of the options
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.
("Response")

Here are the questions:

<INPUT: Question we are trying to respond to>
```

To predict numerical responses, we modified the ending and prompted:

```
[... Same as the categorical response prompt]

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 end of the range. ("Range Interpretation")
Step 2) Write a few sentences reasoning on which option best
predicts the participant's response. ("Reasoning")
Step 3) Predict how the participant will actually respond.
Predict based on the interview and your thoughts. ("Response")
```

```
Here are the questions:

<INPUT: Question we are trying to respond to>
```

Finally, if the agents needed to maintain context from the experimental stimuli for the behavioral experiments, we appended the agents' received stimuli and prior actions in the experiments at the end of the interview transcript and reflections in natural language form.

4. Surveys and Experimental Constructs

To evaluate the fidelity of our generative agents, we aimed to assess their predictive accuracy regarding the attitudes and behaviors of the underlying sample across surveys and experimental constructs from a broad array of social scientific disciplines and methods. To operationalize this, we identified four existing constructs commonly deployed in the social sciences. In this section, we describe them.

The General Social Survey

The General Social Survey (GSS) is a long-running, widely used sociological survey administered biannually to representative cross sections of U.S. adults to collect information encompassing demographic details and respondents' viewpoints on issues such as government spending, race relations, and beliefs concerning the existence and nature of God *(20, 50)*. The survey consists of a repeated "core" module, run every cycle, that covers the more timeless elements of the survey, along with additional modules that are swapped in and out to meet the needs of the year when the survey is administered. Traditionally, the survey was conducted via voice, with the surveyor asking the questions in person or on the phone and later coding the participants' responses into discrete survey question and answer pairs. In recent years (since the COVID-19 pandemic, when in-person contact was challenging), the GSS has been implemented and administered online *(51)*. The ability to predict GSS responses signifies a comprehensive understanding of individuals, particularly in areas of interest to social scientists *(50)*. It also signifies the ability of our agents to predict the participants' responses to survey constructs on topics related to societal issues and personal views.

In our study, we focused on the GSS Core as it represents the most enduring and important set of survey questions within the GSS. While some questions within the GSS allow for qualitative responses or freeform input, our evaluation specifically focused on questions that ask for structured responses in the form of categorical or numerical answers. These questions can be quantitatively assessed, making them the primary focus of our evaluation efforts. Consequently, we excluded: 1) conditional questions that depend on answers to other questions, 2) questions with more than 25 option categories, and (3) questions requiring free-form responses. This refinement process resulted in a final analytic sample of 177 categorical questions and 6 numerical questions. Following recent best practice *(51)*, we administered these questions online through a custom-built Qualtrics survey linked from our study platform.

Big Five Personality Traits

The Big Five personality traits are a widely recognized framework in psychology for understanding human personality *(21)*. The construct encompasses five broad dimensions that capture substantial variability in individuals' personalities: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. Each trait represents a range, with individuals receiving a score for each trait. The Big Five traits have been validated across diverse cultures and are used to predict a wide range of behaviors and life outcomes, from academic and job performance to social relationships and mental health *(21)*. Traditionally, these traits are assessed using self-report questionnaires where respondents rate their agreement with each statement on a Likert scale. The ability to accurately predict an individual's Big Five personality traits is central for psychologists and researchers *(52, 54)*, as it provides insights into human behavior, informs interventions, enhances personal development, and contributes to the understanding of social dynamics.

In our study, we used the 44-item version of the questionnaire for testing the Big Five personality traits (BFI-44), developed by Oliver John and Sanjay Srivastava in 1998 *(21)*. We administered all 44 questions in the BFI-44 through a custom-built Qualtrics form linked from our study platform. Using the participants' responses to these questions, we calculated the scores following the aggregation methods suggested in the original work.

Behavioral Economic Games

Behavioral economic games are a set of experimental tools to study decision-making and social behavior under real stakes. Each game is designed to reveal aspects of human social behavior such as altruism, trust, cooperation, and competition. Participants engage in these games with real financial incentives. In our study, we offered participants a bonus payment based on their choices in the games. This incentive helps ensure that participants' choices reflect genuine preferences and strategies. These economic games are pivotal in understanding the underlying motivations and behavioral patterns in economic decision-making, providing valuable data for social scientists in multiple fields *(52)*.

In our study, we included the following five economic games, chosen for their prominence in the academic community:

- Dictator Game *(22)*: This game measures altruism by allowing one player (the dictator) to decide how to split a sum of money between themselves and another participant. The participants are initially given $5 and decide how to split the $5 between themselves and one other participant. The other participant cannot affect the outcome chosen.
- The Trust Game (First Mover) *(23)*: This game assesses trust, with the First Mover deciding how much of an initial endowment to send to the Second Mover. Player 1 is initially given $3 and then chooses how many dollars, if any, to send to Player 2, another participant in the study. Unlike the Dictator Game, the sum sent is tripled, so for every $1 sent, Player 2 receives $3.
- The Trust Game (Second Mover) *(23)*: This game assesses reciprocity, with the Second Mover deciding how much of the tripled amount received from the First Mover to return. This game continues from the actions of the First Mover in the Trust Game, as described above.

- The Public Goods Game *(24)*: This game explores cooperative behavior by having participants decide how much of their private endowment to contribute to a common pool. Participants are randomly assigned to interact with three other participants, with everyone receiving the same instructions. Each person in the group is given $4 for this interaction. They each decide how much of the $4 to keep for themselves and how much (if any) to contribute to the group's common task. All money contributed to the common task is doubled and then split among the four group members. The money is equally distributed among all players, regardless of individual contributions.
- The Prisoner's Dilemma *(25)*: This game examines the tension between cooperation and self-interest. Each of two participants chooses whether to cooperate or defect. If both cooperate, they receive a moderate payoff of $6. If one defects while the other cooperates, the defector receives a high payoff of $8, and the cooperator receives a low payoff of $2. If both defect, they receive a low payoff of $4.

We administered all five games through a custom-built Qualtrics form linked from our study platform and calculated the bonus amount afterward. At the start of the five games, we informed participants that we would randomly select one of the five studies to calculate their bonuses, but we did not tell them which game would be selected. Each game offered a maximum bonus of $8 to $10. After the study, we randomly paired participants and used the Dictator Game to calculate the bonus amount—the selection of the Dictator Game was randomly determined by us prior to the study. The bonus amount was based on participants' actual earnings in that game.

Replication Studies of Experimental Treatment Effects

Experiments involving randomized controlled trials (RCTs) of interventions ("treatments") with human subjects are standard across the social sciences. In a typical case, study participants are randomly assigned to either a treatment group or a control group, ensuring that outcome differences can be attributed to the intervention and minimizing confounding variables. Somewhat more formally, in this basic example, random assignment ensures that treatment assignment (Z) is independent of the observed outcomes ($Y(1)$ for treatment or $Y(0)$ for control) conditional on any set of observed or unobserved covariates (X) (i.e., $Z \perp \{Y(0), Y(1)\} | X$) *(47)*. Researchers often estimate treatment effects by calculating the difference between the average (mean) outcome in the treatment group and control group. Estimates of treatment effects from rigorous, well-replicated studies support evidence-based practice in social sciences like psychology and guide policy decisions and clinical recommendations *(55, 56)*. By predicting such effects with generative agents, we evaluate whether individual agent behaviors also exhibit (aggregate) responses to interventions in ways that accurately simulate human samples.

Our analytic sample of the replication studies. In our study, we selected a sample of human behavioral experiments from a recent large-scale replication effort of experimental studies that were published in the *Proceedings of the National Academy of Sciences* as curated by Camerer et al. *(26)*. Each study had at least one clear hypothesis and a significant reported effect. Sampling from the studies replicated by Camerer et al. ensured that 1) we did not subconsciously choose studies more favorable to the generative agents; 2) all estimates of treatment effects among human study participants had already been replicated in pre-registered studies conducted by

independent research teams and subjected to peer review; and 3) the interventions we tested were drawn from multiple social scientific disciplines.

The Camerer et al. project replicated 41 studies in total. Among these, we selected studies based on two criteria: first, the study had to be describable in natural language (optionally with images) for processing by a language model; second, the power analysis from the replication effort suggested that the effects would be observable with 1,000 or fewer participants. The criteria ensured that our sample of 1,000 human participants and the corresponding 1,000 generative agents could replicate the effect if present. These filters resulted in the following five studies in our sample of experiments:

- Ames & Fiske, 2015 *(27)*: This study examines how perceived intent affects the evaluation of harm. Participants read a vignette about a nursing home employee who switched patients' medications. One group was told the switch was intentional, while the other was told it was unintentional. After reading the vignette, participants were asked to complete their choice of five tasks, such as providing opinions about how the nurse should be blamed and punished or taking a short quiz about the cost of healthcare in the U.S. The study found that those who read the intentional scenario were more likely to choose tasks related to assigning blame and punishment compared to those who read the unintentional scenario.
- Cooney et al., 2016 *(28)*: This study explores how perceived fairness affects emotional responses. In a modified dictator game, participants believed they were receivers and predicted whether they would feel less upset about not receiving a bonus if the decision was made fairly (by a coin flip) rather than unfairly (by personal choice). The study found that participants expected fairness to influence their feelings, anticipating less upset when the decision was perceived as fair.
- Halevy & Halali, 2015 *(29)*: This study examines the perceived benefits and costs of intervening in conflicts. Participants recalled their personal experience of either intervening or not intervening in a conflict between friends and were asked to assess how beneficial it was to intervene in the conflict. The results showed that those who recalled intervening perceived the intervention as more beneficial and less costly than those who did not intervene.
- Rai et al., 2017 *(30)*: This study explores how dehumanization affects participants' willingness to harm others. In a vignette-based experiment, participants were given a description of a person in a dehumanized manner, simply as a “man,” or a description of a person in a humanized manner with details about the person such as “John is a 29-year-old man with brown hair and brown eyes. People who know him would describe him as ambitious and imaginative...” Participants were then asked whether they would be willing to harm the person for monetary compensation. The study found that participants were more willing to harm a stranger described in dehumanized terms compared to one described in humanized terms when motivated by financial gain.
- Schilke et al., 2015 *(31)*: This study investigates how power influences trust in social exchanges. Participants, imagining themselves as typists, were divided into high-power and low-power groups based on financial need and job availability. In the high-power condition, participants' service was essential for their clients, and they were offering the service to make extra spending money. In the low-power condition, participants' service was non-essential for their clients, and they were offering the service to make ends meet.

Trust was measured by the willingness to provide a free sample of their service to a potential client. The results showed that participants in the high-power condition were less willing to offer a free sample.

Our platform randomly assigned the participants to a condition for each of the five studies and we administered all five experiments through a custom-built Qualtrics form linked from our study platform.

5. Evaluation Methods

Given our survey and experimental constructs, we set out to evaluate the predictive power of the generative agents. In this section, we describe the metrics and evaluation methods used for this purpose. The individual subsection headers in this section are organized to match the presentation in the main document.

Study 1. Predicting Individuals' Attitudes and Behaviors

To determine whether generative agents of the 1,000 human participants accurately predict their respective individuals' behaviors and attitudes, we utilized the GSS, BFI-44, and five economic games. We deployed our generative agents to predict their respective individuals' responses to questions in these surveys and behavioral constructs. However, this measurement poses challenges due to variability in human participants' responses (*36, 37*). To address this, our evaluation employs the following strategy:

1) We use participants' responses from the first phase of participation to assess the accuracy rate of our agents' predictions—the number of answers predicted correctly over the total number of questions.
2) We use the second phase of participation to assess individuals' rate of internal consistency—the participants' rate of prediction on the battery of surveys and experiments used in this study.
3) We then calculate the *normalized accuracy* as follows:

$$normalized\ accuracy = \frac{agent's\ prediction\ accuracy}{internal\ consistency}$$

Conceptually, a normalized accuracy of 1.0 means that the generative agent predicts the individual's responses as accurately as the person replicates their own responses two weeks later.

The diverse response types in our surveys and experimental constructs present a challenge in determining a single metric for assessing our agents' predictive accuracy. For instance, while the categorical-ordinal responses in the GSS are well-suited to accuracy rates, numerical responses in other constructs are better evaluated using Mean Absolute Error (MAE) or correlation coefficients. To address this, we developed a reporting approach that satisfies the following criteria:

1) Report metrics appropriate for each response type (e.g., accuracy rate for categorical, MAE for numerical).

2) Ensure metric interpretability across different community norms (e.g., accuracy/MAE in machine learning literature, correlation in social science literature).
3) Provide a metric allowing comparison across different constructs.

To meet these criteria, we report accuracy rates for evaluation constructs with categorical-ordinal response types, MAE for numerical response types, and Pearson correlation coefficient as a metric comparable across constructs. Additionally, we present these metrics alongside the normalized accuracy to provide a comprehensive evaluation of our agents' performance.

Our analysis primarily focuses on *individual-level analyses*, conducted for each participant and then aggregated across the population. We calculate accuracy measurements for each individual and average these values across all participants. Additionally, we report and discuss *construct-level analyses*, which involve calculating accuracy measurements for the population and measuring the average predictive accuracy for individual questions, dimensions, or games in the constructs, averaged across all agents in the agent bank. In essence, the individual-level analysis answers "For the average person, how accurate are the generative agents?", whereas the construct-level analysis answers "For a specific item in one construct, how accurate are the generative agents?" In the main body of the article, we report the individual-level analysis, as our goal is to generalize over the individuals in our population rather than over items.

Averaging correlation coefficients presents a challenge due to their non-linear nature. To address this, we employ Fisher's z-transformation, which converts correlation coefficients to a scale where they can be averaged linearly (*57*). The process involves:

1) Applying Fisher's z-transformation to each correlation coefficient: $z = \frac{1}{2} ln(\frac{1+r}{1-r})$
2) Calculating the average of the z-values.
3) Applying the inverse Fisher's z-transformation to the average z-value: $r = tanh(z)$

Below, we describe in more detail the evaluation methods and our reporting strategies for the individual constructs.

The General Social Survey. The subset of the core module of the GSS that fits our inclusion criteria, as described in the prior section, includes 183 questions, most of which—177—are categorical or ordinal (*categorical-ordinal)* response types, with the remaining 6 being numerical. As such, we report the accuracy rate and correlation coefficient for the categorical response type variables in the GSS, and separately report the MAE and correlation coefficient for the 6 numerical response type variables.

To calculate the accuracy of the categorical-ordinal GSS questions, we consider our prediction accurate and the participant consistent if the responses match exactly. The accuracy rate is the ratio of correctly predicted items divided by the total number of items. For correlations, we translated categorical and ordinal questions into numerical forms. For categorical variables, each response option is first transformed into a separate binary variable. For example, if a categorical variable has five response options, we create five binary variables, each indicating whether the participant selected that option. These binary variables are coded as 0 or 1. However, to prevent these binary variables from disproportionately influencing the overall correlation due to their higher count, we adjust their weights. Each binary variable is

assigned a weight such that the total weight for all variables combined equals that of the original categorical variable, ensuring fair representation. In this example, each binary variable would receive a weight of 0.25. For ordinal questions, we normalize the responses to a range between 0 and 1 by evenly spacing the options. This ensures that the ordinal responses maintain their inherent order and distance between options while fitting into a numerical scale. We then apply the same weighting principles when calculating the correlation between the participants' original responses and their later self-consistency responses to ensure conformity across variable types.

To calculate the MAE for the numerical GSS questions, we first normalize the participants' responses to a 0 to 1 scale relative to the range of historical responses to the respective question as indicated on the official NORC site for the GSS. For instance, for the question "Your age," if the historical minimum value was 18 and the maximum was 89, an age of 30 would be normalized to 0.17 accordingly. We use these normalized values to calculate the MAE across the questions and responses. Similarly, we calculate the MAE between the participants' responses and predictions, and between the participants' responses and their self-consistency responses.

Using these measurements for our predictions and the participants' internal consistency, we calculate the normalized accuracy. However, we note that normalized accuracy cannot be computed for MAE, as some rows contain internal consistency values of 0 (when there is no variation, meaning the participant gave the same response in both the test and retest), making the denominator 0. Therefore, we report the normalized accuracy only for accuracy and the correlation coefficient.

BFI-44. Each dimension in the five-dimension measurement of the BFI-44 is provided as a scale ranging from 1 to 5. These scales are calculated based on the participants' responses to the 44 questions in BFI-44. To calculate these scales, we first reverse-code certain items as specified in the original work. After reverse coding, we aggregate the responses by taking the average for each dimension. Then, we use these averaged scales to calculate and report the MAE and correlation coefficient of the five scales. As with the GSS, we only report the normalized accuracy for the correlation coefficient.

Economic games. The five economic games include numerical (dictator game, trust games, public goods) and dichotomous (prisoner's dilemma) response types. To normalize these into one construct, we first normalize the responses for each game to a 0 to 1 scale using the minimum and maximum values offered to the participants as min and max. Then, we use the normalized values to calculate the MAE and correlation coefficient. Again, as with the GSS, we only report the normalized accuracy for the correlation coefficient.

Study 2. Replicating Experiments With Generative Agents

Do generative agents predict the behavior of sample average treatment effects? To study this, we use the responses from the human participants in our study to directly run a replication of the five sampled experiments, rather than using the published results. This direct replication is important because our sample might be different, and some studies may not replicate. To the extent that some of these studies are not replicated by the agents, this approach allows us to

distinguish whether the failed replication was due to the original study not replicating, or whether the failure to replicate instead lay with the agents.

In this evaluation, we are interested in three related outcomes: whether the effect direction and significance replicate, and whether there is a correlation between effect sizes. The five studies have different statistics, so we use the same statistical methods as in the original papers to derive the p-values for the effects' significance. We briefly describe these methods here. Consistent with the original replication studies conducted by Camerer et al., we calculate and compare all effect sizes in terms of Cohen's *d*.

- Ames & Fiske, 2015 *(27)*: A chi-square test (χ2-test) of equal proportions to evaluate the hypothesis by comparing the number of participants choosing the blame task between the two experimental conditions.
- Cooney et al., 2016 *(28)*: An F-test on the Outcome x Procedure interaction within a 2x2 ANOVA design, which evaluates how the predicted feelings about fairness (receiving or not receiving a bonus) interact with the allocation procedure (fair or unfair).
- Halevy & Halali, 2015 *(29)*: An independent samples t-test to compare the perceived benefits of intervening in a conflict between two friends across two treatment groups (those who did intervene and those who did not).
- Rai et al., 2017 *(30)*: An independent samples t-test to compare participants' willingness to harm a stranger described in humanized versus dehumanized terms under instrumental motives.
- Schilke et al., 2015 *(31)*: A chi-square test (χ2-test) to compare the levels of trust between participants with high structural power and those with low structural power.

We report which of the studies our human participants and simulated predictions replicated with significant results. To understand how well the effect sizes are correlated, we calculate the Pearson correlation coefficient between the effect sizes from the human participants and simulated participants.

Study 3. Interviews Improve Agents' Prediction Accuracy

Interviews provide qualitative data expressed in participants' own words, but to what extent do interviews improve agents' predictive capabilities? To address this question, we present two sets of analyses:

1) Our main analysis, conducted with all agents in our agent bank and detailed in the main article, compares the predictive performance of interview-based generative agents against agents created using the known practices from recent literature that studied human behavioral simulations with language models.
2) Exploratory analyses conducted on a random subset of 100 agents in the agent bank, investigating a broader range of design spaces. This includes examining generative agents with interview lesions and agents informed by survey data instead of interview data.

The main analysis aims to establish a baseline for predictive performance grounded in prior literature and evaluate whether our agent architecture surpasses this benchmark. In contrast, the exploratory analysis looks to determine whether interviews offer uniquely rich qualitative

insights that outperform other conceivable data types, and if they are shown to be more powerful, elucidate the underlying reasons for their effectiveness.

Main analysis. In this study, we evaluate two alternative agent descriptors: participants' demographic information and persona descriptions. To operationalize demographic information, we reconstruct agent descriptions following a method representative of approaches used in recent literature that employ language models to simulate human behavior. This method, similar to that presented by Argyle et al., constructs a first-person descriptor including political ideology, race, gender, and age (*28*). For example:

> Ideologically, I describe myself as conservative. Politically, I am a strong Republican. Racially, I am white. I am male. In terms of age, I am 50 years old.

We reconstruct these descriptors for our participants using their responses to the GSS. Then, to prompt the language model with this data, we replace the interview content with the demographic descriptors.

To operationalize persona descriptions, we asked participants to write a short paragraph about themselves at the end of their phase 1 participation in surveys and experiments. They were instructed to describe themselves as they might to a stranger, including information about their personal background, personality, and demographics. For instance:

> I am a 20 year old from new york city. I come from a working middle class family, I like sports and nature a lot. I also like to travel and experience new restaurants and cafes. I love to spend time with my family and I value them a lot, I also have a lot of friends which are very kind and caring. I live in the city but I love to spend time outdoors and in nature. I live with both parents and my younger sister.

Given this, we leverage both demographic agents and persona agents to predict participants' responses at the individual level. To compare performance differences between these agent versions, we conducted an ANOVA with post hoc Tukey tests, examining differences between architectures on the main evaluation metrics. This comparison was made between agents created with interviews, demographic information, and persona descriptions.

Exploratory robustness analysis. The primary goal for this exploratory analysis is to understand, in a more exhaustive manner, why the interviews are performant. We conducted this analysis on a random sample of 100 agents from our agent bank, comparing our interview-based generative agents with agents informed by different data sources. Each comparison aims to shed light on a particular question about the efficacy of interview-based data in generating predictive agents.

- Survey and Experiment Agents. To investigate whether interviews are a uniquely powerful medium or if similar information can be captured through surveys and

experiments, we created composite descriptions by compiling participants' responses to all three components used in our study—the GSS, Big Five personality test, and economic games. This approach aligns with traditional social science methods, which have developed and validated surveys and experiments to efficiently gain maximal information from participants. We aimed to determine if these established methods capture comparable information to our interview-based agents. In other words, do the interviews capture the same predictive power, more, or less, than these common validated instruments?

Agents were constructed using these composite descriptions to predict participants' responses to benchmark surveys and experiments. To ensure this was not simply a retrieval task, we excluded question-answer pairs from the same category as the question being predicted. For the GSS, categories were defined by the GSS core documentation, which subdivides the core survey into several subcomponents. For the Big Five, each personality category was considered a separate category, and for economic games, each individual game was its own category. On average, this exclusion process removed 4.00% (std=2.16) of the question-answer pairs per question.

- Maximal Agents. To assess whether the information captured in interviews fully encompasses that from surveys and experiments, we developed maximal agents. These agents were constructed using composite descriptions that integrated all available data sources—not just survey responses and experimental data as with the Survey and Experiment Agents, but also the complete transcript of each participant's interview. By equipping these agents with access to all data sources, we aim to determine our current performance ceiling: do interviews alone capture the most relevant information, or is there added value in combining methods? Similar to how we constructed the survey and experiment agents, we excluded data from the same category as the question being predicted.

- Summary Agents. We investigated whether the predictive power of interviews stems from linguistic cues in the transcript, or from the information in the transcript . To explore this, we created summary agents by prompting GPT-4o to convert interview transcripts into bullet-pointed dictionaries of key response pairs, capturing the factual content while removing most linguistic features (e.g., `{ "childhood_town": "Small town", "siblings": "Only child", "marital_history": "Married twice", "children": "Two children, but they are not living with the interviewee"...}`). By isolating the factual knowledge from the unique linguistic elements, we aimed to determine whether the predictive accuracy of the agents relies on those linguistic nuances. If the summary agents perform worse than the full interview-based agents, it would suggest that linguistic features play a key role in enhancing prediction accuracy.

- Random Lesion Interview Agents. We investigated the efficiency of interviews in providing insights compared to shorter data collection methods. While surveys such as the GSS Core typically take about 30 minutes, our interviews span two hours. To assess how much interview content is necessary to convey meaningful insights, we developed random lesion agents. These agents were created by progressively shortening interviews, randomly removing 0%, 20%, 40%, 60%, and 80% of the question-response pairs. This

approach allowed us to identify the threshold at which the interview becomes too sparse to retain valuable information.

Study 4. Demographic Bias in Agent Predictions

How do individuals' demographic attributes interact with the generative agents' ability to predict their respective behaviors and attitudes? Here, we investigated how individuals' demographic attributes interact with the generative agents' ability to predict their behaviors and attitudes. Specifically, we aimed to assess whether creating individualized models based on interviews reduces performance gaps across demographic groups, compared to relying solely on demographic attributes or personas. To quantify these differences, we employed the Demographic Parity Difference (DPD), a metric used in machine learning fairness literature *(39, 40)*. DPD measures the difference in prediction rates between demographic groups, with lower values indicating more equitable predictions.

We calculated DPD for three constructs—General Social Survey (GSS), Big Five personality traits, and economic games—across three agent conditions: interview-based, demographic-based, and persona-based. The interview-based condition utilized our novel approach, while the demographic and persona conditions served as baselines from prior literature. Using demographic attributes provided by our recruitment vendor, Bovitz, we calculated accuracy metrics for subpopulations across Studies 1 and 2. We compared these metrics across the three agent conditions to determine if certain populations experienced worse predictive accuracy, to quantify the extent of disparities, and to examine how these differences interacted with the agent condition. DPD values were reported alongside individual accuracy results for different population subgroups.

In addition, to assess the statistical significance of the observed differences, we conducted regression analyses. These analyses explored the impact of demographic variables on the predictive performance of generative agents, with separate regression models run for each demographic variable, using predictive performance as the dependent variable.

6. Supplementary Results

In this section, we present a higher-level interpretation supplementary results that, while not central to our main findings, offer valuable insights. The methods are as outlined in the previous section, and detailed tables of these results can be found in Section 8.

Numerical GSS Prediction Results

In addition to our predictive performance on the primary categorical questions of the General Social Survey (GSS), we also evaluated our model's ability to predict responses to the six GSS core numerical questions. The results show similarly strong predictive accuracy, with an average correlation of $r = 0.97$ (std = 0.82) and an average mean absolute error (MAE) of 0.14 (std = 0.15). These metrics indicate a high degree of alignment between our predictions and the actual GSS responses for the numerical questions.

Exploratory Robustness Analysis Results

We present supplementary findings from the exploratory robustness analysis (Study 3), organized by the materials used to inform agent behavior. Detailed results of this analysis are provided in Table 6.

- *Survey and Experiment Agents.* When excluding question-answer pairs from the same category as the predicted question, these agents underperformed compared to interview-based agents across all constructs. Specifically, a subsample of 100 interview agents achieved an average normalized accuracy of 0.85 (std = 0.11) on the GSS, whereas the survey and experiment agents reached only 0.76 (std = 0.12). This suggests that interviews captured richer, more comprehensive information than could be extracted from surveys and experiments alone.
- *Maximal Agents.* Maximal agents, which incorporated information from surveys, experiments, and interviews, achieved a similar performance to interview-based agents, with a normalized accuracy of 0.85 (std = 0.12) on the GSS. This finding suggests that the GSS and other constructs do not appear to be adding additional predictive power above and beyond the interviews.
- *Summary Agents.* The summary agents performed slightly below the interview agents, with a normalized accuracy of 0.83 (std = 0.12) on the GSS. This indicates that while some information may be lost from the linguistic cues during summarization, much of the performance is due to the information in the interview rather than the low level language that participants use.
- *Random Lesion Interview Agents.* Performance declined linearly as we removed increasing portions of the interview data. Starting with a normalized accuracy of 0.85 (std = 0.11) when no information was removed, accuracy dropped to 0.79 (std = 0.11) when 80% of the utterances were excluded. This suggests that although performance decreases as interview length is reduced, even a short interview contains sufficient richness to outperform agents informed solely by surveys and experiments, highlighting the efficiency of interviews in identifying valuable insights.

Demographic Bias in Agent Predictions

We present supplementary findings on bias in agent predictions. The complete results of our regression analysis can be found in Table 4, with demographic parity differences (DPD) detailed in Table 5. In addition to the DPD findings outlined in the main document, our regression analysis indicates that significant demographic performance discrepancies across constructs were relatively rare. However, notable exceptions emerged for political ideology, party affiliation, and sexual orientation when predicting GSS responses. Specifically, predictive performance was noticeably stronger for participants identifying as strong liberals, strong Democrats, and non-heterosexual/straight, compared to those identifying as more conservative, Republican, or heterosexual/straight.

7. Research Access for the Agent Bank

In this section, we outline a framework that defines the key elements of our research access and present a plan for providing scientific access to the agent bank. Access to the agent bank offers value to the scientific community, with important implications for two key domains:

- In social science, agents from the agent bank can be used to develop simulations involving individual or multiple agents. For example, how might a new government policy impact economic behavior? Would a social media intervention reduce political polarization? What factors influence whether institutions foster or erode prosocial behavior as a group grows? These models may allow social scientists to explore a wide range of American individual perspectives, and create bottom-up simulations that analyze the emergent behaviors of different social groups.
- In machine learning, the agent bank can serve as both a benchmark and a training resource for developing new models, prompts, and agent architectures that mimic the original participants. Much like how ImageNet contributed to the development of computer vision techniques (*56*), the agent bank may enable researchers to refine model prompts for improved predictive accuracy and assess how newly developed models can enhance these capabilities.

However, as the use of agent banks extends beyond our specific context, it is essential to strike a balance between the benefits they offer and the risks they may pose. This is especially important given the inherent uncertainty of future advancements in generative AI—such as enhanced reasoning abilities—which could introduce unforeseen vulnerabilities. For instance, unrestricted access to the agent bank might lead to privacy risks, including data leaks or the misuse of participants' identities. In a worst-case scenario, someone may manipulate agent responses to falsely attribute harmful or defamatory statements to individuals represented in the agent bank, creating significant reputational damage.

Overview of the Agent Bank

At a high level, Stanford University plans to provide controlled research-only API access to agent behaviors, allowing researchers to submit queries—such as questions they wish to ask our agents or prompts they want to run—and specify a target population to our agent bank server. The server will then return corresponding agent responses. The details of this plan are outlined in the subsection on Plan and Desiderata below. Our documentation will include general demographic information about the agents in the dataset so that researchers can know which populations are available.

The full agent bank consists of both data and code: the data includes detailed interview transcripts, along with survey and experimental responses from participants, while the code consists of Python scripts and language model prompts that generate agent behaviors from this data. However, granting unrestricted access to the raw data poses privacy and safety risks for participants, even with their consent to share the data for research purposes. Therefore, instead of providing the actual interviews, we will release a sample of the interview data to illustrate its structure and format. Access to the full agent behaviors will be offered via an API. By offering API access, we can maintain tighter control over what is shared and make adjustments based on

community feedback and ongoing observations. This approach balances flexibility with the necessary safeguards to protect participants' privacy.

Strategic Framework for Access

To accommodate the diverse use cases outlined above, we need to consider varying levels of access to the agent bank. What are the key dimensions along which we can structure this access? And what opportunities and risks emerge as we adjust these levels? We propose framing this discussion along three axes: 1) What types of tasks can be submitted? 2) How are agent responses presented? and 3) Who can access the system? This forms the design space, and we will present our proposed plan in the next section.

What types of tasks can be submitted? This axis examines the range of queries users can submit to our agents. At one end of the spectrum, users could submit any query, receiving responses in various forms, from discrete (e.g., multiple-choice) to open-ended (e.g., qualitative interview responses). This flexibility carries the risk that participants could potentially be identified through probing questions, even if raw data is not directly available. Moving along the axis, we could implement more structured tasks, where queries are predefined (e.g., as surveys or experiments from our benchmark) and responses are constrained to specific formats. In this scenario, users could still submit suggestions for new queries and response types, subject to review and approval. This more controlled access would allow users to replicate our findings while providing them an opportunity to explore their own interests, albeit with a slightly slower feedback loop due to the approval process.

How are agent responses presented? This axis addresses the form of the responses users will receive. At one end, users could access individual agent responses, allowing them to see exactly how each agent responded. While this offers detailed insight, it also poses privacy risks, as individual responses could reveal sensitive information. Moving along the axis, we could provide increasingly aggregated results—from lightly processed individual responses to summary statistics that represent the collective responses of the agents. While aggregation enhances safety, it may limit the utility of the Agent Bank for building bottom-up simulations, which rely on individual agent data. Nevertheless, aggregated results could still support machine learning benchmarks (e.g., by reporting accuracy statistics based on new models) and provide social scientists with valuable insights at a population level.

Who can access the system? The final axis concerns the range of users who can access the agent bank API. At one end, we could offer broad access to all researchers, conditioned by the terms outlined in our IRB. Researchers would still need to sign agreements restricting commercial use and adhere to ethical guidelines. This approach would maximize reach and scientific impact, allowing the academic community to quickly leverage the agents for testing theories and developing studies. However, it also introduces higher risks, as not all researchers would be subject to institutional oversight (e.g., IRBs), meaning that misconduct could pose an immediate threat to participant safety. Moving along the axis, we could implement increasingly stringent access controls, potentially requiring researchers to obtain IRB approval from their institution before accessing the dataset. While tighter controls provide a higher level of

protection, they also increase the burden on users and may slow adoption, particularly in fields like machine learning, where IRB approval is not typically required.

Plan and Desiderata

We aim to provide access to the agent bank to bridge the potential for novel research and accessible benefits for the scientific community. However, fully open access to individual agent responses on open tasks, while offering the most flexibility, introduces significant privacy risks for participants. To address this, we will implement a two-pronged access system:

- Open access to aggregated responses on fixed tasks: This category of access is designed to provide fast and flexible access while maintaining participant privacy. Academic researchers who sign our usage agreement will be granted access to aggregated data from fixed tasks, such as the surveys and experiments conducted in our study. Researchers can query subpopulations of the Agent Bank and receive responses in the form of aggregate data. While this access is somewhat limited, it is intended to facilitate meaningful investigations into agent behaviors and attitudes across broad population segments. In addition, researchers can submit suggestions for new queries or discrete response options, which will be reviewed on a rolling basis and potentially incorporated into the system.
- Restricted access to individualized responses on open tasks: For researchers aiming to fully explore the capabilities of the Agent Bank, we offer a more flexible, but controlled, access strategy. Researchers will need to provide a clear statement of their project's research purpose and its potential harms to the participants in the Agent Bank. Once approved, they may request access from us, and upon approval, will be granted API access to query the agents with custom questions. In this category of access, users will receive individualized agent responses, allowing them to maximize the potential of the Agent Bank for more in-depth research.

We recognize that as research evolves, so will the complexity of potential use cases, risks, opportunities, and the possibility of jailbreaking. Therefore, we will maintain an audit log of how these APIs are used and continue to monitor their usage. This may lead us to expand access—for instance, allowing users to submit new prompts for generating agent behaviors or integrating new models—or, alternatively, to restrict access further if necessary. The plan outlined here serves as a starting point, and we will adapt it based on community feedback and ongoing observations.

8. Supplementary Tables

Age	18 to 24	25 to 34	35 to 44	45 to 54	55 to 64	65 to 74
	11.03%	13.88%	17.49%	19.77%	21.48%	13.50%
	75 or more					
	2.85%					
Census division	New England	Middle Atlantic	E.N. Central	W.N. Central	South Atlantic	E.S. Central
	6.65%	12.83%	18.73%	8.08%	10.08%	11.5%
	W.S. Central	Mountain	Pacific	Foreign		
	8.65%	5.13%	15.78%	2.57%		
Education	Less than high school graduate	High school graduate	Associate/junior college	Bachelor's degree	Graduate degree	
	2.28%	38.88%	17.59%	26.9%	14.35%	
Race	White	Black	Other			
	75.95%	14.26%	9.79%			
Ethnicity	White/Caucasian	Black/African American	Asian	Native Hawaiian or Pacific Isl.	American Indian or Alaskan nat.	Other ethnicity
	79.18%	14.64%	5.04%	0.48%	2.76%	5.80%
Sexuality	Heterosexual /straight	Gay or lesbian	Bisexual	Asexual	Pansexual	Other sexual orientation
	82.2%	4.36%	8.04%	1.72%	2.41%	1.15%
Gender	Female	Male				
	56.37%	43.63%				
Income	Less than $25,000	$25,000 to $34,999	$35,000 to $49,999	$50,000 to $74,999	$75,000 to $99,999	$100,000 to $124,999
	18.83%	11.83%	13.89%	20.44%	14.7%	8.04%
	$125,000 to $149,999	$150,000 to $174,999	$175,000 to $199,999	$200,000 to $249,999	$250,000 or more	
	5.05%	2.18%	1.61%	1.38%	2.18%	
Neighborhood	Urban	Suburban	Rural			
	30.88%	48.11%	21.13%			
Political ideology	Extremely Liberal	Liberal	Slightly Liberal	Moderate	Slightly conservative	Conservative
	11.31%	19.01%	9.32%	28.8%	8.94%	16.83%
	Extremely conservative					
	5.8%					
Political party preference	Strong Democrat	Democrat	Independent, close to Dem.	Independent	Independent, close to Rep.	Republican
	21.96%	13.31%	11.88%	15.59%	8.46%	11.6%
	Strong Republican	Other				
	14.83%	2.38%				

Table 1. Demographic distribution of our 1,052 participants. Collectively, they represent a stratified sample of the U.S. demographic across age, gender, race, region of residence, level of education, and political identity. Note that for ethnicity, the participants could choose more than one option.

[General Social Survey: Agent Architecture Comparison (ANOVA)**]**

Source	Sum of Squares (SS)	df	F	p-value
Group	10.032	2	989.62	< .001
Residual	15.981	3153	N/A	N/A

Group 1	Group 2	Mean Difference	p-value	Lower Bound	Upper Bound	Reject Null
Demographics	Interview	0.1186	< .001	0.1113	0.1258	Yes
Demographics	Persona	-0.0021	0.7852	-0.0093	0.0052	No
Interview	Persona	-0.1206	< .001	-0.1279	-0.1133	Yes

[Big Five Personality Traits: Agent Architecture Comparison (ANOVA)**]**

Source	Sum of Squares (SS)	df	F	p-value
Group	15.298	2	39.46	< .001
Residual	611.110	3153	N/A	N/A

Group 1	Group 2	Mean Difference	p-value	Lower Bound	Upper Bound	Reject Null
Demographics	Interview	0.1705	< .001	0.1255	0.2155	Yes
Demographics	Persona	0.0843	< .001	0.0393	0.1293	Yes
Interview	Persona	-0.0862	< .001	-0.1312	-0.0412	Yes

[Economic Games: Agent Architecture Comparison (ANOVA)**]**

Source	Sum of Squares (SS)	df	F	p-value
Group	1.812	2	2.12	0.120
Residual	1344.295	3153	N/A	N/A

Group 1	Group 2	Mean Difference	p-value	Lower Bound	Upper Bound	Reject Null
Demographics	Interview	0.0585	0.0996	-0.0083	0.1253	No
Demographics	Persona	0.0333	0.4704	-0.0334	0.1001	No
Interview	Persona	-0.0252	0.6508	-0.0919	0.0416	No

Table 2. Comparative predictive performance of agents built using different descriptions. Agents constructed from interview transcripts outperformed both demographic-based and persona-based agents across multiple tasks. Specifically, interview-based agents showed significant improvement in predicting responses to the General Social Survey (in accuracy) and Big Five Personality Traits (in correlation), as confirmed by ANOVA tests ($p < 0.001$ for both). In contrast, no significant differences were observed between agent types for the Economic Games (in correlation), indicating that interviews were particularly valuable for tasks requiring deeper, personal insights.

[**General Social Survey:** Construct-Level Analysis]

General Social Survey	Accuracy	Normalized accuracy	Correlation	Normalized Correlation
natspac/y	0.33	0.45	0.17	0.25
natenvir/y	0.6	0.74	0.51	0.69
natheal/y	0.7	0.91	0.36	0.73
natcity/y	0.48	0.75	0.32	0.62
natdrug/y	0.56	0.78	0.29	0.5
nateduc/y	0.68	0.84	0.36	0.54
natrace/y	0.64	0.84	0.67	0.87
natarms/y	0.56	0.75	0.46	0.63
nataid/y	0.7	0.85	0.34	0.53
natfare/y	0.67	0.84	0.5	0.71
natroad	0.44	0.61	0.14	0.24
natsoc	0.56	0.72	0.21	0.36
natmass	0.52	0.71	0.37	0.62
natpark	0.54	0.75	0.22	0.43
natchld	0.58	0.78	0.43	0.69
natsci	0.48	0.69	0.32	0.55
natenrgy	0.54	0.78	0.43	0.7
uswary	0.51	0.62	0.15	0.24
prayer	0.75	0.91	0.4	0.7
courts	0.56	0.74	0.48	0.7
discaffw	0.43	0.7	0.37	0.59
discaffm	0.39	0.67	0.33	0.6
fehire	0.33	0.58	0.48	0.71
fechld	0.48	0.75	0.29	0.48
fepresch	0.43	0.64	0.29	0.43
fefam	0.43	0.65	0.45	0.62
fepol	0.78	0.88	0.28	0.39
reg16	0.67	0.85	0.6	0.81
mobile16	0.75	0.85	0.72	0.86
famdif16	0.89	0.97	0.75	0.89
incom16	0.53	0.71	0.53	0.73
dwelown16	0.76	0.82	0.59	0.67
paeduc*	0.86	0.96	0.66	0.85
padeg*	0.8	0.91	0.74	0.87
maeduc*	0.91	0.98	0.72	0.92
madeg*	0.84	0.97	0.8	0.96
mawrkgrw	0.97	1.05	0.91	1.2
marital	0.97	1.02	0.96	1.03
widowed	0.98	1.0	0.84	0.95
divorced	0.94	0.97	0.86	0.92
martype*	0.84	0.97	0.78	0.94
posslq/y	0.95	1.02	0.93	1.03

wrkstat	0.76	0.94	0.76	0.94
evwork	0.96	1.01	0.66	0.94
wrkgovt1	0.88	0.95	0.56	0.74
wrkgovt2	0.74	0.89	0.3	0.51
partfull	0.81	0.99	0.69	0.93
wksub1	0.83	1.05	0.68	1.06
wksup1	0.83	1.0	0.69	0.95
conarmy	0.49	0.66	0.28	0.43
conbus	0.58	0.81	0.34	0.56
conclerg	0.63	0.83	0.6	0.83
coneduc	0.56	0.82	0.28	0.47
confed	0.5	0.7	0.23	0.38
confinan	0.52	0.72	0.2	0.31
conjudge	0.54	0.74	0.41	0.59
conlabor	0.54	0.78	0.38	0.64
conlegis	0.55	0.75	0.22	0.39
conmedic	0.6	0.85	0.43	0.67
conpress	0.62	0.84	0.42	0.64
consci	0.63	0.83	0.48	0.69
contv	0.58	0.87	0.33	0.65
vetyears	0.97	0.99	0.91	1.03
joblose	0.76	0.97	0.59	0.94
jobfind	0.59	0.84	0.32	0.67
happy	0.67	0.92	0.54	0.84
hapmar	0.77	0.94	0.5	0.81
satjob	0.67	0.96	0.43	0.83
speduc*	0.84	0.98	0.8	0.96
spdeg*	0.65	0.76	0.37	0.49
spwrksta	0.74	0.88	0.56	0.78
spjew	0.71	0.88	0.08	0.13
spfund	0.76	0.94	0.42	0.74
unemp	0.74	0.89	0.45	0.68
union1	0.94	0.99	0.57	0.83
spkath/y	0.74	0.87	0.23	0.44
colath	0.64	0.8	0.19	0.4
spkrac/y	0.56	0.71	0.04	0.07
colrac	0.71	0.9	0.17	0.34
librac/y	0.54	0.72	0.14	0.29
spkcom/y	0.68	0.8	0.21	0.32
colcom/y	0.67	0.85	0.2	0.39
libcom/y	0.75	0.89	0.19	0.34
spkhomo/y	0.9	0.98	0.28	0.65
colhomo	0.83	0.88	0.21	0.38
libhomo/y	0.85	0.96	0.39	0.67
cappun	0.67	0.77	0.37	0.51
polhitok/y	0.65	0.8	0.14	0.24

polabuse/y	0.94	1.0	-0.01	-0.02
polattak/y	0.67	0.86	0.08	0.16
grass	0.73	0.77	0.31	0.36
gunlaw	0.7	0.83	0.35	0.56
owngun	0.76	0.8	0.37	0.43
hunt1	0.89	0.94	0.21	0.3
class	0.62	0.77	0.61	0.79
satfin	0.67	0.97	0.67	0.96
finalter	0.63	0.89	0.54	0.83
finrela	0.53	0.77	0.71	0.97
race*	0.93	0.95	0.92	0.95
racdif1	0.81	0.95	0.61	0.86
racdif2	0.93	1.0	0.08	0.17
racdif3	0.71	0.89	0.42	0.7
racdif4	0.75	0.9	0.38	0.63
wlthwhts	0.33	0.6	0.14	0.35
wlthblks	0.15	0.28	0.11	0.29
wlthhsps	0.31	0.62	0.06	0.15
racwork	0.62	0.84	0.37	0.61
letin1a	0.49	0.71	0.6	0.75
getahead	0.59	0.8	0.31	0.5
aged	0.53	0.75	0.15	0.31
parsol	0.36	0.68	0.46	0.71
kidssol	0.44	0.79	0.17	0.45
spanking	0.39	0.53	0.46	0.54
divlaw	0.48	0.64	0.34	0.52
sexeduc	0.78	0.84	0.34	0.47
pillok	0.44	0.67	0.49	0.64
xmarsex	0.56	0.8	0.22	0.37
homosex	0.68	0.86	0.61	0.74
marhomo	0.53	0.69	0.6	0.7
discaff	0.57	0.8	0.47	0.78
abdefect	0.83	0.9	0.41	0.58
abnomore	0.8	0.9	0.55	0.72
abhlth	0.92	0.96	0.27	0.4
abpoor	0.83	0.91	0.59	0.74
abrape	0.9	0.95	0.48	0.64
absingle	0.79	0.88	0.54	0.68
abany	0.79	0.88	0.56	0.71
letdie1	0.74	0.82	0.35	0.48
suicide1	0.79	0.89	0.4	0.56
suicide2	0.75	0.85	0.05	0.07
suicide3	0.75	0.85	0.14	0.21
suicide4	0.66	0.77	0.22	0.31
pornlaw	0.7	0.85	0.48	0.68
fair	0.43	0.63	0.18	0.3

helpful	0.42	0.65	0.24	0.44
trust	0.47	0.66	0.24	0.41
tax	0.58	0.73	0.26	0.46
vote16	0.84	0.92	0.76	0.88
pres16	0.77	0.83	0.72	0.79
if16who	0.81	0.89	0.76	0.86
polviews	0.55	0.66	0.84	0.88
partyid	0.74	0.9	0.71	0.89
news	0.31	0.46	0.37	0.48
relig*	0.76	0.85	0.63	0.73
jew	0.72	0.96	0.17	0.36
relig16*	0.62	0.7	0.55	0.63
jew16*	0.75	1.0	0.18	0.44
attend	0.56	0.75	0.75	0.84
pray	0.5	0.69	0.75	0.83
postlife	0.82	0.9	0.59	0.75
bible	0.66	0.75	0.65	0.75
reborn	0.68	0.83	0.48	0.64
savesoul	0.82	0.9	0.57	0.72
relpersn	0.63	0.79	0.79	0.91
sprtprsn	0.56	0.78	0.73	0.9
born	0.99	1.0	0.89	1.02
granborn	0.62	0.82	0.43	0.84
uscitzn*	1.0	1.0	1.0	1.27
fucitzn	0.99	1.0	0.42	1.02
mnthsusa	0.77	0.8	0.85	0.88
educ*	0.94	0.98	0.37	0.76
degree*	0.89	0.98	0.94	0.99
income	0.42	0.67	0.34	0.56
visitors	0.86	0.93	0.25	0.61
rvisitor	0.98	1.0	0.57	1.1
dwelown	0.94	1.0	0.92	1.02
zodiac	0.9	0.92	0.89	0.92
othlang	0.85	0.92	0.55	0.71
sex*	0.99	1.0	0.97	1.0
hispanic	0.98	1.01	0.92	1.03
health	0.6	0.75	0.64	0.81
compuse*	0.97	1.01	0.12	0.31
webmob	0.98	1.01	0.15	0.62
xmovie	0.66	0.75	0.23	0.31
usewww*	0.98	1.0	-0.01	-0.05
life	0.61	0.78	0.33	0.5
richwork	0.33	0.82	0.34	0.49

[Big Five: Construct-Level Analysis**]**

Big Five	MAE	-	Correlation	Correlation-Replication ratio
extraversion	0.72	-	0.45	0.51
agreeableness	0.6	-	0.35	0.43
conscientiousness	0.63	-	0.52	0.59
neuroticism	0.75	-	0.68	0.77
openness	0.62	-	0.39	0.46

[Economic Games: Construct-Level Analysis**]**

Economic Games	MAE	-	Correlation	Correlation-Replication ratio
game1_DG	0.23	-	0.11	0.24
game2_TF1	0.35	-	0.08	0.16
game3_TF2	0.17	-	0.03	0.06
game4_PG	0.52	-	-0.05	-0.71
game5_PD	0.36	-	0.1	0.32

Table 3. The construct-level analysis of predictive accuracy across the General Social Survey, Big Five personality traits, and econimic games. For each construct, we provide accuracy and correlation metrics, along with replication ratios. The analysis highlights the performance of generative agents in predicting specific dimensions within these constructs, with metrics showing varied levels of predictive accuracy across different social, personality, and game-based items. This evaluation complements the individual-level analysis, which is of our primary interest in this work, by offering a detailed look at the accuracy of agents for specific constructs and items.

[General Social Survey: Regression]

Age

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6957	0.0039	177.365	<0.001
affiliation_18 - 24	0.0029	0.0067	0.427	0.669
affiliation_25 - 34	-0.0147	0.0063	-2.343	0.019
affiliation_35 - 44	-0.0205	0.0059	-3.510	<0.001
affiliation_45 - 54	-0.0199	0.0057	-3.517	<0.001
affiliation_65 - 74	0.0108	0.0063	1.717	0.086
affiliation_75 or more	0.0224	0.0115	1.954	0.051

Note: $R^2 = 0.045$, $F(6.0, 1045.0) = 8.18$, $p < 0.001$. Reference category: 55 - 64.

Census Division

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6892	0.0043	161.841	<0.001
affiliation_e. sou. central	-0.0095	0.0069	-1.370	0.171
affiliation_foreign	-0.0060	0.0123	-0.485	0.627
affiliation_middle atlantic	0.0176	0.0067	2.643	0.008
affiliation_mountain	-0.0139	0.0092	-1.515	0.130
affiliation_new england	0.0073	0.0083	0.878	0.380
affiliation_pacific	-0.0075	0.0063	-1.187	0.235
affiliation_south atlantic	0.0032	0.0072	0.438	0.662
affiliation_w. nor. central	0.0004	0.0078	0.049	0.961
affiliation_w. sou. central	-0.0067	0.0076	-0.878	0.380

Note: $R^2 = 0.022$, $F(9.0, 1042.0) = 2.55$, $p < 0.001$. Reference category: e. nor. central.

Political Ideology

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6703	0.0031	218.263	<0.001
affiliation_conservative	-0.0081	0.0051	-1.605	0.109
affiliation_extremely conservative	-0.0007	0.0075	-0.099	0.921
affiliation_extremely liberal	0.0704	0.0058	12.181	<0.001
affiliation_liberal	0.0497	0.0049	10.208	<0.001
affiliation_slightly conservative	-0.0030	0.0063	-0.468	0.640
affiliation_slightly liberal	0.0269	0.0062	4.336	<0.001

Note: $R^2 = 0.215$, $F(6.0, 1045.0) = 47.70$, $p < 0.001$. Reference category: moderate.

Political Party

Variable	Coefficient	Std. Error	t-value	p-value
const	0.7165	0.0037	196.221	<0.001
affiliation_independent (neither)	-0.0397	0.0057	-7.006	<0.001
affiliation_independent, close to democrat	0.0022	0.0062	0.349	0.727
affiliation_independent, close to republican	-0.0516	0.0069	-7.459	<0.001
affiliation_not very strong democrat	-0.0171	0.0059	-2.884	0.004

affiliation_not very strong republican	-0.0508	0.0062	-8.178	<0.001
affiliation_other party	-0.0408	0.0117	-3.494	<0.001
affiliation_strong republican	-0.0577	0.0058	-10.024	<0.001

Note: R^2 = 0.155, F(7.0, 1044.0) = 27.28, p < 0.001. Reference category: strong democrat.

Education

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6828	0.0029	233.029	<0.001
affiliation_associate/junior college	-0.0024	0.0053	-0.448	0.655
affiliation_bachelor's	0.0140	0.0046	3.063	0.002
affiliation_graduate	0.0225	0.0056	3.985	<0.001
affiliation_less than high school	-0.0354	0.0124	-2.844	0.005

Note: R^2 = 0.034, F(4.0, 1047.0) = 9.11, p < 0.001. Reference category: high school.

Race

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6922	0.0021	327.316	<0.001
affiliation_black	-0.0208	0.0053	-3.917	<0.001
affiliation_other	-0.0067	0.0063	-1.078	0.281

Note: R^2 = 0.015, F(2.0, 1049.0) = 7.83, p < 0.001. Reference category: white.

Ethnicity

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6926	0.0021	334.816	<0.001
affiliation_American Indian or Alaskan native	-0.0055	0.0113	-0.484	0.628
affiliation_Asian	-0.0027	0.0085	-0.317	0.752
affiliation_Black/African American	-0.0183	0.0052	-3.492	<0.001
affiliation_Native Hawaiian or Pacific Islander	-0.0417	0.0268	-1.558	0.119
affiliation_Other race or ethnicity	-0.0103	0.0084	-1.232	0.218

Note: R^2 = 0.013, F(5.0, 1122.0) = 3.02, p < 0.001. Reference category: White/Caucasian.

Gender

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6909	0.0025	279.783	<0.001
affiliation_male	-0.0054	0.0037	-1.432	0.153

Note: R^2 = 0.002, F(1.0, 1050.0) = 2.05, p < 0.001. Reference category: female.

Income

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6837	0.0046	149.055	<0.001
affiliation_$100,000 to $124,999	-0.0023	0.0086	-0.263	0.792
affiliation_$125,000 to $149,999	-0.0023	0.0103	-0.221	0.825
affiliation_$150,000 to $174,999	0.0389	0.0148	2.631	0.009
affiliation_$175,000 to $199,999	-0.0106	0.0170	-0.623	0.533
affiliation_$200,000 to $249,999	0.0329	0.0183	1.800	0.072

affiliation_$25,000 to $34,999	-0.0067	0.0076	-0.889	0.374
affiliation_$250,000 or more	0.0038	0.0148	0.255	0.799
affiliation_$35,000 to $49,999	0.0109	0.0072	1.515	0.130
affiliation_$75,000 to $99,999	0.0039	0.0071	0.547	0.585
affiliation_Less than $25,000	-0.0026	0.0066	-0.394	0.694

Note: $R^2 = 0.019$, $F(10.0, 861.0) = 1.70$, $p < 0.001$. Reference category: $50,000 to $74,999.

Neighborhood

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6908	0.0030	230.902	<0.001
affiliation_Rural	-0.0122	0.0054	-2.255	0.024
affiliation_Urban	-0.0076	0.0048	-1.586	0.113

Note: $R^2 = 0.007$, $F(2.0, 869.0) = 2.91$, $p < 0.001$. Reference category: Suburban.

Sexual Orientation

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6783	0.0022	305.815	<0.001
affiliation_Asexual	0.0708	0.0155	4.574	<0.001
affiliation_Bisexual	0.0328	0.0074	4.415	<0.001
affiliation_Gay or lesbian	0.0316	0.0099	3.199	0.001
affiliation_Other sexual orientation	0.0375	0.0189	1.984	0.048
affiliation_Pansexual	0.0596	0.0131	4.539	<0.001

Note: $R^2 = 0.071$, $F(5.0, 864.0) = 13.25$, $p < 0.001$. Reference category: Heterosexual/straight.

[Big Five: Regression**]**

Age

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6712	0.0178	37.639	<0.001
affiliation_18 - 24	0.0291	0.0306	0.949	0.343
affiliation_25 - 34	0.0167	0.0285	0.588	0.557
affiliation_35 - 44	-0.0047	0.0266	-0.178	0.858
affiliation_45 - 54	-0.0030	0.0258	-0.118	0.906
affiliation_65 - 74	-0.0516	0.0287	-1.798	0.072
affiliation_75 or more	-0.1028	0.0521	-1.973	0.049

Note: $R^2 = 0.011$, $F(6.0, 1045.0) = 1.87$, $p < 0.001$. Reference category: 55 - 64.

Census Division

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6520	0.0191	34.093	<0.001
affiliation_e. sou. central	0.0585	0.0310	1.888	0.059
affiliation_foreign	-0.0095	0.0551	-0.172	0.863
affiliation_middle atlantic	-0.0208	0.0300	-0.694	0.488

affiliation_mountain	0.0361	0.0412	0.877	0.381
affiliation_new england	-0.0181	0.0374	-0.486	0.627
affiliation_pacific	0.0213	0.0283	0.753	0.452
affiliation_south atlantic	0.0313	0.0323	0.968	0.333
affiliation_w. nor. central	0.0524	0.0348	1.503	0.133
affiliation_w. sou. central	-0.0209	0.0340	-0.613	0.540

Note: R^2 = 0.011, F(9.0, 1042.0) = 1.28, p < 0.001. Reference category: e. nor. central.

Political Ideology

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6805	0.0154	44.052	<0.001
affiliation_conservative	-0.0068	0.0254	-0.267	0.790
affiliation_extremely conservative	0.0151	0.0377	0.400	0.689
affiliation_extremely liberal	-0.0241	0.0291	-0.828	0.408
affiliation_liberal	-0.0436	0.0245	-1.779	0.076
affiliation_slightly conservative	-0.0041	0.0317	-0.130	0.897
affiliation_slightly liberal	-0.0363	0.0312	-1.162	0.246

Note: R^2 = 0.005, F(6.0, 1045.0) = 0.84, p < 0.001. Reference category: moderate.

Political Party

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6656	0.0176	37.744	<0.001
affiliation_independent (neither)	0.0365	0.0274	1.333	0.183
affiliation_independent, close to democrat	-0.0464	0.0298	-1.558	0.120
affiliation_independent, close to republican	0.0105	0.0334	0.315	0.753
affiliation_not very strong democrat	-0.0379	0.0287	-1.321	0.187
affiliation_not very strong republican	-0.0101	0.0300	-0.338	0.736
affiliation_other party	0.0701	0.0564	1.242	0.215
affiliation_strong republican	0.0222	0.0278	0.799	0.424

Note: R^2 = 0.012, F(7.0, 1044.0) = 1.81, p < 0.001. Reference category: strong democrat.

Education

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6368	0.0132	48.170	<0.001
affiliation_associate/junior college	0.0504	0.0237	2.127	0.034
affiliation_bachelor's	0.0353	0.0207	1.708	0.088
affiliation_graduate	0.0416	0.0255	1.635	0.102
affiliation_less than high school	0.1890	0.0562	3.366	<0.001

Note: R^2 = 0.014, F(4.0, 1047.0) = 3.77, p < 0.001. Reference category: high school.

Race

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6742	0.0095	71.125	<0.001
affiliation_black	-0.0676	0.0238	-2.834	0.005
affiliation_other	0.0093	0.0280	0.332	0.740

Note: $R^2 = 0.008$, $F(2.0, 1049.0) = 4.27$, $p < 0.001$. Reference category: white.

Ethnicity

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6740	0.0093	72.481	<0.001
affiliation_American Indian or Alaskan native	0.0728	0.0507	1.436	0.151
affiliation_Asian	-0.0134	0.0380	-0.353	0.724
affiliation_Black/African American	-0.0650	0.0235	-2.761	0.006
affiliation_Native Hawaiian or Pacific Islander	0.1443	0.1204	1.199	0.231
affiliation_Other race or ethnicity	0.0464	0.0377	1.230	0.219

Note: $R^2 = 0.012$, $F(5.0, 1122.0) = 2.80$, $p < 0.001$. Reference category: White/Caucasian.

Gender

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6490	0.0110	58.917	<0.001
affiliation_male	0.0377	0.0167	2.264	0.024

Note: $R^2 = 0.005$, $F(1.0, 1050.0) = 5.12$, $p < 0.001$. Reference category: female.

Income

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6991	0.0204	34.276	<0.001
affiliation_$100,000 to $124,999	0.0170	0.0384	0.443	0.658
affiliation_$125,000 to $149,999	-0.0581	0.0458	-1.268	0.205
affiliation_$150,000 to $174,999	-0.0526	0.0657	-0.801	0.423
affiliation_$175,000 to $199,999	-0.0166	0.0755	-0.220	0.826
affiliation_$200,000 to $249,999	0.0787	0.0812	0.970	0.333
affiliation_$25,000 to $34,999	-0.0181	0.0337	-0.537	0.591
affiliation_$250,000 or more	-0.0505	0.0657	-0.769	0.442
affiliation_$35,000 to $49,999	-0.0496	0.0321	-1.546	0.123
affiliation_$75,000 to $99,999	-0.0571	0.0315	-1.812	0.070
affiliation_Less than $25,000	-0.0210	0.0295	-0.714	0.476

Note: $R^2 = 0.010$, $F(10.0, 861.0) = 0.88$, $p < 0.001$. Reference category: $50,000 to $74,999.

Neighborhood

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6619	0.0132	49.984	<0.001
affiliation_Rural	0.0273	0.0240	1.137	0.256
affiliation_Urban	0.0163	0.0212	0.770	0.442

Note: $R^2 = 0.002$, $F(2.0, 869.0) = 0.73$, $p < 0.001$. Reference category: Suburban.

Sexual Orientation

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6798	0.0102	66.621	<0.001
affiliation_Asexual	-0.0499	0.0712	-0.700	0.484

affiliation_Bisexual	-0.0359	0.0342	-1.050	0.294
affiliation_Gay or lesbian	0.0150	0.0455	0.329	0.742
affiliation_Other sexual orientation	-0.0853	0.0869	-0.981	0.327
affiliation_Pansexual	-0.0603	0.0605	-0.997	0.319

Note: R^2 = 0.004, F(5.0, 864.0) = 0.69, p < 0.001. Reference category: Heterosexual/straight.

[**Economic Games:** Regression]

Age

Variable	Coefficient	Std. Error	t-value	p-value
const	0.2859	0.0593	4.822	<0.001
affiliation_18 - 24	-0.0007	0.1017	-0.007	0.994
affiliation_25 - 34	-0.0081	0.0945	-0.085	0.932
affiliation_35 - 44	0.1149	0.0884	1.300	0.194
affiliation_45 - 54	0.0898	0.0857	1.049	0.295
affiliation_65 - 74	0.0104	0.0955	0.109	0.913
affiliation_75 or more	0.0504	0.1755	0.287	0.774

Note: R^2 = 0.003, F(6.0, 1041.0) = 0.54, p < 0.001. Reference category: 55 - 64.

Census Division

Variable	Coefficient	Std. Error	t-value	p-value
const	0.2977	0.0636	4.683	<0.001
affiliation_e. sou. central	0.1580	0.1032	1.531	0.126
affiliation_foreign	-0.0672	0.1827	-0.368	0.713
affiliation_middle atlantic	-0.0062	0.0998	-0.062	0.950
affiliation_mountain	0.0058	0.1368	0.042	0.966
affiliation_new england	0.0033	0.1239	0.026	0.979
affiliation_pacific	0.0994	0.0939	1.059	0.290
affiliation_south atlantic	-0.0280	0.1073	-0.261	0.794
affiliation_w. nor. central	-0.0012	0.1156	-0.010	0.992
affiliation_w. sou. central	-0.0131	0.1133	-0.116	0.908

Note: R^2 = 0.005, F(9.0, 1038.0) = 0.57, p < 0.001. Reference category: e. nor. central.

Political Ideology

Variable	Coefficient	Std. Error	t-value	p-value
const	0.3642	0.0509	7.153	<0.001
affiliation_conservative	-0.0534	0.0840	-0.635	0.526
affiliation_extremely conservative	0.2830	0.1252	2.259	0.024
affiliation_extremely liberal	-0.0917	0.0959	-0.956	0.339
affiliation_liberal	-0.0976	0.0810	-1.205	0.228
affiliation_slightly conservative	-0.0875	0.1046	-0.836	0.403
affiliation_slightly liberal	-0.0995	0.1030	-0.966	0.334

Note: R^2 = 0.010, F(6.0, 1041.0) = 1.76, p < 0.001. Reference category: moderate.

Political Party

Variable	Coefficient	Std. Error	t-value	p-value
const	0.2616	0.0587	4.456	<0.001
affiliation_independent (neither)	0.1738	0.0909	1.913	0.056
affiliation_independent, close to democrat	0.0344	0.0988	0.348	0.728
affiliation_independent, close to republican	0.0215	0.1110	0.193	0.847
affiliation_not very strong democrat	0.0117	0.0953	0.123	0.903
affiliation_not very strong republican	0.0204	0.0996	0.205	0.838
affiliation_other party	0.0712	0.1871	0.381	0.703
affiliation_strong republican	0.1708	0.0926	1.846	0.065

Note: $R^2 = 0.007$, $F(7.0, 1040.0) = 1.01$, $p < 0.001$. Reference category: strong democrat.

Education

Variable	Coefficient	Std. Error	t-value	p-value
const	0.3909	0.0440	8.887	<0.001
affiliation_associate/junior college	-0.0941	0.0787	-1.195	0.232
affiliation_bachelor's	-0.1202	0.0688	-1.748	0.081
affiliation_graduate	-0.1055	0.0850	-1.241	0.215
affiliation_less than high school	-0.0636	0.1866	-0.341	0.733

Note: $R^2 = 0.004$, $F(4.0, 1043.0) = 0.95$, $p < 0.001$. Reference category: high school.

Race

Variable	Coefficient	Std. Error	t-value	p-value
const	0.3365	0.0315	10.686	<0.001
affiliation_black	-0.0358	0.0793	-0.452	0.652
affiliation_other	-0.0610	0.0935	-0.653	0.514

Note: $R^2 = 0.001$, $F(2.0, 1045.0) = 0.28$, $p < 0.001$. Reference category: white.

Ethnicity

Variable	Coefficient	Std. Error	t-value	p-value
const	0.3335	0.0298	11.173	<0.001
affiliation_American Indian or Alaskan native	-0.0601	0.1625	-0.370	0.712
affiliation_Asian	-0.0909	0.1219	-0.746	0.456
affiliation_Black/African American	-0.0342	0.0757	-0.452	0.651
affiliation_Native Hawaiian or Pacific Islander	-0.0499	0.3859	-0.129	0.897
affiliation_Other race or ethnicity	-0.0491	0.1219	-0.403	0.687

Note: $R^2 = 0.001$, $F(5.0, 1118.0) = 0.18$, $p < 0.001$. Reference category: White/Caucasian.

Gender

Variable	Coefficient	Std. Error	t-value	p-value
const	0.3190	0.0365	8.742	<0.001
affiliation_male	0.0148	0.0554	0.266	0.790

Note: $R^2 = 0.000$, $F(1.0, 1046.0) = 0.07$, $p < 0.001$. Reference category: female.

Income

Variable	Coefficient	Std. Error	t-value	p-value
const	0.2884	0.0731	3.947	<0.001
affiliation_$100,000 to $124,999	-0.0201	0.1373	-0.147	0.883
affiliation_$125,000 to $149,999	-0.0204	0.1638	-0.125	0.901
affiliation_$150,000 to $174,999	-0.0917	0.2347	-0.391	0.696
affiliation_$175,000 to $199,999	-0.0945	0.2699	-0.350	0.726
affiliation_$200,000 to $249,999	0.0108	0.2900	0.037	0.970
affiliation_$25,000 to $34,999	0.0166	0.1205	0.138	0.890
affiliation_$250,000 or more	0.0266	0.2347	0.113	0.910
affiliation_$35,000 to $49,999	0.0252	0.1147	0.220	0.826
affiliation_$75,000 to $99,999	-0.0138	0.1128	-0.122	0.903
affiliation_Less than $25,000	0.2508	0.1054	2.381	0.018

Note: $R^2 = 0.011$, $F(10.0, 860.0) = 0.95$, $p < 0.001$. Reference category: $50,000 to $74,999.

Neighborhood

Variable	Coefficient	Std. Error	t-value	p-value
const	0.3270	0.0475	6.884	<0.001
affiliation_Rural	0.0676	0.0862	0.785	0.433
affiliation_Urban	-0.0300	0.0760	-0.394	0.693

Note: $R^2 = 0.001$, $F(2.0, 868.0) = 0.56$, $p < 0.001$. Reference category: Suburban.

Sexual Orientation

Variable	Coefficient	Std. Error	t-value	p-value
const	0.3417	0.0365	9.366	<0.001
affiliation_Asexual	-0.0639	0.2545	-0.251	0.802
affiliation_Bisexual	-0.0330	0.1222	-0.270	0.787
affiliation_Gay or lesbian	-0.0567	0.1624	-0.349	0.727
affiliation_Other sexual orientation	-0.0750	0.3107	-0.241	0.809
affiliation_Pansexual	-0.0866	0.2160	-0.401	0.689

Note: $R^2 = 0.000$, $F(5.0, 863.0) = 0.08$, $p < 0.001$. Reference category: Heterosexual/straight.

Table 4. Regression analyses of predictive performance across demographic variables. This table presents the results of regression analyses examining the influence of demographic variables on the predictive performance of generative agents, across the General Social Survey, Big Five Personality Traits, and economic games. Separate models were run for each demographic factor, with predictive performance as the dependent variable. Coefficients represent the relative difference in performance for each demographic group, with positive values indicating higher predictive accuracy. Significant discrepancies are observed for political ideology, party affiliation, and sexual orientation, with stronger performance for participants identifying as strong liberals, strong Democrats, and non-heterosexual/straight individuals.

[General Social Survey: Demographic Parity Difference**]**

	Agents w/ Interview	Agents w/ Demog. Info.	Agents w/ Persona Desc.
Age	DPD=4.30 Min: 35 - 44 (67.51%) Max: 75 or more (71.81%)	DPD=4.76 Min: 45 - 54 (55.3%) Max: 75 or more (60.06%)	DPD=3.23 Min: 45 - 54 (55.98%) Max: 18 - 24 (59.21%)
Census Division	DPD=3.16 Min: mountain (67.52%) Max: middle atlantic (70.68%)	DPD=2.74 Min: foreign (55.51%) Max: middle atlantic (58.25%)	DPD=3.99 Min: foreign (54.76%) Max: middle atlantic (58.75%)
Political Ideology	DPD=7.86 Min: conservative (66.22%) Max: extremely liberal (74.07%)	DPD=12.35 Min: conservative (50.74%) Max: extremely liberal (63.09%)	DPD=11.91 Min: extremely conservative (50.43%) Max: extremely liberal (62.34%)
Political Party	DPD=5.98 Min: strong republican (65.89%) Max: independent, close to democrat (71.87%)	DPD=9.79 Min: strong republican (51.96%) Max: independent, close to democrat (61.74%)	DPD=9.34 Min: strong republican (52.08%) Max: independent, close to democrat (61.41%)
Education	DPD=5.79 Min: less than high school (64.74%) Max: graduate (70.52%)	DPD=5.15 Min: less than high school (53.2%) Max: graduate (58.35%)	DPD=6.41 Min: less than high school (52.28%) Max: graduate (58.7%)
Race	DPD=2.08 Min: black (67.13%) Max: white (69.22%)	DPD=3.33 Min: black (54.19%) Max: white (57.52%)	DPD=3.19 Min: black (54.13%) Max: white (57.32%)
Ethnicity	DPD=2.15 Min: Other race or ethnicity (67.11%) Max: White/Caucasian (69.26%)	DPD=4.42 Min: Other race or ethnicity (53.77%) Max: Native Hawaiian or Pacific Islander (58.19%)	DPD=3.63 Min: Other race or ethnicity (53.75%) Max: White/Caucasian (57.37%)
Gender	DPD=0.54 Min: male (68.55%) Max: female (69.09%)	DPD=0.56 Min: male (56.68%) Max: female (57.24%)	DPD=0.61 Min: male (56.45%) Max: female (57.06%)
Income	DPD=4.95 Min: $175,000 to $199,999 (67.31%) Max: $150,000 to $174,999 (72.26%)	DPD=5.43 Min: Less than $25,000 (55.4%) Max: $200,000 to $249,999 (60.83%)	DPD=6.61 Min: Less than $25,000 (55.21%) Max: $200,000 to $249,999 (61.82%)
Neighborhood	DPD=1.22 Min: Rural (67.85%) Max: Suburban (69.08%)	DPD=1.62 Min: Urban (56.0%) Max: Suburban (57.62%)	DPD=2.89 Min: Rural (54.68%) Max: Suburban (57.57%)
Sexual Orientation	DPD=7.08 Min: Heterosexual/straight (67.83%) Max: Asexual (74.92%)	DPD=7.50 Min: Heterosexual/straight (55.85%) Max: Pansexual (63.36%)	DPD=9.88 Min: Heterosexual/straight (55.73%) Max: Asexual (65.61%)

[Big Five: Demographic Parity Difference**]**

	Agents w/ Interview	**Agents w/ Demog. Info.**	**Agents w/ Persona Desc.**
Age	DPD=0.22 Min: 18 - 24 (0.56) Max: 75 or more (0.78)	DPD=0.37 Min: 18 - 24 (0.23) Max: 65 - 74 (0.6)	DPD=0.33 Min: 18 - 24 (0.37) Max: 65 - 74 (0.7)
Census Division	DPD=0.19 Min: mountain (0.49) Max: w. sou. central (0.68)	DPD=0.24 Min: mountain (0.36) Max: foreign (0.6)	DPD=0.23 Min: mountain (0.39) Max: foreign (0.63)
Political Ideology	DPD=0.06 Min: moderate (0.6) Max: extremely conservative (0.67)	DPD=0.17 Min: extremely liberal (0.34) Max: moderate (0.51)	DPD=0.14 Min: extremely liberal (0.49) Max: slightly conservative (0.63)
Political Party	DPD=0.17 Min: other party (0.49) Max: independent, close to democrat (0.66)	DPD=0.22 Min: other party (0.3) Max: independent, close to democrat (0.53)	DPD=0.22 Min: other party (0.39) Max: strong republican (0.61)
Education	DPD=0.23 Min: less than high school (0.44) Max: graduate (0.68)	DPD=0.24 Min: less than high school (0.3) Max: graduate (0.54)	DPD=0.3 Min: less than high school (0.33) Max: graduate (0.63)
Race	DPD=0.11 Min: other (0.59) Max: black (0.7)	DPD=0.17 Min: other (0.4) Max: black (0.57)	DPD=0.08 Min: other (0.52) Max: black (0.6)
Ethnicity	DPD=0.21 Min: Native Hawaiian or Pacific Islander (0.49) Max: Black/African American (0.71)	DPD=0.62 Min: Native Hawaiian or Pacific Islander (-0.05) Max: Black/African American (0.57)	DPD=0.42 Min: Native Hawaiian or Pacific Islander (0.31) Max: American Indian or Alaskan native (0.73)
Gender	DPD=0.01 Min: male (0.62) Max: female (0.63)	DPD=0.01 Min: female (0.46) Max: male (0.46)	DPD=0.02 Min: female (0.54) Max: male (0.55)
Income	DPD=0.32 Min: $200,000 to $249,999 (0.41) Max: $250,000 or more (0.73)	DPD=0.29 Min: $200,000 to $249,999 (0.31) Max: $250,000 or more (0.6)	DPD=0.38 Min: $200,000 to $249,999 (0.35) Max: $150,000 to $174,999 (0.73)
Neighborhood	DPD=0.03 Min: Urban (0.58) Max: Rural (0.62)	DPD=0.05 Min: Rural (0.43) Max: Urban (0.48)	DPD=0.04 Min: Suburban (0.51) Max: Urban (0.54)
Sexual Orientation	DPD=0.23 Min: Other sexual orientation (0.43) Max: Pansexual (0.66)	DPD=0.29 Min: Other sexual orientation (0.17) Max: Heterosexual/straight (0.46)	DPD=0.45 Min: Other sexual orientation (0.1) Max: Heterosexual/straight (0.55)

[Economic Games: Demographic Parity Difference**]**

	Agents w/ Interview	Agents w/ Demog. Info.	Agents w/ Persona Desc.
Age	DPD=0.18 Min: 75 or more (0.21) Max: 45 - 54 (0.39)	DPD=0.19 Min: 35 - 44 (0.18) Max: 55 - 64 (0.36)	DPD=0.8 Min: 35 - 44 (0.15) Max: 55 - 64 (0.94)
Census Division	DPD=0.2 Min: mountain (0.25) Max: foreign (0.45)	DPD=0.21 Min: e. nor. central (0.19) Max: new england (0.4)	DPD=0.65 Min: mountain (0.15) Max: pacific (0.8)
Political Ideology	DPD=0.19 Min: conservative (0.24) Max: extremely liberal (0.44)	DPD=0.5 Min: extremely conservative (-0.03) Max: extremely liberal (0.47)	DPD=0.53 Min: moderate (0.37) Max: slightly liberal (0.91)
Political Party	DPD=0.22 Min: independent (neither) (0.21) Max: strong democrat (0.43)	DPD=0.4 Min: strong republican (0.03) Max: strong democrat (0.43)	DPD=0.58 Min: independent (neither) (0.31) Max: independent, close to democrat (0.88)
Education	DPD=0.14 Min: high school (0.29) Max: less than high school (0.43)	DPD=0.12 Min: associate/junior college (0.21) Max: graduate (0.33)	DPD=0.44 Min: less than high school (0.31) Max: graduate (0.76)
Race	DPD=0.04 Min: black (0.3) Max: white (0.34)	DPD=0.04 Min: other (0.24) Max: white (0.29)	DPD=0.42 Min: other (0.37) Max: black (0.79)
Ethnicity	DPD=0.52 Min: Other race or ethnicity (0.06) Max: Native Hawaiian or Pacific Islander (0.58)	DPD=0.63 Min: Other race or ethnicity (0.07) Max: Native Hawaiian or Pacific Islander (0.7)	DPD=1.54 Min: Asian (0.12) Max: Native Hawaiian or Pacific Islander (1.66)
Gender	DPD=0.03 Min: female (0.32) Max: male (0.35)	DPD=0.03 Min: female (0.27) Max: male (0.29)	DPD=0.05 Min: female (0.55) Max: male (0.6)
Income	DPD=0.51 Min: $200,000 to $249,999 (0.21) Max: $175,000 to $199,999 (0.72)	DPD=0.46 Min: $200,000 to $249,999 (0.19) Max: $150,000 to $174,999 (0.65)	DPD=0.71 Min: $125,000 to $149,999 (0.22) Max: $175,000 to $199,999 (0.93)
Neighborhood	DPD=0.14 Min: Urban (0.25) Max: Suburban (0.38)	DPD=0.11 Min: Urban (0.2) Max: Suburban (0.32)	DPD=0.52 Min: Urban (0.35) Max: Rural (0.88)
Sexual Orientation	DPD=0.32 Min: Asexual (0.23) Max: Other sexual orientation (0.55)	DPD=0.41 Min: Heterosexual/straight (0.24) Max: Other sexual orientation (0.66)	DPD=0.74 Min: Bisexual (0.2) Max: Pansexual (0.93)

Table 5. Demographic Parity Difference (DPD) Results. This table summarizes the results of regression analyses measuring demographic parity differences (DPD) across three tasks (GSS, Big Five, and economic games) for agents using demographic information, interview data, and persona-based profiles. Interview-based agents consistently reduced bias compared to demographic-based agents across political ideology, race, and gender.

[Robustness Analysis]

	GSS	GSS Num.	Big Five	Economic Games
Interview	Acc.=67.96 (std=6.67) Nrm. Acc.=0.85 (std=0.11) -- Correl.=0.65 (z std=0.2) Nrm. Correl.=0.85 (std=0.32)	MAE=0.14 (std=0.15) -- Correl.=0.97 (z std=0.82) Nrm. Correl.=2.16 (std=9.17)	MAE=0.67 (std=0.27) -- Correl.=0.79 (z std=0.73) Nrm. Correl.=0.77 (std=0.76)	MAE=0.51 (std=2.01) -- Correl.=0.52 (z std=0.9) Nrm. Correl.=0.33 (std=1.22)
Survey and Experiments	Acc.=61.08 (std=7.0) Nrm. Acc.=0.76 (std=0.12) -- Correl.=0.55 (z std=0.18) Nrm. Correl.=0.72 (std=0.33)	MAE=0.32 (std=0.19) -- Correl.=0.75 (z std=0.62) Nrm. Correl.=1.74 (std=8.94)	MAE=0.68 (std=0.25) -- Correl.=0.71 (z std=0.75) Nrm. Correl.=0.64 (std=0.61)	MAE=0.47 (std=2.0) -- Correl.=0.52 (z std=0.93) Nrm. Correl.=0.31 (std=1.22)
Maximal	Acc.=68.0 (std=6.65) Nrm. Acc.=0.85 (std=0.12) -- Correl.=0.65 (z std=0.2) Nrm. Correl.=0.85 (std=0.34)	MAE=0.28 (std=0.21) -- Correl.=0.78 (z std=0.71) Nrm. Correl.=1.85 (std=9.19)	MAE=0.68 (std=0.26) -- Correl.=0.77 (z std=0.75) Nrm. Correl.=0.72 (std=0.78)	MAE=0.47 (std=2.0) -- Correl.=0.55 (z std=0.93) Nrm. Correl.=0.3 (std=1.29)
Summary	Acc.=66.8 (std=6.58) Nrm. Acc.=0.83 (std=0.12) -- Correl.=0.64 (z std=0.2) Nrm. Correl.=0.84 (std=0.36)	MAE=0.19 (std=0.18) -- Correl.=0.93 (z std=0.82) Nrm. Correl.=0.92 (std=2.77)	MAE=0.69 (std=0.3) -- Correl.=0.76 (z std=0.76) Nrm. Correl.=0.7 (std=0.67)	MAE=0.5 (std=2.0) -- Correl.=0.49 (z std=0.85) Nrm. Correl.=0.41 (std=1.1)
Random lesion (0% removal)	Acc.=67.96 (std=6.67) Nrm. Acc.=0.85 (std=0.11) -- Correl.=0.65 (z std=0.2) Nrm. Correl.=0.85 (std=0.32)	MAE=0.14 (std=0.15) -- Correl.=0.97 (z std=0.82) Nrm. Correl.=2.16 (std=9.17)	MAE=0.67 (std=0.27) -- Correl.=0.79 (z std=0.73) Nrm. Correl.=0.77 (std=0.76)	MAE=0.51 (std=2.01) -- Correl.=0.52 (z std=0.9) Nrm. Correl.=0.33 (std=1.22)
Random lesion (20% removal)	Acc.=67.3 (std=6.94) Nrm. Acc.=0.84 (std=0.12) -- Correl.=0.64 (z std=0.2) Nrm. Correl.=0.84 (std=0.33)	MAE=0.16 (std=0.13) -- Correl.=0.95 (z std=0.81) Nrm. Correl.=1.46 (std=3.71)	MAE=0.66 (std=0.28) -- Correl.=0.79 (z std=0.74) Nrm. Correl.=0.75 (std=0.72)	MAE=0.52 (std=2.02) -- Correl.=0.58 (z std=0.89) Nrm. Correl.=0.46 (std=1.18)
Random lesion (40% removal)	Acc.=66.5 (std=6.79) Nrm. Acc.=0.83 (std=0.12) -- Correl.=0.63 (z std=0.2) Nrm. Correl.=0.82 (std=0.32)	MAE=0.16 (std=0.14) -- Correl.=0.96 (z std=0.75) Nrm. Correl.=2.14 (std=9.15)	MAE=0.65 (std=0.28) -- Correl.=0.78 (z std=0.74) Nrm. Correl.=0.75 (std=0.74)	MAE=0.52 (std=2.02) -- Correl.=0.63 (z std=0.9) Nrm. Correl.=0.5 (std=1.21)
Random lesion (60% removal)	Acc.=66.44 (std=6.69) Nrm. Acc.=0.83 (std=0.12) -- Correl.=0.63 (z std=0.18) Nrm. Correl.=0.82 (std=0.33)	MAE=0.16 (std=0.12) -- Correl.=0.95 (z std=0.81) Nrm. Correl.=1.96 (std=9.02)	MAE=0.65 (std=0.29) -- Correl.=0.77 (z std=0.74) Nrm. Correl.=0.73 (std=0.68)	MAE=0.48 (std=2.0) -- Correl.=0.57 (z std=0.91) Nrm. Correl.=0.41 (std=1.33)
Random lesion (80% removal)	Acc.=63.7 (std=6.96) Nrm. Acc.=0.79 (std=0.11) -- Correl.=0.6 (z std=0.19) Nrm. Correl.=0.76 (std=0.25)	MAE=0.18 (std=0.15) -- Correl.=0.96 (z std=0.76) Nrm. Correl.=2.0 (std=9.09)	MAE=0.64 (std=0.26) -- Correl.=0.75 (z std=0.72) Nrm. Correl.=0.73 (std=0.74)	MAE=0.52 (std=2.01) -- Correl.=0.62 (z std=0.93) Nrm. Correl.=0.46 (std=1.19)

Table 6. Robustness analysis. This table presents the results of an exploratory robustness analysis comparing different agent architectures informed by various data sources, including interviews, surveys, experiments, and summaries. Performance is evaluated across four constructs: the General Social Survey (GSS), GSS Numeric (GSS Num.), Big Five Personality Traits (Big Five), and economic games. Interview-based agents consistently outperform others, achieving high accuracy (0.85, std=0.11) on the GSS, indicating that interviews provide richer, more comprehensive information than surveys and experiments. Maximal agents, which integrate data from all sources, show similar performance. Summary agents perform slightly below interview-based agents, with minor losses in accuracy (0.83, std = 0.12). A progressive decline is observed for random lesion agents as portions of interview data are removed, with accuracy dropping from 0.85 to 0.79 as 80% of the utterances are excluded, suggesting that even shortened interviews retain valuable insights compared to survey-only agents.

[Interview Script]

Script	Time Limit (sec)
Hi, <participant's name>! My name is Isabella, and I'm an AI assistant who will be conducting today's interview with you. Thank you so much for choosing to participate in our study! Before we start, please take a moment to familiarize yourself with the controls on your screen. The button labeled 'Show Isabella's Subtitles' will display subtitles for my dialogue, while the pause button will take you back to the homepage that led you to this interview screen. Please note, however, that this interview is intended to be completed in one sitting. While it's okay to pause the interview if unexpected events occur or if you need a break, we recommend trying to complete the interview in one sitting. Please also be aware that if you pause the interview, you will be asked to resume from the last saved checkpoint. You will know when you have passed a checkpoint by a flash of green light around the circle displaying my avatar, accompanied by a message that says 'Saved.' If we lose connection for more than a minute during the interview, a red button will appear, allowing you to refresh the page and resume from the last saved checkpoint. This is a semi-structured interview expected to take roughly two hours. For completeness, I may ask a few questions that might seem repetitive along the way. As described in the consent form, this interview will cover your life experiences as well as your views on various social topics. The content of this interview will be a part of our dataset that will be shared with scientific researchers to better understand the lived experiences of those being interviewed. Finally, you may speak when a microphone icon appears in the circle displaying my avatar. Please stop the interview now if any of this is not ok. If all this sounds good, let's get started!	0
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.	625
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?	235
Some people tell us they made a conscious choice or decision in moments like these, while others say it 'just happened'. What about you?	50
Do you think another person or an organization could have lent a helping hand during moments like this?	40
Moving to present time, tell me more about family who are important to you. Do you have a partner, or children?	310
Are there anyone else in your immediate family whom you have not mentioned? Who are they, and what is your relationship with them like?	25
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?	80
Now let's talk about your current neighborhood. Tell me all about the neighborhood and area in which you are living now.	105
Some people say they feel really safe in their neighborhoods, others not so much. How about for you?	40
Living any place has its ups and downs. Tell me about what it's been like for you living here.	40
Tell me about the people who live with you right now, even people who are staying here temporarily.	80
Is anyone in your household a temporary member, or is everyone a permanent member of the household?	10
Tell me about anyone else who stays here from time to time, if there is anyone like that	10
Is there anyone who usually lives here but is away – traveling, at school, or in a hospital?	10

Right now, across a typical week, how do your days vary?	105
At what kind of job or jobs do you work, and when do you work?	105
Do you have other routines or responsibilities that you did not already share?	80
Tell me about any recent changes to your daily routine.	80
If you have children, tell me about a typical weekday during the school year for your children. What is their daily routine? And what are after school activities your children participate in?	80
Some people we've talked to tell us about experiences with law enforcement. How about for you?	80
What other experiences with law enforcement stand out in your mind?	40
Some people tell us about experiences of arrest – of loved ones, family members, friends, or themselves. How about for you?	40
Some people say they vote in every election, some tell us they don't vote at all. How about you?	80
How would you describe your political views?	310
Tell me about any recent changes in your political views.	115
One topic a lot of people have been talking about recently is race and/or racism and policing. Some tell us issues raised by the Black Lives Matter movement have affected them a lot, some say they've affected them somewhat, others say they haven't affected them at all. How about for you?	105
How have you been thinking about the issues Black Lives Matter raises?	105
How have you been thinking about race in the U.S. recently?	105
How have you responded to the increased focus on race and/or racism and policing? Some people tell us they're aware of the issue but keep their thoughts to themselves, others say they've talked to family and friends, others have joined protests. What about you?	50
What about people you're close to? How have they responded to the increased focus on race and/or racism and policing?	80
Now we'd like to learn more about your health. First, tell me all about your health.	80
For you, what makes it easy or hard to stay healthy?	65
Tell me about anything big that has happened in the past two years related to your health: any medical diagnoses, flare-ups of chronic conditions, broken bones, pain – anything like that.	80
Sometimes, health problems get in the way. They can even affect people's ability to work or care for their children. How about you?	50
Sometimes, it's not your health problem, but the health of a loved one. Has this been an issue for you?	80
Tell me what it has been like trying to get the health care you or your immediate family need. Have you ever had to forgo getting the health care you need?	80
Have you or your immediate family ever used alternative forms of medicine? This might include indigenous, non-western, or informal forms of care.	80
During tough times, some people tell us they cope by smoking or drinking. How about for you?	80
Other people say they cope by relying on prescriptions, pain medications, marijuana, or other substances. How about for you and can you describe your most recent experience of using them, if any?	80
How are you and your family coping with, or paying for, your health care needs right now?	80
Some people are excited about medical vaccination, and others, not so much. How about you?	50
What are your trusted sources of information about the vaccine?	80
Now we're going to talk a bit more about what life was like for you over the past year. Tell me all about how	115

you have been feeling.	
Tell me a story about a time in the last year when you were in a rough place or struggling emotionally.	80
Some people say they struggle with depression, anxiety, or something else like that. How about for you?	80
How has it been for your family?	50
Some people say that religion or spirituality is important in their lives, for others not so much. How about you?	155
Some people tell us they use Facebook, Instagram, or other social media to stay connected with the wider world. How about for you? Tell me all about how you use these types of platforms.	115
Some people say they use Facebook or Twitter to get (emotional or financial) support during tough times. What about you?	50
Tell me about any recent changes in your level of stress, worry, and your emotional coping strategies.	50
This might be repetitive, but before we go on to the next section, I want to quickly make sure I have this right: who are you living with right now?	25
Any babies or small children?	10
Anyone who usually lives here but is away – traveling, at school, or in a hospital?	10
Any lodgers, boarders, or employees who live here?	10
Who shares the responsibility for the rent, mortgage, or other household expenses?	5
Now we'd like to talk about how you make ends meet and what the monthly budget looks like for you and your family. What were your biggest expenses last month?	50
How much did your household spend in the past month? Is that the usual amount? And if not, how much do you usually spend and on what?	50
Does your household own or rent your home?	50
Bills can fluctuate over the course of a year. Seasons change, there are holidays and special events, school starts and ends, and so on. Tell me all about how your bills have fluctuated over the past year.	50
Tell me all about how you coped with any extra expenses in the past months.	40
Some people have a savings account, some people save in a different way, and some people say they don't save. How about you?	80
Some people save for big things, like a home, while others save for a rainy day. How about for you (in the last year)?	80
Some people have student loans or credit card debt. Others take out loans from family or friends or find other ways of borrowing money. Tell me about all the debts you're paying on right now.	50
Tell me about any time during the past year that you haven't had enough money to buy something that you needed or pay a bill that was due.	65
What is or was your occupation? In this occupation, what kind of work do you do and what are the most important activities or duties?	155
Was your occupation covered by a union or employee association contract?	40
In the past year, how many weeks did you work (for a few hours, including paid vacation, paid sick leave, and military service)?	40
How many hours each week did you usually work?	15
In the past week, did you work for pay (even for 1 hour) at a job or business?	15
Did you have more than one job or business, including part time, evening, or weekend work?	25

Now, in the past month, how many jobs did you work, including part time, evening, or weekend work?	25
For your job or occupations, how much money did you receive in the past month?	50
How often do you get paid?	25
In total, how much did your household make in the past month?	50
Switching gears a bit, let's talk a little about tax time. Did you file taxes last year (the most recent year)?	40
Tell me more about your workplace. How long have you been at your current job? How would you describe the benefits that come with your job? (This could include things like health benefits, paid time off, vacation, and sick time.)	15
How would you describe your relationships at work? (How is your relationship with your manager or boss? How are your relationships with your coworkers?)	15
Tell me about how predictable your work schedule is. How would you describe your job in terms of flexibility with your hours?	80
Some people say it's hard to take or keep a job because of child care. How about for you?	15
Do you receive any payments or benefits from SNAP, food stamps, housing voucher payments, supplemental security income, or any other programs like that?	25
Tell me about times over the last year when your income was especially low. And tell me about all the things you did to make ends meet during that time.	50
What would it be like for you if you had to spend $400 for an emergency? Would you have the money, and if not, how would you get it?	50
Overall, how do you feel about your financial situation?	105
Are you now married, widowed, divorced, separated, or have you never been married? If you are not currently married, are you currently living with a romantic partner?	40
For our records, we also need your date of birth. Could you please provide that?	25
Were you born in the United States? If you were not born in the U.S., what country were you born, and what year did you first come to the U.S. to live?	25
What city and state were you born in?	25
Are you of Hispanic, Latinx, or Spanish origin?	15
What race or races do you identify with?	15
What is the highest degree or grade you've completed?	15
Are you enrolled in school?	15
Have you been enrolled in school during the past 3 months?	15
Have you ever served on active duty in the U.S. Armed Forces, Reserves, or National Guard?	15
What religion do you identify with, if any?	15
Generally speaking, do you usually think of yourself as a Democrat, a Republican, an Independent, or what? And how strongly do you associate with that party?	80
Do you think of yourself as closer to the Democratic Party or to the Republican Party?	50
What was the city and state that you lived in when you were 16 years old?	25
Did you live with both your own mother and father when you were 16?	25
Who else did you live with?	25
What's the highest degree or grade your dad completed?	15

What's the highest degree or grade your mom completed?	15
Did your mom work for pay for at least a year while you were growing up?	15
What was her job? What kind of work did she do? What were her most important activities or duties?	50
What kind of place did she work for? What kind of business was it? What did they make or do where she worked?	50
Did your dad work for pay for at least a year while you were growing up?	15
What was his job? What kind of work did he do? What were his most important activities or duties?	50
What kind of place did he work for? What kind of business was it? What did they make or do where he worked?	50
We all have hopes about what our future will look like. Imagine yourself a few years from now. Maybe you want your life to be the same in some ways as it is now. Maybe you want it to be different in some ways. What do you hope for?	155
What do you value the most in your life?	80
And that was the last question I wanted to ask today. Thank you so much again for your time <participant's name>. It was really wonderful getting to know you through this interview. Now, we will take you back to the Home Screen so that you may finish the rest of the study!	0

Table 7. The interview script used to guide two-hour conversations with participants, adapted from the American Voices Project's interview. The script covers a wide range of topics, from participants' life stories to their views on social, political, and personal values. Select portions were abbreviated to ensure a manageable session length, while still capturing the breadth of experiences and perspectives essential for building nuanced generative agents.

General Social Survey	Accuracy	Normalized Accuracy	Correlation	Normalized Correlation
Participant Replication	*81.25% (std=8.11)*	*1.00 (std=0.00)*	*0.83 (std=0.30)*	*1.00 (std=0.00)*
Agents w/ Interview	68.85% (std=6.01)	0.85 (std=0.11)	0.66 (std=0.19)	0.83 (std=0.31)
Agents w/ Demog. Info.	57.00% (std=7.45)	0.71 (std=0.11)	0.51 (std=0.19)	0.63 (std=0.26)
Agents w/ Persona Desc.	56.79% (std=7.76)	0.70 (std=0.11)	0.50 (std=0.20)	0.62 (std=0.25)

Big Five	Mean Absolute Error	-	Correlation	Normalized Correlation
Participant Replication	*0.30 (std=0.17)*	-	*0.95 (std=0.76)*	*1.00 (std=0.00)*
Agents w/ Interview	0.67 (std=0.27)	-	0.78 (std=0.70)	0.80 (std=1.88)
Agents w/ Demog. Info.	0.76 (std=0.32)	-	0.61 (std=0.70)	0.55 (std=2.25)
Agents w/ Persona Desc.	0.74 (std=0.32)	-	0.71 (std=0.73)	0.75 (std=2.59)

Economic Games	Mean Absolute Error	-	Correlation	Normalized Correlation
Participant Replication	*0.25 (std=0.88)*	-	*0.99 (std=1.00)*	*1.00 (std=0.00)*
Agents w/ Interview	0.32 (std=0.89)	-	0.66 (std=0.95)	0.66 (std=2.83)
Agents w/ Demog. Info.	0.34 (std=0.89)	-	0.57 (std=0.91)	0.48 (std=2.90)
Agents w/ Persona Desc.	0.33 (std=0.88)	-	0.60 (std=0.93)	0.57 (std=2.84)

Table 8. Generative agents' predictive performance. The consistency rate between participants and the predictive performance of generative agents is evaluated across various constructs and averaged across individuals. For the General Social Survey (GSS), accuracy is reported due to its categorical response types, while the Big Five personality traits and economic games report mean absolute error (MAE) due to their numerical response types. Correlation is reported for all constructs. Normalized accuracy is provided for all metrics, except for MAE, which cannot be calculated for individuals whose MAE is 0 (i.e., those who responded the same way in both phases). We find that generative agents predict participants' behavior and attitudes well, especially when compared to participants' own rate of internal consistency. Additionally, using interviews to inform agent behavior significantly improves the predictive performance of agents for both GSS and Big Five constructs, outperforming other commonly used methods in the literature.