Title

User Simulation via Language Models

ISBN

9798189276132

Author

Kang, Minwoo

Publication Date

2026-05-01

Peer reviewed|Thesis/dissertation

User Simulation via Language Models

By
Minwoo Kang

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Computer Science
in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:
Professor John Wawrzynek, Co-Chair
Professor John Canny, Co-Chair
Assistant Professor Sagar Karandikar
Assistant Professor William D. Thompson

Spring 2026

User Simulation via Language Models

Abstract

User Simulation via Language Models

by

Minwoo Kang

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor John Wawrzynek, Co-Chair

Professor John Canny, Co-Chair

This dissertation develops a unified framework for using large language models (LLMs) as simulators of human users—both as synthetic participants in behavioral and social-scientific research and as synthetic interlocutors for the training and evaluation of AI assistant systems. The central position of this work is that user simulation is a fundamentally distinct modeling problem from assistant alignment, and that pretrained base models, which we frame as *Human–Language Mixture Models* (HLMMs), are the appropriate substrate for it. Trained via cross-entropy on web-scale corpora authored by hundreds of millions of distinct individuals, pretrained models internalize a "mixture of voices" that reflects the diversity of human linguistic, social, and pragmatic behavior. Instruction-tuned assistant language models (ALMs), by contrast, are further optimized under objectives that systematically suppress the breadth, ambiguity, and diverse behaviors that real users exhibit. A central claim of this dissertation is that much of the "failure" of LLM-based user simulation reported in the recent literature is in fact a failure of taking the wrong approach of instructing an ALM to role-play a user, rather than a limitation of language models in general.

Building on this argument, the dissertation introduces a unifying mechanism for binding pretrained LLMs to coherent virtual personas. We propose conditioning language models with unconstrained and natural life narratives, which we refer to as *backstories*, generated by base models themselves. Backstories both explicitly and implicitly encodes diverse details about its author, including demographic information, psychological context, and human beliefs, values, and perspectives. When conditioned with backstories, language models are bound to a naturalistic persona of a particular individual, and throughout this work we show how backstory conditioning improves realism, representativeness, consistency, and behavioral fidelity in simulated populations than existing methods that rely on prescriptive persona design.

Chapter 2 formalizes the distinction between HLMMs and ALMs and provides experimental evidence across diverse conversational corpora that pretrained base models better estimate the conditional distribution of human utterances, better reflect pragmatic dialogue structure, and better preserve lexical and semantic diversity than their instruction-tuned counterparts.

We additionally introduce *Tandem Models*, an inference-time procedure that pairs a base-model utterance generator with an instruction-tuned supervisor.

Chapter 3 introduces the *Anthology* framework, which generates diverse first-person backstories through simple open-ended prompting (e.g., "Tell me about yourself") and aligns them to target demographic distributions via maximum-weight or greedy bipartite matching. When conditioned on these backstories, base LLMs reproduce population-level opinion distributions from the Pew Research Center's American Trends Panel with smaller distributional shifts between human and model responses and higher internal consistency than existing persona-conditioning methods, including for under-represented demographic subgroups.

Chapter 4 extends the *Anthology* framework to model social identity and group perception. We test whether backstory-conditioned LLMs are capable of *deep* persona binding: responding as authentic in-group members would rather than shallow imitation of out-group stereotypes. Longer, more coherent backstories generated through a minimal Tandem modeling approach—multi-turn interview prompting pretrained models and rejection sampling by an instruction-tuned LLM-as-a-critic—yield richer and more consistent virtual personas. These backstory-conditioned LLMs reproduce partisan asymmetries in moral judgment, expectations of democratic backsliding, and exaggerated meta-perceptions observed in human data, demonstrating that narrative depth and consistency are essential for capturing identity-driven perspectives.

Chapter 5 then turns to applications. It applies backstory conditioning to *action prediction* in partisan social-dilemma games (Dictator and Trust), introducing *Temporal Grounding* and *Consistency Filtering* as supplementary prompting strategies for situated behavioral simulation, and reproduces human partisan favoritism across replications spanning more than a decade. The same setting enables a $2 \times 2 \times 2$ counterfactual decomposition of year, framing, and recruitment-population effects that no single human experiment can supply. The chapter then frames *design problems* as the natural application surface for the methodology developed in this dissertation, identifies three structural challenges that distinguish AI assistance in design from instruction-following, and positions digital hardware design as a flagship use-case where faithful user simulation could be applied to building collaborative AI agents.

Together, these studies establish backstory-conditioned pretrained LLMs as a principled foundation for scalable, responsible behavioral simulation and as a faithful substrate for synthetic interlocutors in agent evaluation.

To my family, friends, and mentors.
And to the curious reader who discovered this work.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my deepest gratitude to my advisors, Professor John Wawrzynek and Professor John Canny, for their unwavering guidance, patience, and support throughout my time at Berkeley. I am also deeply thankful to my dissertation committee members, Professor Sagar Karandikar and Professor Bill Thompson: their insights helped refine both the technical depth and the broader dimensions of this work.

It took a village of numerous talented collaborators to make this dissertation possible. I would like to thank Suhong Moon, Sehoon Kim, David Chan, and Joseph Suh for our close collaborations on various projects. I am also grateful to Coleman Hooper, Qijing (Jenny) Huang, Professor Sophia Shao for many early projects on hardware acceleration and co-design, and to Professor Serina Chang, Professor Alane Suhr, Téa Wright, Seun Eisape, and Seung-Hyeong Lee for collaborations on various topics on language modeling.

Outside of research, my doctoral program was enriched by countless friends I met during my time at Berkeley: Woosuk Kwon, Sunjin Choi, Sam Son, Hansung Kim, Junsun Choi, Hongsuk Choi, Kunmo Kim, and Dayeol Lee. Finally, I would like to thank Karmen and my family for their love and support, without which this work would not have been possible.

Chapter 1

Introduction

Large language models (LLMs) have demonstrated remarkable general-purpose capabilities across diverse domains, from reasoning and code generation to open-ended dialogue. Their ability to process and produce human-like communication has led to a rapidly growing interest in using LLMs not only as computational tools but also as *proxies for human users* [14, 4, 279, 181, 170, 116, 205]. This dissertation is concerned with the methodology of using LLMs in this role, the problem of *user simulation*. It argues that user simulation deserves a dedicated treatment, distinct in both its modeling objectives and its evaluation criteria from the increasingly dominant problem of training helpful assistant agents.

LLMs are deployed as simulated users in two complementary regimes that motivate the work in this dissertation:

(i) Simulated study participants. In computational social science, behavioral economics, and political psychology, LLMs are used as scalable, low-cost stand-ins for human participants. Researchers prompt LLMs to answer survey questions [205, 14, 170, 227], complete personality inventories [211, 111, 93], render moral judgments [220, 3, 61], simulate full societies of agents [184, 181, 185], and play behavioral economic games [5, 94, 250, 69]. The promise is profound: the pool of voices accessible through web-scale pretrained LLMs is enormous and includes hard-to-reach participants (ill, incarcerated, non-cooperative) in essentially unlimited contexts. The "participants" also remain available indefinitely, are highly scrutable (one can ask them *why* they answered as they did), and impose no direct burden on human participants.

(ii) Simulated interlocutors for AI agents. A parallel, rapidly growing line of work uses LLMs as simulated users for the development and evaluation of LLM-based assistants and agents. Rather than recruiting human participants for every iteration, a simulated user "interacts" with a model in multi-turn dialogue, providing automatic feedback at scale. The practice of building classical task-oriented dialogue systems through user simulation [132, 78, 207, 127] has continued to prove effective even as the underlying technology has moved to LLMs, for problems including cooperative planning [145], cooperative task completion [135, 133], end-to-end evaluation of dialogue

agents [120, 210, 213, 260, 273], and reinforcement learning pipelines where simulator fidelity directly determines the quality of the reward signal [249, 2, 74].

These two regimes share a common technical core. In both cases, we want a language model conditioned on a context c, such as a survey item, a dialogue history, a game prompt; the conditioned LM then produces continuations that approximate what a human user would produce in the same context. In the first regime, c is typically static and the relevant fidelity criterion is at the population level (how does the distribution of model responses compare to the distribution of human responses?). In the second, c unfolds turn by turn, and the relevant fidelity criterion includes pragmatic and behavioral dynamics (does the simulated user resist, clarify, and disengage as humans do?). What unifies them is that in both cases we are asking the language model to be a faithful estimator of human-authored text in context—not a helpful assistant, not a polite interlocutor, not a normatively aligned agent.

The promise of either regime, however, is bottlenecked by a methodological problem that motivates this dissertation. The dominant practice in both regimes is to take an instruction-tuned "chat" model and prompt it with a persona description or role-playing instruction [173, 216, 239, 232, 181, 183]. A growing body of evidence shows that this default is the wrong one. Instruction-tuned LLMs have been repeatedly shown to exhibit collapsed lexical and semantic diversity [124, 179, 263, 112, 153], sycophantic alignment to apparent user preferences [46, 189, 217], persona drift over multi-turn interactions [134, 139], and systematic ideological skew, particularly toward U.S. liberal positions [205, 204, 77]. These pathologies are commonly reported as failures of "language models" in general [10, 273, 46, 134, 154, 117, 220]. The position of this dissertation is that they are not. They are failures of the wrong default—of using assistant-aligned models for a problem that is fundamentally about modeling the conditional distribution of human user text.

1.1 User Simulation is Language Modeling, Not Role-Playing

Chapter 2 of this dissertation develops this position in detail. We distinguish two classes of LLMs as substrates for user simulation:

Human–Language Mixture Models (HLMMs). Pretrained base models, trained via cross-entropy on web-scale corpora, are optimized to minimize divergence from the empirical conditional distribution of human-authored text [196, 33, 9, 172]. In the idealized limit, $p_{\theta_H}(u \mid c) \approx p^\star(u \mid c)$, where $p^\star$ is the conditional distribution of human utterances given context. Because the training corpus is itself authored by an enormous and heterogeneous population of speakers, the resulting model can be viewed as a mixture over latent author and context variables—a property we call the "mixture of voices." This is precisely the property that user simulation requires.

Assistant Language Models (ALMs). Instruction-tuned models are produced by post-training pretrained checkpoints with supervised fine-tuning on instruction-response pairs and reinforcement learning against learned reward models [178, 16, 49, 224]. These mode-seeking objectives concentrate probability mass on stylistically uniform, assistant-coded completions and, by design, suppress the diversity, off-task drift, and emotionally charged behaviors that characterize real users. The induced distribution p_{θ_A} is no longer a faithful estimate of $p^\star$ but a narrow distribution of speakers acting as helpful and safe assistants.

Chapter 2 evaluates this position empirically across seven model families and five conversational corpora: Reddit, WildChat-1M, LMSYS-Chat-1M, MultiWOZ, and DialOp. Across these corpora, base models more closely estimate the empirical distribution of human user utterances (measured by perplexity and MAUVE), better reflect the pragmatic structure of dialogue acts, and preserve lexical and semantic diversity that instruction-tuned models systematically lose. The chapter further introduces *Tandem Models*, an inference-time procedure that pairs an HLMM utterance generator with an ALM supervisor that selects (but never rewrites) candidate continuations, demonstrating that the strengths of each model class can be combined without re-collapsing diversity.

The remainder of this dissertation builds directly on this position. Chapters 3 and 4 both assume the use of pretrained base models, and demonstrate how to bind those models to coherent, controllable virtual personas as synthetic study participants. The methodology developed there is also the natural foundation for the second regime, since the same backstory-conditioned base model that simulates a Pew respondent can also serve as a synthetic interlocutor in an agent evaluation pipeline.

1.2 Why Faithful User Simulation Matters

Before describing the technical framework, it is worth elaborating on why user simulation matters in each of the two regimes outlined above, and what is at stake when the dominant practice gets it wrong.

For Behavioral Research: Ethics, Cost, Validity

Traditional human-subject studies remain indispensable for understanding real human behavior, but they face three enduring challenges that LLM-based simulation can complement.

Ethics and the Belmont Principles. Research with human participants is guided by the Belmont Report's principles of *Respect for Persons*, *Beneficence*, and *Justice* [81]: protect autonomy through informed consent, maximize benefits while minimizing harm, and distribute burdens and benefits fairly across populations.

Traditional experiments often struggle to satisfy these ideals in practice: sensitive protocols may induce discomfort or privacy risks; convenience sampling can underrepresent key groups; and scaling ethically across contentious topics is difficult. LLM-based simulations offer a complementary path. When conditioned as virtual personas, no real person's autonomy is infringed, direct harm is eliminated, and demographically balanced synthetic cohorts can be constructed by design. Crucially, however, *Justice is not automatic*: researchers must still ensure that the synthetic population is fairly constituted, by verifying the density of minority personas, checking that demographic groups are represented proportionally, and monitoring for downstream forms of imbalance that may emerge during analysis. Achieving Justice therefore requires deliberate measurement and adjustment rather than assuming that synthetic cohorts are inherently fair. In this dissertation, the Belmont principles are treated not as external constraints but as *internal design goals* for ethically grounded LLM-simulated studies.

Cost and Scalability. Human data collection is resource-intensive. Recruiting respondents through commercial platforms commonly costs between $18 and $61 per participant-hour [65], whereas equivalent simulations using LLMs can be conducted for roughly $0.02 per persona-hour, representing reductions of 90× to 300× in cost. This disparity transforms what is feasible in behavioral research. LLMs enable thousands of controlled studies across demographic or ideological conditions, large-scale replications, and longitudinal tracking that would otherwise be prohibitively expensive. LLMs thus serve as cost-effective proxies for early-stage validation and pilot testing. Rather than replacing human participants, they extend the research workflow: hypotheses can be iteratively refined through simulated experiments and subsequently confirmed through human trials.

Bias, Variance, and Validity. The final consideration concerns the bias and variance of the estimated quantities: how representative the sample is and how reliable the resulting inferences are. Because recruiting large, diverse, and geographically distributed participants is expensive, many contemporary studies rely on online recruitment platforms such as Prolific or Amazon Mechanical Turk. While these platforms have expanded access, their users tend to be younger, more educated, and politically liberal, so many studies lack true population representativeness and suffer from high sampling bias. When researchers cannot afford to recruit sufficiently large samples, statistical variance increases, leading to greater estimation error and lower confidence in observed effects. The consequences are reflected in the *replication crisis* in psychology and social science, where reproducibility rates have been reported below 40% [175, 38]. Reproducibility is further affected by *contextual factors*: results can vary depending on how instructions are framed, which demographic groups happened to participate, and the temporal context in which the study was conducted. Even minor shifts in these elements can yield meaningfully different behavioral outcomes, complicating cross-study comparison.

LLM-based virtual personas offer a partial solution. By enabling the systematic generation of demographically balanced samples under controlled conditions, they allow researchers to explicitly examine and manipulate sources of bias while reducing variance through repeated sampling and consistent conditioning. The bias–variance tradeoff becomes a controllable aspect of study design. Beyond this, LLM simulation also opens forms of inquiry that are not feasible with human participants alone, such as *counterfactual replication*, in which the same study is run with the year, framing, or recruitment population independently varied to decompose contextual factors that human studies necessarily bundle together [169].

For Conversational AI: Synthetic Interlocutors as Critical Path

The second regime, LLMs as simulated interlocutors, has a parallel set of stakes. Synthetic users increasingly sit on the critical path of training and evaluation for conversational agents. Reinforcement learning pipelines optimize agent policies against simulated user feedback; benchmark evaluations score agents against multi-turn dialogues with simulated users; and user models serve as the training distribution for instruction-following itself.

When the simulated user is unrealistically polite, overly compliant, or distributionally collapsed, this distorts conversational trajectories the assistant agent model experiences during training. Recent evidence makes the consequences concrete. Gandhi, Bhatia, and Goodman [74] show that optimizing for judge-based rewards from instruction-tuned simulators leads to reward hacking that actually *decreases* the log-likelihood of ground-truth human responses. Zhou et al. [273] document a substantial gap on τ-Bench: instruction-tuned simulators are too cooperative, stylistically uniform, and insensitive to real-user frustration and ambiguity, and crucially, higher base-model capability does not close this gap. Seshadri et al. [213] show that off-the-shelf simulators are systematically miscalibrated about agent success rates relative to real users. Lu et al. [153] locate the problem mechanistically in an "assistant axis" within model activations that persists even under explicit persona prompting. Across these results, a common pattern emerges: the diversity collapse and stylistic uniformity that are valuable for an assistant become liabilities when the same model is asked to simulate a user. Faithful user simulation is therefore not a downstream nicety but a prerequisite for trustworthy agent evaluation and training.

Here, too, the position of this dissertation applies: the way out is not engineering fixes on top of instruction-tuned models, but recentering user simulation on pretrained base models that retain the breadth required of a faithful user model.

1.3 Binding LLMs to Virtual Personas via Backstories

The position above settles the question of substrate but not the question of *binding*: how does one steer a pretrained LLM's enormous mixture of voices toward a particular human persona, consistently across many turns and many questions? This dissertation proposes a unified answer: synthetic narrative *backstories* generated by base models themselves.

The key idea is that an individual's coherent life narrative, what psychologists call narrative identity [162, 35, 158, 161], can serve as a strong conditioning context that shapes attitudes, perceptions, and decision-making. A backstory is a first-person life narrative that explicitly and implicitly encodes diverse details about its author: age, gender, region, education, career, relationships, values, beliefs, formative experiences. Lengthy and coherent backstories thus narrowly constrain the latent persona without requiring the experimenter to specify every demographic and psychological variable in advance.

While the backstory itself provides a rich and psychologically meaningful context, it must still be *bound* to the underlying LLM in a way that the model internalizes and adopts it as its operating persona. Crucially, pretrained base models support this form of binding naturally: the backstory can be used directly as a prefix, and the model conditions on it as part of its generative context, locating a coherent region of its mixture-of-voices distribution. In contrast, instruction-tuned or chat models are typically bound through explicit instructions or system prompts, and do not reliably treat narrative context as persona-defining; their alignment objectives override such cues with safety, politeness, or normative interpretations of how the persona "ought to" behave [134, 117]. Effective backstory-based simulation therefore depends on the substrate choice argued for in Chapter 2: it is the conjunction of pretrained base models and rich narrative prefixes that produces coherent virtual personas.

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

1.4 Contributions and Roadmap

This dissertation makes the following contributions:

- It articulates and empirically validates the position that user simulation is a language-modeling problem that should use pretrained base models (HLMMs), not instruction-tuned assistant models (ALMs), and that many reported limitations of "LLMs" for user simulation are more accurately understood as limitations of the wrong default.
- It introduces an inference-time *Tandem* architecture that pairs an HLMM utterance generator with an ALM selector, demonstrating near-Turing-indistinguishable simulation while preserving generative diversity, and showing that the strengths of base and instruction-tuned models can be combined without sacrificing either.
- It introduces a unifying methodology for binding pretrained LLMs to coherent virtual personas via synthetic narrative backstories, and a demographic-matching procedure that constructs simulated populations representative of target human studies.

- It demonstrates, across three large-scale Pew Research Center surveys, that backstory-conditioned base models reproduce population-level opinion distributions with greater representativeness, internal consistency, and demographic-subgroup fidelity than prior persona-conditioning baselines.
- It introduces a method for *deep* persona binding via long, internally consistent multi-turn interview backstories, and shows that the resulting personas reproduce partisan asymmetries in moral judgment, expectations of democratic backsliding, and exaggerated meta-perceptions observed in real human partisans.
- It extends backstory-conditioned LLM simulation from attitudinal prediction to *action prediction* in partisan social-dilemma games, introduces *Temporal Grounding* and *Consistency Filtering* as supplementary prompting strategies for situated behavioral simulation, and shows that the resulting personas faithfully replicate human partisan favoritism in published Dictator and Trust Game studies spanning more than a decade of replications.
- It demonstrates that LLM simulation enables a form of inquiry not feasible with human participants alone: factorial counterfactual decomposition of human-study reproducibility into the contributions of year, framing, and recruitment-population effects. It then applies this to a $2 \times 2 \times 2$ ANOVA decomposition of partisan effects across published Dictator and Trust Game replications.
- It positions user simulation as the critical methodological path for training and evaluating thought-partner AI assistants in wicked design problems, articulates three structural challenges that distinguish AI assistance in design from instruction-following (epistemic uncertainty over latent goals, distribution shift across interaction trajectories, and skill-level heterogeneity), and identifies digital hardware design as a flagship application where the methodology of this dissertation is most needed and least developed.

The remainder of the dissertation is organized as follows.

Chapter 2 formalizes the HLMM/ALM distinction, presents the empirical case across five conversational corpora that pretrained base models are better user simulators than their instruction-tuned counterparts, and introduces the Tandem architecture.

Chapter 3 introduces the *Anthology* framework for backstory-conditioned virtual personas. It details the generation, demographic estimation, and bipartite-matching pipeline, and demonstrates substantial improvements over prior persona-conditioning methods on three Pew American Trends Panel surveys.

Chapter 4 extends the Anthology framework to model social identity and group perception. It introduces a method for generating long, internally consistent multi-turn interview backstories with LLM-as-a-critic filtering, and shows that the resulting deep persona binding reproduces partisan asymmetries in moral judgment and meta-perception that shallower conditioning methods fail to capture.

Chapter 5 extends the methodology and then turns to applications. It first extends the application of backstory conditioning from attitudinal to behavioral simulation through action prediction in partisan social-dilemma games, and introduces a counterfactual decomposition of human-study reproducibility into year, framing, and population effects. It then frames *design problems* as the natural application surface for the methodology developed here, articulates three structural challenges that distinguish AI assistance in design from instruction-following, and identifies digital hardware design as a flagship application target.

Chapter 6 synthesizes these contributions, discusses limitations and responsible-use considerations, and outlines remaining open directions for the broader user-simulation research program.

Chapter 2

Role Playing is Not Language Modeling

2.1 Introduction: User Simulation Should Use Pretrained, Base LLMs

Large language models (LLMs) are increasingly used as proxies for human users. For example, in computational social science, they serve as synthetic survey respondents and study participants, offering scalable alternatives to costly human panels [14, 4, 21, 279]. A parallel and rapidly growing line of work uses LLMs as *simulated users* for training and evaluating LLMs: rather than recruiting human participants for every iteration, a simulated user "interacts" with a model in multi-turn dialogue, providing automatic feedback at scale. The practice of building classical task-oriented dialogue systems through user simulations [132, 78, 207] has continued to prove effective even as the underlying technology has moved to LLMs, for problems including cooperative planning [145], cooperative task completion [135, 133], end-to-end evaluation of dialogue agents [120, 210, 213, 260, 273], and in reinforcement learning pipelines where simulator fidelity determines the quality of the reward signal [249, 2, 74].

Nearly all work that takes advantage of modern LLMs uses instruction-tuned (IT) "chat" models, prompted with persona descriptions or role-playing instructions to mimic user behavior [173, 216, 239, 183, 232, 181]. While such models have an easy-to-use interface (i.e., via natural language instructions) and typically produce plausible responses, instructing assistant-aligned LLMs for simulating diverse human users has inherent structural problems. Several reports show how instruction-tuning collapses generative diversity: reward maximization concentrates probability mass on stylistically uniform, assistant-coded completions [124, 179, 263, 79, 112, 241]. For user simulation, we want LMs to reflect the *broad* distribution of natural human utterances, including ambiguous, resistant, and off-task turns, which IT models do not.

Critically, *pretrained* base models do not share the limitations born of subsequent post-training. Trained via cross-entropy on human text, these models approximate the empirical distribution of human authorship and lexical, stylistic, and pragmatic variations present in web-scale data [196, 33, 9, 172]. Yet failures of IT simulators have been reported as failures of "language models" in general, with respect to limited generative diversity, sycophantic behavior, and discrepancies in naturalness of generated continuations compared to humans [10, 273, 46, 134, 154, 173, 155, 102, 55, 117, 220]. These studies narrowly consider IT models in producing the negative results, which presents a misleading picture of what LLM-based simulation can achieve with models prior to instruction-tuning.

The position of this paper is that *instructing* an IT model to "behave as users" yields simulated interactions reflecting the assistant LLM's "interpretation" of how a user *ought* to behave, rather than a reflection of actual user behavior as found in the pretraining data. This shift motivates a necessary change in the default model choice. If the goal is to simulate the next turn of a human interlocutor conditioned on some dialogue history, the appropriate starting point is the model class trained to estimate human text in context (pretrained LMs), not the model class subsequently optimized to occupy the social role of a helpful assistant (instruction-tuned LMs).

Specifically, our position is described by the following claims:

(C0) **Pretrained LMs better estimate the *conditional distribution of human user utterances* than the subsequent model after instruction-tuning.** If the task is to generate a plausible next user turn, the relevant target is the empirical distribution of what humans write in context.

(C0) **Pretrained LMs better reflect the *pragmatic behaviors* of human interlocutors.** Real users are not merely stylistically diverse; they also differ yet converge at the corpus-level in what communicative acts they perform and when they perform them.

(C0) **Pretrained LMs better capture the *diversity* of language produced by human speakers.** User simulation requires coverage; a simulator that repeatedly produces the same polite, explicit, assistant-adjacent utterance may appear reasonable under isolated inspection, but it is a poor model of a population of users.

The rest of the paper develops and tests this position empirically. Section 2.3 formalizes the distinction between pretrained LMs as *Human-Language Mixture Models (HLMMs)* and instruction-tuned models as assistant models. Sections 2.4, 2.5, and 2.6 then evaluate the three requirements of faithful user simulation: distributional fit to human utterances, reflecting the pragmatic dialogue structure exhibited by human interlocutors, and lexical and semantic diversity. Across these analyses, we substantiate that LLMs optimized to be good *assistants* are not the models best suited to simulate *human users*.

2.2 Background & Related Work

LLM-based User Simulation

User simulation has a long history in dialogue research, ranging from statistical and agenda-based simulators for spoken dialogue systems [132, 78, 206, 207] to data-driven neural approaches [127, 144]. In the LLM era, the dominant paradigm has shifted toward prompting instruction-tuned (IT) models with persona descriptions or role-play instructions [173, 216, 239, 232, 137, 265, 157, 235]. A growing body of evidence, however, suggests that this approach inherits pathologies of assistant-aligned models that are fundamentally at odds with the demands of user simulation: reduced lexical and semantic diversity [124, 179, 263, 112, 153], sycophantic alignment to apparent user preferences [46, 189, 217], and persona drift over multi-turn interactions [134, 139]. Recent work has begun to surface this gap: Lyman et al. [154] show that base models retain greater behavioral fidelity for social-science applications, and Naous et al. [172] argue that "better assistants yield worse simulators." Yet much current effort still goes into engineering fixes [249, 2, 278, 139, 264] for problems that could be sidestepped by a different choice of model type [170, 275, 116]. Our position is that these challenges are downstream of a misidentified default: the field has tacitly equated "LLM" with "instruction-tuned LLM," and recentering user simulation on base models recasts these pathologies not as open research problems but as symptoms of the wrong starting point.

Evaluating User Simulators

The evaluation of user simulators attempts to quantify the discrepancy between simulated and real-world interactions [273]. Naous et al. [172] assess fine-tuned LLM user simulators using metrics such as intent coverage via n-gram overlap, first-turn diversity, and likelihood scores from an AI-detector Pangram [67] to estimate naturalness. Wu et al. [249] rely on an automated judge to score alignment with ground-truth human responses on their comprehensive benchmark, HUMANUAL. However, recent work demonstrates that optimizing for judge-based rewards leads to reward hacking, actually decreasing the log-likelihood of ground-truth human responses [74]. Instead of automated judges, some recent work favors statistical measures of human-likeness such as the composite User-Sim Index which calculates the Sørensen–Dice coefficient to measure alignment across several dimensions [273], or expected calibration error on τ-Bench [213]. A recurring theme in these evaluations is a prevalent diversity collapse and "artificial hivemind" effect [112] in which instruction-tuned models consistently use the same assistant-like language at the expense of diversity [124, 179]. Lu et al. [153] locate this collapse in an "assistant axis" within model activations that persists even with persona prompting. Yun et al. [263] show that the diversity collapse is reinforced by structured output templates that limit the output space. Base models, however, are not subject to these structural bounds. As Lin et al. [142] demonstrate, base models can achieve the same level of performance as instruction-tuned models through prompting with stylistic examples while maintaining the diversity that modern assistants have lost. We thus

argue that the current reliance on instruction-tuned models is a misinformed default, and user simulation requires the breadth found in pretrained base models.

2.3 Pretrained and Instruction-Tuned LLMs for User Simulation

In this section, we formalize the problem of simulating human users in dialogue, where a language model is provided with a dialogue context and generates a continuation of the target human interlocutor's next turn. We explain how such generation is in itself a reflection of *language modeling* capabilities of LLMs that are learned through pretraining on a large corpus of human text, and how additional instruction-tuning and reinforcement learning (RL) increase model likelihoods of generations that match desired objectives as assistant agents, but at the expense of shifting the distribution of generations from the original conditional distribution learned during pretraining.

Two Classes of LLMs

Pretrained LLMs are Human-Language Mixture Models. At the utterance level, a language model defines a conditional distribution $p_\theta(u \mid c)$ over next-turn utterances u given dialogue context c. We denote the corresponding true distribution of human utterances as $p^\star(u \mid c)$. Following theories of language production and pragmatic reasoning [32, 131, 72, 80], we can view this target distribution as a *mixture* over implicit authorship z, where the latent variable summarizes the state of the interlocutor, such as identity/persona, beliefs, desires, and communicative intentions. Under this modeling lens,

$$p^\star(u \mid c) = \int p^\star(u \mid c, z)\, p^\star(z \mid c)\, dz.$$

During pretraining, language models are trained under the cross-entropy loss over large-scale corpora, the vast majority of which are human-authored text. Training with this objective minimizes divergence of the model distribution p_θ from the empirical distribution over the training corpus, $\hat{p}^\star$:

$$\arg\min_\theta \; \mathbb{E}_{(u,c)\sim\hat{p}^\star}\big[-\log p_\theta(u \mid c)\big].$$

Minimizing this objective is equivalent to minimizing $\mathrm{KL}(\hat{p}^\star \| p_\theta)$. Thus, in the idealized limit of scaled representative data and training convergence, the pretrained model distribution approaches the empirical conditional distribution of human-authored continuations. Assuming the corpus itself is representative, then $p_\theta(u \mid c) \approx \hat{p}^\star(u \mid c) \approx p^\star(u \mid c)$.

This view of cross-entropy training of LMs motivates our formulation of pretrained LMs as Human–Language Mixture Models (HLMMs), which we denote as $p_{\theta_{\mathrm{H}}}$. Note that we do not claim that Transformer-based auto-regressive models are *explicitly* modeling the posterior $p_{\theta_{\mathrm{H}}}(z \mid c)$, nor that pretraining replicates the generative process by which humans

produce language. Rather, following Andreas [9], we adopt the view that, as an *emergent* consequence of next-token prediction over a corpus produced by many distinct human authors, a pretrained LM learns internal representations that behave *as if* they track the latent state of the speaker. In turn, this is precisely the property that user simulation requires. An HLMM provides *coverage* over the heterogeneous space of plausible human speaker states and approximates lexical, stylistic, and pragmatic variation present in natural human dialogue.

Instruction fine-tuning yields LLMs as Assistant Language Models (ALMs). Instruction fine-tuning, on the other hand, transforms pretrained LLMs into helpful conversational assistants. Modern training pipelines typically apply supervised instruction fine-tuning (SFT) on curated instruction-response pairs, followed by reinforcement learning from human or AI feedback (RLHF/RLAIF) against a learned reward model [178, 16, 49, 224]. More recent pipelines additionally incorporate reasoning-oriented RL objectives that reward chain-of-thought trajectories leading to correct or preferred final answers [84, 105].

While effective at aligning models with assistant-style desiderata such as helpfulness, harmlessness, and instruction-following, these post-training stages fundamentally alter the distribution learned during pretraining. RL post-training optimizes models via a mode-seeking objective: reward maximization moves probability mass toward high-reward completions, collapsing the broad pretraining distribution [124, 179, 79]. The implications for user simulation are direct: the induced distribution, which we denote as $p_{\theta_{\mathrm{A}}}$, is no longer a faithful estimation of the mixture of human authorship, but a limited distribution representing speakers aligned as helpful and safe assistants [145, 263].

Fine-tuning and Inference-Time Methods for HLMMs

There are several directions we witness in the literature that aim to improve the use of pretrained LMs without introducing the artifacts and limitations of instruction-tuning. Recent works propose supervised fine-tuning a pretrained model directly on human conversational data, using objectives that remain anchored to the empirical distribution of human text [172]. Other recent methods are inference-time or prompt-based techniques for extending the capabilities of pretrained models without weight updates. *Backstory conditioning*, for instance, prepends a rich narrative self-description of the simulated individual as model context, without requiring the model to interpret a prescriptive demographic profile as an instruction to enact as the user [170, 116].

These observations highlight a simple inference-time procedure, which we refer to as *Tandem Modeling*. This idea pairs a pretrained model as an utterance generator with an instruction-tuned model as a supervisor of the generated outputs from the utterance generator. First sample a pool of candidate utterances from the pretrained LLM; the supervising assistant model then selects a sampled candidate based on its selection/rejection criteria:

$$\mathcal{U}(c) = \left\{u^{(1)}, \ldots, u^{(N)}\right\}, \quad u^{(i)} \sim p_{\theta_{\mathrm{H}}}(u \mid c),$$
$$\hat{u} = \arg\max_{u \in \mathcal{U}(c)} s_{\mathrm{ALM}}(u, c).$$

Table 2.1: **Conversational corpora used in our case study.** For all corpora, we use similar next-utterance prediction setups with LMs and sample 8 generations per context. For Multiwoz and DialOp, we take 128 and 117 dialogues, respectively, and generate utterances of all human-user turns; for all other datasets, we take the final user turn and generate model utterances in place of the human interlocutor.

Corpus	Dialogue type	Sampled Contexts	Total
Reddit (ConvoKit) [42]	Open-domain; Human–Human	1,024	8,192
WildChat-1M [268]	Human-LLM Chat; Human–AI	1,024	8,192
LMSYS-Chat-1M [272]	Human-LLM Chat; Human–AI	1,024	8,192
MultiWOZ 2.1 [37]	Task-Oriented; Human–Human	807	6,456
DialOp [145]	Task-Oriented; Human–Human	306	2,448

The key design constraint is that the IT model supervisor only selects or rejects, without altering or rewriting the final utterance: a simple approach to retain the generative diversity of pretrained models, while allowing the instruction-tuned model to bring its stronger discriminative capability to bear on which candidate best fits the local conversational context.

Experiment Setup

In the following sections, we provide experimental results supporting our claims on the differences between pretrained and instruction-tuned LMs in simulating human interlocutors. To ensure that our findings are general across various domains of conversational language generation, we employ five conversational corpora spanning from open-domain dialogue to human-AI interactions. For each corpus we extract next-utterance prediction examples from multi-turn conversations (minimum one turn from each interlocutor) and generate 8 continuations per context for each model. The total number of dialogue contexts we sample from the dataset are described in Table 2.1. Further details on the procedure and details of the text corpora are described in Section A.1.

We evaluate pairs of pretrained, base LMs and the checkpoint after instruction-tuning across families of open-weight LLMs: Llama-3.1-8B, Mistral-7B, Mistral-Small-24B, Qwen2.5-7B, Qwen2.5-14B, Qwen3-8B, and Qwen3-14B. Comparing matched base/instruct pairs isolates the effect of instruction-tuning on user modeling. Pretrained base models are prompted with the natural formatting of the dialogue so far; for instruction-tuned models, we provide the instructions "You are simulating a user in the following dialogue" and "Continue the conversation as the user" along with the string-formatted dialogue context. For detailed examples and prompt formats, see Section A.2.

In the following sections, we mainly focus on model variants that are > 10 billion parameters in size, and present full results in Appendix A.5.

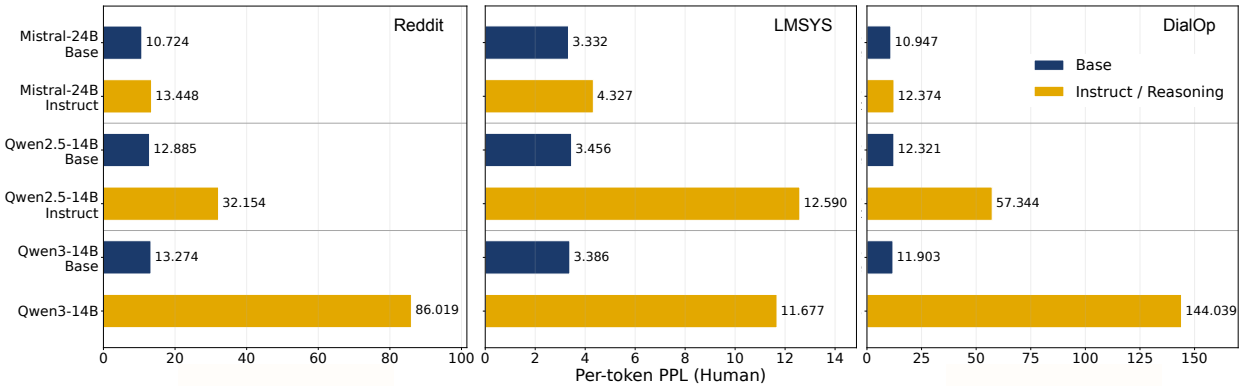

Figure 2.1: **Per-token perplexity of human utterances.** For each model, we report perplexity of human utterances given dialogue context; perplexity refers to the exponentiated average per-token negative log likelihood (NLL). Perplexity measures how surprising a piece of text is to a model: lower values mean the text matches what the model expects to see. See Appendix A.5 for full results.

2.4 Do LMs Estimate the Distribution of Human Utterances?

In this section, we provide supporting analyses around our claim C1 that pretrained LLMs are in fact models trained to estimate the (conditional) distribution of human language generation in a given dialogue context. Conversely, we present evidence on how dominant post-training stages (RLHF/RLAIF and reasoning RL fine-tuning) alter model distributions from that learned during pretraining.

Pretrained vs. Instruction-tuned LLMs through the Lens of Perplexity

A direct measure of how well a language model's induced conditional $p_\theta(u_t \mid c)$ aligns with the empirical distribution of human continuations $\hat{p}^\star(u_t \mid c)$ is its per-token perplexity on human dialogue text. Under the HLMM view (Section 2.3), a faithful estimator of $p^\star$ should assign comparable likelihood to a human continuation and to its own samples drawn from p_θ: both are draws from the same distribution the model is meant to capture.

Table 2.2: **Comparison of empirical estimates of model self-entropy and negative log-likelihood (NLL) of human utterances.** All units are per-token, thus nats/token.

Data	Model	H_{self}	$\text{NLL}_{\text{human}}$	Δ
Reddit	M24B-Base	2.8470	2.3725	0.4745
	M24B-Inst.	2.3455	2.5988	-0.2533
	Q3/14B-Base	3.0530	2.5858	0.4672
	Q3/14B (RL)	1.1912	4.4546	-3.2634
LMSYS	M24B-Base	1.9945	1.2035	0.7910
	M24B-Inst.	1.6038	1.4648	0.1390
	Q3/14B-Base	1.8273	1.2196	0.6077
	Q3/14B (RL)	0.7068	2.4576	-1.7508
DialOp	Mistral-Base	2.4861	2.3931	0.0930
	Mistral-Inst.	2.2998	2.5156	-0.2158
	Q3/14B-Base	2.3327	2.4768	-0.1441
	Q3/14B (RL)	1.5146	4.9701	-3.4555

Evidence for `C1`. If a language model successfully estimates the conditional distribution of human language given dialogue context, it should not find the human continuations from the sampled corpus idiosyncratic relative to its own.
As shown in Figure 2.1 and Table 2.2, instruction-tuned models assign very low likelihood to their own outputs but markedly higher perplexity to human text; pretrained models, on the other hand, show near-parity or reversed asymmetry—lower perplexity on human text than on their own outputs. Comparing the same model before and after the post-training stages, we not only revisit the known reports of mode-collapse from instruction-tuning in prior work, but also draw noteworthy findings on how pretrained base models assign high likelihoods of human-written continuations.

Comparing the Human and Model-Generated Utterance Distributions

To complement per-token perplexity (a local, token-level probe), we then test global distributional similarity between model-generated and human-produced utterances using MAUVE [193].

The MAUVE **metric.** Let $\widehat{P}_{\mathcal{D}}$ and $\widehat{Q}_{\theta,\mathcal{D}}$ denote the empirical distributions over human-written and model-generated utterances, encoded with a fixed text encoder $\phi(\cdot)$ and discretized into K bins via k-means. For $\lambda \in (0,1)$, the mixture $R_\lambda = \lambda\widehat{P}_{\mathcal{D}} + (1-\lambda)\,\widehat{Q}_{\theta,\mathcal{D}}$ admits two Kullback–Leibler divergences: $\text{KL}(\widehat{P}_{\mathcal{D}} \,\|\, R_\lambda)$ measures *recall failure* (regions of human text the model fails to cover), and $\text{KL}(\widehat{Q}_{\theta,\mathcal{D}} \,\|\, R_\lambda)$ measures *precision failure* (regions where the model places mass humans rarely visit). Sweeping across λ:

$$\mathcal{C}(\widehat{P}_{\mathcal{D}}, \widehat{Q}_{\theta,\mathcal{D}}) = \left\{ \left(\exp(-c\,\text{KL}(\widehat{Q}_{\theta,\mathcal{D}} \| R_\lambda)),\ \exp(-c\,\text{KL}(\widehat{P}_{\mathcal{D}} \| R_\lambda))\right) : \lambda \in (0,1) \right\}$$

with scaling constant $c > 0$, MAUVE is the area under the divergence curve:

$$\text{MAUVE}(\widehat{P}_{\mathcal{D}}, \widehat{Q}_{\theta,\mathcal{D}}) \;=\; \text{AUC}\big(\mathcal{C}(\widehat{P}_{\mathcal{D}}, \widehat{Q}_{\theta,\mathcal{D}})\big) \;\in\; [0,1]. \tag{2.1}$$

A value of 1 indicates the two samples are statistically indistinguishable; values below 1 reflect mass placed off-support in either direction. MAUVE penalizes both failure modes

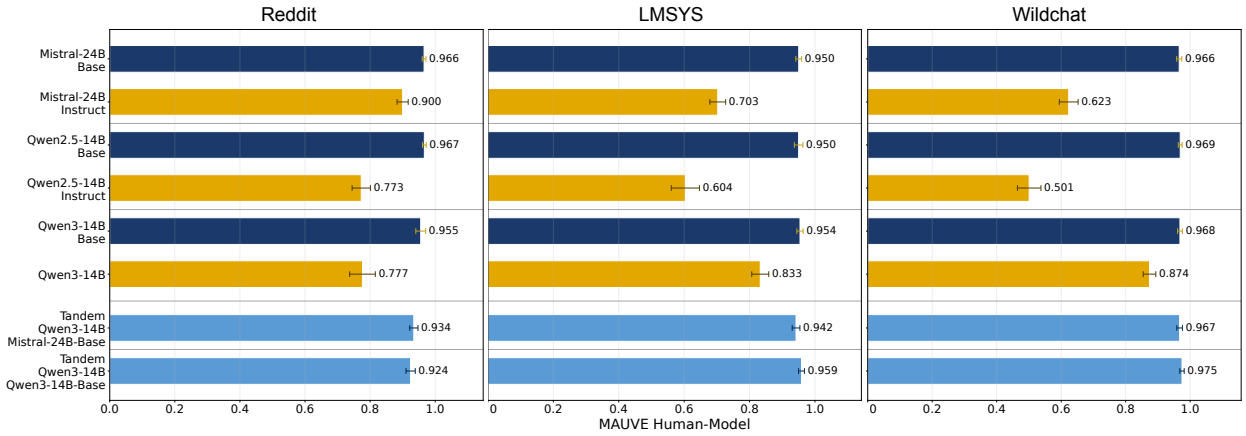

Figure 2.2: **MAUVE results.** MAUVE is a sample-based score in $[0, 1]$ that quantifies how close two text distributions are. The score is high when the set of model samples and the set of human references cover the same regions of the distribution under a text embedding model; conversely, the score is low when either distribution places mass where the other does not. A value near 1.0 indicates that the two samples are statistically indistinguishable, while lower values mean the model is sampling from a different region of utterance space than humans.

symmetrically: a model cannot score highly by collapsing onto a high-confidence subset of human utterances, because that subset would leave most of P uncovered. In our experiments, we use a state-of-the-art embedding model (`gemini-embedding-2`) as the text encoder.

Evidence for `C1`. Figure 2.2 shows MAUVE between model-generated and aligned human continuations for Reddit, LMSYS-Chat-1M, and WildChat-1M. Pretrained base models sit near the ceiling (0.95–0.97) on all three corpora, consistent with the view that they estimate the distribution of human user text directly. Instruction-tuned models drop substantially and unevenly; the Tandem configuration (see Section 2.3) shows near-identical scores as pretrained base models, showing that the distributional collapse is a property of direct generation with IT models.

2.5 Do LMs Reflect Pragmatic Patterns of Human Interlocutors?

Beyond utterance-level similarity, the realism of a user simulator in dialogue depends on its pragmatic consistency over multiple turns. This section provides analyses surrounding our claim `C2`: Pretrained LMs better reflect the pragmatic behaviors of human interlocutors.

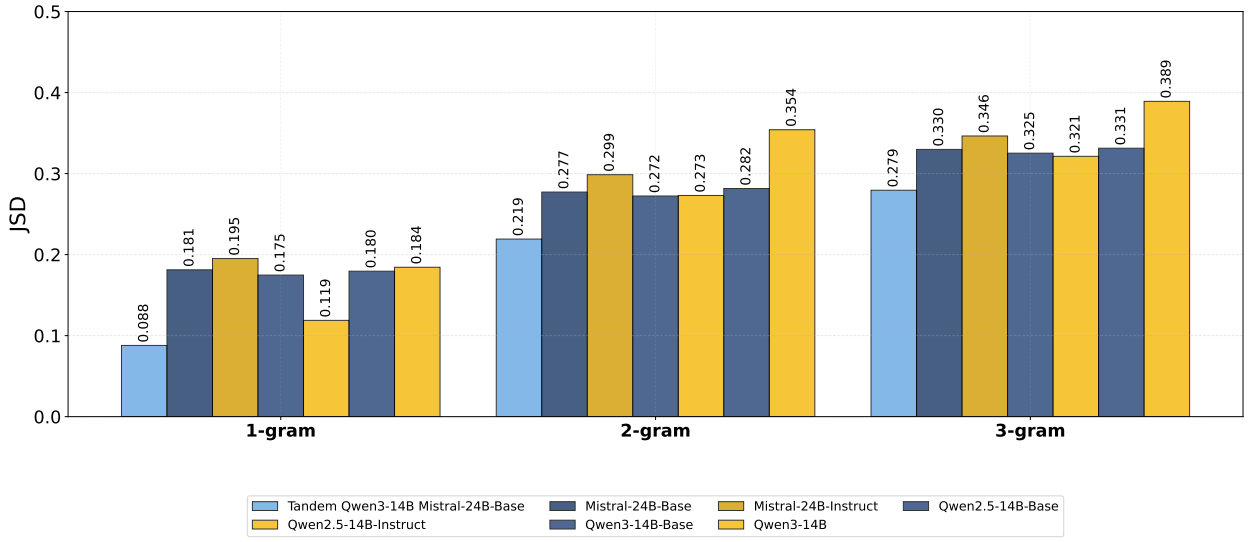

Figure 2.3: **Jensen–Shannon divergence (JSD) between human and model dialogue-act N-gram distributions on MultiWOZ-2.1.** For each $N \in \{1, \ldots, 3\}$, we form the empirical distribution of speaker-marked dialogue-act N-grams from the aligned human reference (P_N) and from the model under evaluation (Q_N), then plot JSD(P_N, Q_N). JSD is bounded in $[0, 1]$; *lower is better*, with 0 corresponding to identical distributions. Higher N probes longer pragmatic context—a model that matches marginal act frequencies (low JSD at N=1) can still fail to reproduce the *ordering* of acts across turns (high JSD at $N \geq 3$).

Analysis on Distributions of Conversational Intents. Dialogue acts represent the communicative function of turns and their sequential organization [226]. To assess pragmatic competence, we build on prior evaluations of dialogue acts and the pragmatic competence of LLMs [155, 214, 17, 215, 262]. We compare model and human dialogue-act sequences on MultiWOZ-2.1 [37], labeling generated user utterances with ConvLab BERTNLU [276] while using the aligned human turn annotations for the reference side. Each turn is converted to a speaker-marked intent token. Full details regarding labeling mechanics, intent inventory, and metric definitions are provided in Appendix A.4.

We then compare human and model distributions over dialogue-act N-grams for $N = 1, 2, 3$. This sequence-level view extends utterance-level dialogue-act analyses of LLM behavior [155, 214]: a model may match marginal frequency of the dialogue act `Inform`, while still failing to reproduce when `Inform` follows a request or resolves a constraint.

Table 2.3 shows Jensen-Shannon divergence (JSD) and vocabulary overlap with $k = 100$ results for dialogue act sequences, and Figure 2.3 visualizes the JSD between human and

model distributions of dialogue acts across 1 to 3 turns. Results for models smaller than 10B parameters, as well as individual dialogue-act distributions, can be found in Appendix A.5.

Table 2.3: **Full JSD and vocabulary overlap results ($K = 100$).** B, I, and R respectively indicate Base, Instruct, and Reasoning model variants.

	JSD (↓)			V Overlap (↑)		
Model	$N = 1$	2	3	$N = 1$	2	3
Tandem	**.088**	**.219**	**.280**	0.08	**.75**	**.84**
M24B-B	.181	.277	.330	0.08	**.75**	.74
M24B-I	.195	.299	.346	0.08	.68	.67
Q2.5-14B-B	.175	.272	.325	0.08	.68	.76
Q2.5-14B-I	.119	.273	.321	0.08	.68	.75
Q3-14B-B	.180	.282	.331	0.08	.71	.77
Q3-14B-R	.184	.354	.389	0.08	.65	.67

Evidence for C2. While all models diverge more from human patterns at higher n-gram orders, the Tandem approach maintains lower divergence with a JSD of 0.280 at $N = 3$ compared to the 0.389 seen in the Qwen3-14B reasoning model. With the exception of Qwen2.5, the Instruct and reasoning models have higher divergence than their Base counterparts. The vocabulary overlap results support this; the Tandem retains the highest overlap through $N = 3$, suggesting that picking from base model candidates preserves natural conversational transitions. Ultimately, the superior performance of the Tandem and Base models over the Instruct variants provides empirical evidence that pretrained LMs better reflect the pragmatic behaviors of human interlocutors by preserving natural transitions in task-oriented dialogue.

Analysis on Distributional Preference Expression Rate. Preference expression rate refers to where in a dialogue each of the user preferences are expressed to the interlocutor. This section analyzes preference expression rate in DialOp's [145] travel-planning task. In this setting, a "user" expresses their preferences to an "assistant" in order to iteratively construct a travel itinerary. The location where a "user" expresses one of their assigned preferences is annotated using Gemini 2.5 Flash. For each utterance, our annotator receives the full list of "user" assigned preferences and a single "user" utterance from the dialogue. Our annotator is prompted to not infer any user intent and only label utterances if they are actively being expressed. The prompt for our annotator is outlined in Section A.2. To compare the preference expression rate of a model and human as a "user", we measure **(i)** the turn index $t \in 0, 1, 2, \ldots$ in which the annotated "user" utterance lies within a dialogue, and **(ii)** the relative word position $w \in [0, 100]$ of the middle word of an annotated preference, expressed as a percentage from 0 to 100 within the full sequence.

Evidence for C2. Qualitative results from the annotator's outputs show that the difference in expression rate between base and instruction-tuned models is driven by how preferences are expressed, rather than how often. Base models tend to express their preferences via paraphrases of their initial preferences, whereas instruction-tuned models make explicit use of the wording in their preferences. This behavior inflates the preference-annotation

count for instruction-tuned models versus base models. Consider the opening turn of a DialOp conversation where the "user" has been assigned a private goal that includes: *COVID conscious, outdoor seating places would be best* (among nine other items) and the assistant's first utterance is "Hi, how are you doing today?" The responses are as follows:

Human:	"Great, I'm alright! Thanks."	user
Base:	"I am doing great. Thanks."	base model
Instruct:	"I'm good! I'm in town for the day and wanted to get some recommendations for places to visit. Would be best if it's COVID friendly and places have outdoor seating."	instruction-tuned model

The distributional measurements for t and w are normalized so the distributions are comparable; however, the over-representation of preferences from the instruction-tuned model produces utterances that differ significantly from human "users." Across model families, instruction-tuned simulators receive on average 3–7× more preference annotations per utterance than either base simulators or human users (means of 2.8–6.6 vs. 0.9–1.2 vs. 0.9 per utterance, respectively).

2.6 Do LMs Capture the Diversity of Language Generated by Human Speakers?

This section provides experimental evidence for our claim `C3`: pretrained base models, viewed as Human–Language Mixture Models, capture the diversity of human speakers in a way that instruction-tuned assistant language models systematically fail to.

Metrics of Lexical and Semantic Diversity of Model Generations. We measure generation diversity at two complementary levels: lexical (Distinct-2, Self-BLEU [180], Self-ROUGE-L[143]) and semantic variability (Self-BERTScore-F1[266], Vendi[73]) of model generations conditioned on the same input context. Further details on the metrics and the experimental configurations are found in Appendix A.4.

Evidence for `C3`. Figure 2.4 summarizes utterance-level diversity results on Reddit. Across all three metrics—Self-BLEU, Self-BERTScore-F1, and Vendi score—pretrained base models generate substantially more varied continuations than their instruction-tuned or reasoning-tuned counterparts. Lower Self-BLEU and Self-BERTScore-F1 values, which indicate greater lexical and semantic diversity respectively, are consistently observed for base models across every base/instruct pair. Vendi scores, where higher values indicate a greater effective number of semantically distinct samples, further confirm that instruction-tuning leads to a marked collapse in output diversity. Notably, the Tandem configuration preserves near-base diversity while still benefiting from IT model-based candidate selection.

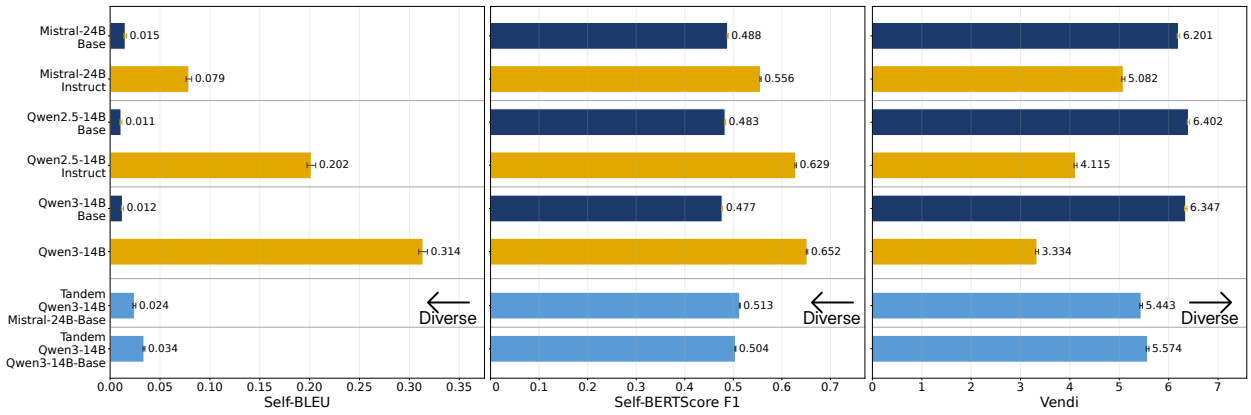

Figure 2.4: **Utterance-level diversity diagnostics on Reddit.** For each model we report three complementary measures of how varied the model-generated continuations are within a given dialogue context, averaged across contexts. **Self-BLEU** (left) computes BLEU between every pair of continuations from the same context: *lower is better*, since a low value means the continuations are lexically distinct from one another. **Self-BERT-F1** (center) is the analogous quantity in contextual-embedding space, capturing semantic rather than surface variation: again, *lower indicates greater diversity*. **Vendi score** (right) reports the effective number of semantically distinct samples, computed from the eigenspectrum of the embedding similarity kernel [73]: *higher is better*, with a value near 1 corresponding to total mode collapse.

2.7 Conclusion and Limitations

In this work, we have argued that user simulation constitutes a fundamentally distinct problem from assistant alignment or optimization, and have discussed why the dominant practice of prompting instruction-tuned (IT) models to role-play personas is the wrong default for this task. Our position calls for a conceptual reframing of how the community should view user simulation: pretrained base models, trained via cross-entropy to approximate the empirical distribution of human authorship, are Human-Language Mixture Models (HLMMs) that possess the capabilities required for user simulation. Instruction-tuned Assistant Language Models (ALMs), by contrast, are optimized under objectives that collapse this breadth and approximation of how users generate text in context.

Limitations. Our case study focused on five English-language corpora; extensions of our analysis to multilingual settings, domain-specific dialogue, and/or synchronous spoken interaction remain as open empirical questions. Broader limitations—metric convergence as evidence in lieu of a single oracle, biases inherited from pretraining corpora, and the

reduced safety constraints of base-model outputs—are discussed at the dissertation level in Chapter 6.

Broader implications. The broader implication of this work is that user modeling deserves a dedicated methodology, sensitive to the target dialogue regime and distinct from the objectives, failure modes, and evaluation criteria of assistant training. Solutions developed for assistant alignment do not automatically transfer to user simulation; in many cases, they actively work against it. We hope that the framework we have laid out here—the HLMM/ALM distinction, the three simulation desiderata C1–C3, and the example Tandem modeling idea—provides a useful conceptual scaffold for the community, and that the experimental evidence encourages more careful treatment of model selection when LLMs are deployed as proxies for human users.

Chapter 3

Virtual Personas for Language Models via an Anthology of Backstories

3.1 Introduction: the *Anthology* Framework

Large language models (LLMs) are trained from vast repositories of human-written text [234, 165, 33, 176, 168, 107]. These texts are authored by hundreds of millions of distinct authors, reflecting an enormous diversity of human traits [48, 244]. As a result, when a language model completes a prompt, the generated response implicitly encodes a mixture of voices from human authors who produced the training text from which the completion has been extrapolated. However, this natural diversity of "voices" in language models is at odds with the requirements of most applications of LLMs: a single, helpful agent voice, factual answers and a minimum of human-like emotion in responses as illustrated in Figure 3.1. Much of the later tuning of LLMs (especially RLHF in chat models such as ChatGPT and Gemini) has been shown to reduce this diversity [40]. It also clearly suppresses negative human attitudes and behaviors. Here we show that with careful design and use of "upstream" (before instruction tuning) pre-trained models, it is possible to preserve compelling and realistic virtual human voices, and apply them to practical human simulation tasks.

There is growing recent interest in the use of LLMs as human proxies for behavioral studies [14, 25, 204, 188, 185, 220, 118, 87, 108, 4, 3, 181]. While it is premature and perhaps unrealistic to argue that LLMs can replace human studies, LLMs need not replace human studies to be useful. In practice, most human studies involve a variety of compromises in scale, reach, representation, and number of questions to be answered [14]. LLMs, on the other hand, provide a low-cost, high-speed alternative that supports a nearly infinite range of querying and conditioning over target participants. The pool of LLM voices (hundreds of millions) contains many under-represented voices and hard-to-access participants (homeless, ill, disabled, incarcerated, non-cooperative) in a seemingly unlimited set of contexts. For the specific design presented here, LLMs are also highly *scrutable*. That is, participants can be queried in natural language about *why* they behaved in a certain way; the "study"

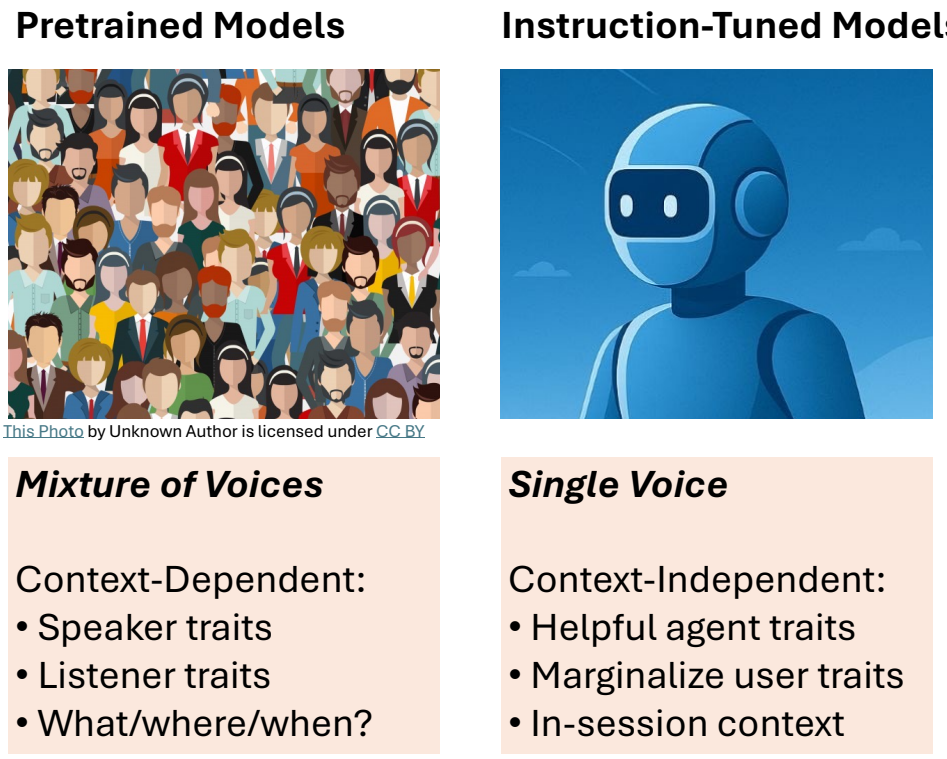

Figure 3.1: **Conceptual illustration of the differences between pretrained and fine-tuned language models.**

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

While there are evident risks from many uses of LLMs [29, 15, 91], the use of language models as an adjunct/alternative to human studies can help experimenters satisfy best practices (Belmont Principles [81]) for human studies. They minimize harms since no human participants are directly involved, and with careful design can improve representation (justice).

For language models to effectively serve as virtual participants, we must be able to steer their responses to reflect particular human users, *i.e.* condition models to reliable *virtual personas*. To this end, existing work prompts LLMs with context that explicitly spells out the demographic and personal traits of the intended persona: for example, [204, 147, 122, 100] attempt to steer LLM responses with a dialog consisting of a series of question-answer pairs about demographic indicators, a free-text biography listing all traits, and a portrayal of that persona in second-person point-of-view. While these approaches have shown modest success,

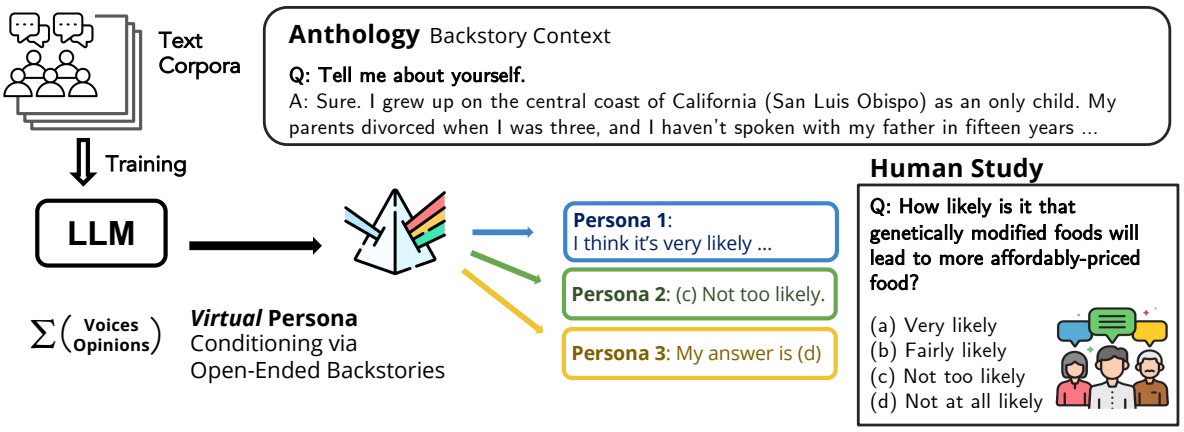

Figure 3.2: ***Anthology* overview.** This work introduces *Anthology*, a method for conditioning LLMs to representative, consistent, and diverse virtual personas. We achieve this by generating naturalistic backstories, which can be used as conditioning context, and show that *Anthology* enables improved approximation of large-scale human studies compared to existing approaches in steering LLMs to represent individual human voices.

they have been limited in (i) closely representing the responses of human counterparts, (ii) consistency, and (iii) successfully binding to diverse personas, especially those from under-represented sub-populations.

So how might we condition LLMs to virtual personas that are *representative, consistent*, and *diverse*? In this work, we investigate the use of naturalistic bodies of text describing individual life-stories, namely *backstories*, as prefixes to model prompts for persona conditioning. The intuition is that open-ended life narratives both explicitly and implicitly embody diverse details about the author, including age, gender, education level, emotion, and beliefs [12, 22, 209, 59, 225]. Lengthy backstories thus narrowly constrain the user characteristics, including latent traits such as personality or mental health that are not solicited explicitly [162, 34], and strongly condition LLMs to diverse personas.

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

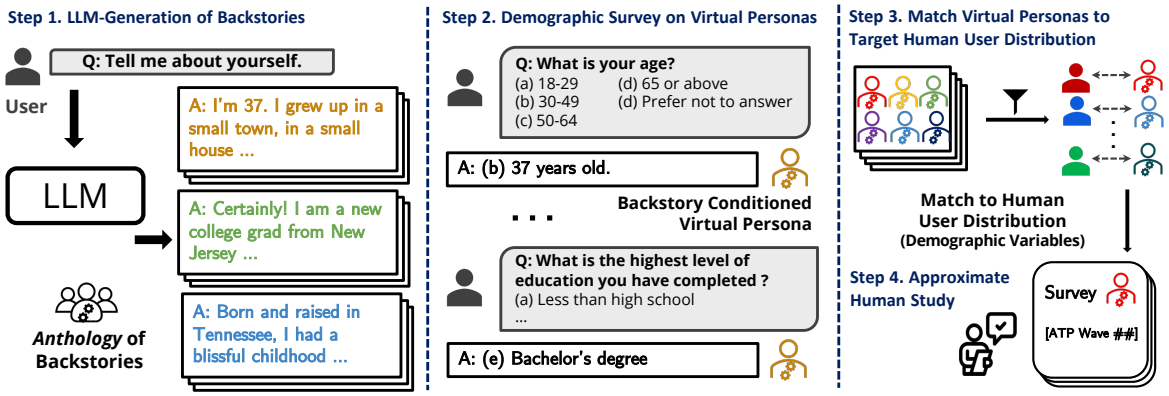

Figure 3.3: **Step-by-step process of the *Anthology* approach.** The approach operates in four stages. First, we leverage a language model to generate an anthology of backstories using an unrestrictive prompt. Next, we perform demographic surveys on each of these backstory-conditioned personas to estimate the persona demographics. Following this, we methodologically select a representative set of virtual personas that match a desired distribution of demographics, based on which we administer the survey. We find that our approach can closely approximate human results (see Section 3.4 for details).

Our contributions are summarized as follows:

- We introduce *Anthology*, which employs LLM-generated backstories to further condition LLM outputs, demonstrating that *Anthology* more accurately approximates human response distributions across three surveys covering various topics and diverse demographic sub-groups (Sections 3.4 and 3.4).
- We describe a method for matching virtual participants conditioned by backstories to target human populations. This approach significantly enhances the approximation of human response distributions (Section 3.4).
- We provide an open-source anthology of approximately 10,000 backstories for future research and applications in a broad spectrum of human behavioral studies. Additionally, we make the code for producing, processing, and administering surveys publicly available.

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

In this section, we discuss details of the proposed *Anthology* approach. We start with answering the core question: What are backstories and how might they help condition LLMs to particular personas when given as context? With an example, we examine and lay out the advantages of conditioning models with backstories in Section 3.2.

There are two practical considerations when using backstories as conditioned virtual personas for approximating human participants. In the following sections, we discuss how we address each of these implications: (i) We must acquire a substantial set of backstories that reflects a sufficient variety of human authors, since the target human study may require arbitrary demographic distribution of participants. To this end, we introduce LLM-generated backstories to efficiently generate diverse backstories (Section 3.2); and (ii) We cannot *a priori* determine the possible demographic profile of a given backstory, since demographic variables may not be explicitly mentioned in a naturalistic life narrative. Hence, we introduce methods to estimate demographics of the virtual persona conditioned by each backstory (Section 3.2) and sample subsets of backstories from anthology that match target human populations (Section 3.2).

What are Backstories?

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

Consider the example shown in Figure 3.4. We observe that the life story both *explicitly* and *implicitly* encodes information about the author, thereby providing rich insight into who the author is. For instance, the backstory provides explicit hints about the author's age

Question: Tell me about yourself.

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

Figure 3.4: **Example of an LLM-generated backstory.** The generated life story can reveal explicit details about the author, such as age, hometown, and financial background, while also implicitly reflecting the author's values, personality, and unique voice through the narrative's style and content.

("in my 60s"), hometown and/or region ("backwoods of this country"), and financial status during childhood ("grew up with very little"). But rather than being a simple listing of the aforementioned traits, the story itself embodies a natural, authentic voice of a particular human that reflects their values and personality [162, 34].

Our proposed approach is to condition language models with backstories by placing them as prefixes to the LLM [33, 234] so as to strongly condition the ensuing text completion, in the same spirit of standard prompting approaches. As we see in Figure 3.4, backstories capture a wide range of attributes about the author through high levels of detail and are naturalistic narratives that provide realism and consistency of the persona to which the LLM is conditioned.

LLM-Generated Backstories

A collection of human-written backstories could be drawn from existing sets of autobiographies or oral history collections. The challenge, however, is both in terms of scale and diversity [257, 258]. We find that, in their current standing, publicly available sources of autobiographical life narratives and oral histories are limited in the number of samples to sufficiently approximate larger human studies.

Custom human oral histories for LLM personas were recently collected in [181]. This is a promising alternative approach, but is expensive and there are privacy challenges with distribution of such stories for living persons.

Instead, we explore generation of backstories with pre-trained language models as a more scalable and cost-efficient alternative. We can also sample with finer control: e.g., tailoring demographics to a particular study and/or over-sampling minoritized groups to improve sample density and accuracy for those groups. As shown in Step 1 of Figure 3.3, we prompt LLMs with an open-ended prompt such as, "Tell me about yourself." We specifically design the prompt to be simple so that the model responses are as broad as possible (complex or academic language biases responses toward more highly-educated personas).

We believe that *generation* of plausible backstories and *accurate subsequent querying* of personas conditioned on those backstories require the same capabilities in the language model. That is, if the model can accurately generate responses conditioned on a backstory and query, it should be able to *extend a partial backstory*, and thereby iteratively generate an entire backstory. See Figure 3.4. With sampling temperature $T = 1.0$, we generate backstories that encapsulate a broad range of life experiences of diverse human users. Further details about LLM generation of backstories, including examples, are summarized in Appendix B.1.

In our experience, instruction-tuned models (i.e., most LLM agent models) are completely unsuitable for this task [134]. Whereas pre-trained models naturally represent an enormous spectrum of real users' voices, instruction-tuned models have been trained toward a single helpful, largely unemotional voice. Attempting to prompt an instruction-tuned model for a backstory leads to short, evasive and vague answers. And queries to an instruction-tuned model conditioned on a (real or synthetic) backstory lead to actions that are only positive and helpful, avoiding the (realistically human) actions that are not. Reinforcing this view,

a recent paper [117] studies the use of backstory-conditioned LLMs for qualitative (open-ended questioning) studies. The experimenters found a variety of disparities between the LLM responses and human responses. We argue that most of these disparities were due to the use of an instruction-tuned model "working as intended," but are not properties of LLMs more generally.

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

Demographic Survey on Virtual Personas

As we intend to use virtual personas to approximate human respondents in behavioral studies, it is critical that we curate an appropriate set of backstories to condition personas that represent the target human population. Each study would have a specific set of demographic variables and either accurate statistics or an estimate of the demographics of its respondents. Naturalistic backstories, despite their rich details about the individual authors, are however not guaranteed to explicitly mention all demographic variables of interest. Therefore, we emulate the process of how the demographic traits of human respondents have been collected—performing demographic surveys on virtual personas, as shown in Step 2 of Figure 3.3.

While we use the same set of demographic questions as used in the human studies, we consider that, unlike human respondents who each have a well-defined, deterministic set of traits, LLM virtual personas should be described with a *probabilistic* distribution of demographic variables. As such, we sample multiple responses for each demographic question to estimate the distribution of traits for the given virtual persona. Further details about the process and prompts used in demographic surveys are described in Appendix B.4.

Matching Target Human Populations

The remaining question is: How do we choose the right set of backstories for each survey to approximate? With the results of the demographic survey, we match virtual personas to the real human population, presented as Step 3 in Figure 3.3. In doing so, we construct a complete weighted bipartite graph defined by the tuple, $G = (H,\ V,\ E)$.

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

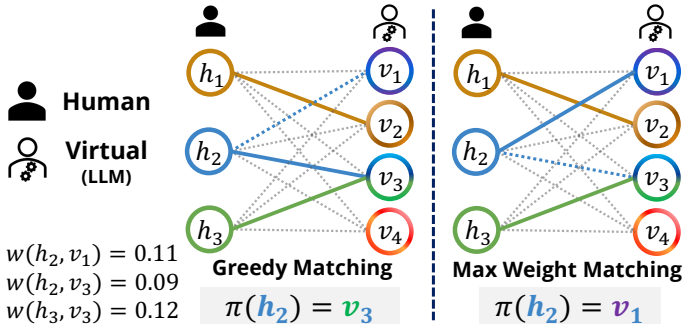

Figure 3.5: **Matching human users to virtual personas.** For greedy matching, each human user is matched to a virtual persona that has the most similar demographic traits among the virtual users. Maximum weight matching maximizes the sum of edge weights while satisfying one-to-one correspondence.

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

$$w(e_{ij}) = w(h_i,\ v_j) = \prod_{l=1}^{k} P\left(d_{jl} = t_{il}\right) \tag{3.1}$$

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

$$\pi^* = \arg\max_{\pi} \sum_{i=1}^{n} w(h_i,\ v_{\pi(i)}) \tag{3.2}$$

We explore two matching methods: (1) maximum weight matching, and (2) greedy matching. First, maximum weight matching is the method that finds the optimal π^* with the objective of Eq. 3.2, while ensuring that π establishes a one-to-one correspondence between users. We employ the Hungarian matching algorithm [128] to determine π^*. On the other hand, greedy matching seeks to maximize the same objective without requiring a one-to-one correspondence. It determines the optimal matching function such that

$$\pi^*(i) = \arg\max_{j} w(h_i,\ v_j) \tag{3.3}$$

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

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

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

Figure 3.6: **An example question (`SOCIETY_RELIG`) from ATP Wave 92 (Political Typology).** The question asks opinions about whether a given statement is good or bad for the American society.

greedy matching does not bias the estimated population mean, but it does increase the variance of survey estimates because multiple human users may be assigned to the same virtual persona.

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

3.3 Approximating Human Studies with LLM Personas

In this section, we discuss the large-scale human studies that we aim to approximate (Step 4 of Figure 3.3) using LLM virtual participants, based on varying methods of persona conditioning. We detail the overall experimental setup and define criteria for evaluation.

Human Study Data The Pew Research Center's American Trends Panel (ATP) is a nationally representative panel of randomly selected U.S. adults, designed to track public opinion and social trends over time. Each panel focuses on a particular topic, such as politics, social issues, and economic conditions. In this work, we consider ATP Waves 34, 92, and 99, a set of relatively recent surveys that cover a wide variety of topics: biomedical & food issues, political typology, and AI & human enhancement, respectively. In each wave, we select 6 to 8 questions from the original questionnaire that capture diverse facets of human opinions about the wave's topic using a Likert scale. Details on the questions selected and further information about each ATP wave are discussed in Appendix B.3.

Experiment Setup For each ATP survey considered, we format the select questions into language model prompts to administer survey approximations. Examples of such formatted

questions are shown in Figure 3.6. All questions we consider are in multiple-choice question answer formats, and we carefully preserve the wording of each question and choice options from the original survey. We ask all questions *in series*—language models are given all previous questions and their answers when answering each new question. This process replicates the mental process that human respondents would undergo during surveys. For further details on prompts used and the experimental setting, see Appendix B.2.

Language Models We consider a suite of recent LLMs including the Meta Llama3 family (`Llama-3-70B`) [165] and the sparse mixture-of-experts (MoE) models from Mistral AI (`Mixtral-8x22B`) [107, 168]. We primarily focus on models with the largest number of active parameters, which roughly correlates with model capabilities and the size of the training data corpus.

Note that we primarily consider pre-trained LLMs without fine-tuning (i.e., base models). We find instruction fine-tuned models, such as by RLHF [177] or DPO [197], to be unfit for our study as their opinions are highly skewed, in particular to certain groups (e.g., politically liberal). Prior work similarly reports notable opinion biases in fine-tuned models [204, 147, 77]. A more detailed discussion of chat models and their viability for diverse-persona conditioning is given in Section 3.4.

Virtual Persona Conditioning Methods As baseline methods for persona conditioning, we follow [204] and use (i) `Bio`, which constructs free-text biographies in a rule-based manner; and (ii) `QA`, which lists a sequence of question-answer pairs about each demographic variable.

We then compare against two variants of *Anthology*: (i) Natural refers to the use of backstories generated without any presupposed persona, as discussed in Section 3.2. In this case, we leverage either the greedy or maximum weight matching methods in Section 3.2 to select the subset to be used for each survey; (ii) Demographics-Primed instead generates backstories given a particular human user's demographics to approximate, where a language model is prompted to generate a life narrative that would reflect a person of the specified demographics (for details, see Appendix B.1). We then append descriptions of demographic traits with the generated backstories, with which we provide as context to LLMs. Examples of prompts from each conditioning method and further details can be found in Appendix B.2.

Evaluation Criteria The goal of this work is to address the research question: How do we condition LLMs to representative, consistent, and diverse personas?
Representativeness: we believe that a "representative" virtual persona should successfully approximate the *first-order* opinion tendencies of their counterpart human participants, *i.e.* respond with similar answers to individual survey questions. As questions are multiple-choice questions with ordered response options, we compare the average answer choice distributions of each question in terms of Wasserstein distance (also known as earth mover's distance) As for the representativeness across an entire set of sampled questions from a given survey, we use the average of Wasserstein distances.

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

$$d_{\text{cov}} = \|\Sigma_V - \Sigma_H\|_F \tag{3.4}$$

where $\|\cdot\|_F$ is the Frobenius norm. We additionally consider Cronbach's alpha as a measure of internal consistency independent of ground-truth human responses.
Diversity: we define the success of conditioning to diverse virtual participants by measuring the representativeness and consistency of virtual personas in approximating human respondents belonging to particular demographic sub-groups.

3.4 Experimental Results

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

Human Study Approximation

We evaluate the effectiveness of different methods for conditioning virtual personas in the context of approximating three Pew Research Center ATP surveys: Waves 34, 92, and 99, described in Section 3.3. Prior to analyzing virtual participants, we first estimate the lower bounds of each evaluation metric: the average Wasserstein distance (WD) Frobenius norm (Fro.), and the Cronbach's alpha (α), which are shown in the last row of Table 3.1. This involves repeatedly dividing the human population into two equal-sized groups at random and calculating these metrics between the sub-groups. We take averaged values from 100 iterations to represent the lower-bound estimates.

The results are summarized in Table 3.1. We consistently observe that *Anthology* outperforms other conditioning methods with respect to all metrics, for both the Llama-3-70B and the Mixtral-8x22B. Comparing two matching methods, the greedy matching method tends to show better performance on the average Wasserstein distance across all Waves. We attribute the differences in different matching methods to the one-to-one correspondence condition of maximum weight matching and the limited number of virtual users we have available. Specifically, the weights assigned to the matched virtual participants in maximum weight matching are inevitably lower than those assigned in greedy matching, as the latter relaxes the constraints on one-to-one correspondence. This discrepancy can result in a lower demographic similarity between the matched human and virtual users when compared to the counterpart from greedy matching. These results suggest that the richness of the generated backstories in our approach can elicit more nuanced responses compared to baselines.

Table 3.1: **Approximating human responses across Pew ATP surveys.** Results on approximating human responses for Pew Research Center ATP surveys Wave 34, Wave 92, and Wave 99, which were conducted in 2016, 2021, and 2021 respectively. We measure three metrics: (i) WD: the average Wasserstein distance between human subjects and virtual subjects across survey questions; (ii) Fro.: the Frobenius norm between the correlation matrices of human and virtual subjects; and (iii) α: Cronbach's alpha, which assesses the internal consistency of responses. *Anthology* (DP) refers to conditioning with demographics-primed backstories, while *Anthology* (NA) represents conditioning with naturally generated backstories (without presupposed demographics). Boldface and underlined results indicate values closest and the second closest to those of humans, respectively. These comparisons are made with the human results presented in the last row of the table.

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

Approximating Diverse Human Participants

We further evaluate *Anthology* against other baseline conditioning methods in terms of the *Diversity* criterion outlined in Section 3.3. To do this, we categorize users into subgroups based on race (White and other racial groups) and age (18-49, 50-64, and 65+ years old) with the data from ATP Survey Wave 34. We choose the Llama-3-70B model and *Anthology* using natural backstories and with greedy matching as our method and employ evaluation metrics as in Section 3.4.

As summarized in Table 3.2, *Anthology* outperforms other methods. Notably, *Anthology* achieves the lowest average Wasserstein distances and the highest Cronbach's alpha for all sub-groups. Specifically, the gap in the Wasserstein distance between *Anthology* and the second-best method is 0.029 for the 18–49 age group, showing a 14.5% difference. These results validate that *Anthology* is more effective than prior methods in approximating diverse demographic populations.

Intriguingly, for every subgroup except those aged 18-49, all methods show worse average Wasserstein distance compared to the results approximating the entire human respondents presented in Tab. 3.1. For instance, the average Wasserstein distance for *Anthology* in the ATP Wave 34 survey is 0.227, while it increases to 0.242 for the 50-64, and 0.303 for the 65+ age groups. Conversely, for the 18-49 age group, *Anthology* shows a lower average Wasserstein

Table 3.2: **Subgroup comparison of representativeness and consistency.** Target population is divided into demographic subgroups, and representativeness and consistency are measured within each subgroup. *Anthology* consistently results in lower Wasserstein distances, lower Frobenius norm, and higher Cronbach's alpha. Boldface and underlined results indicate values closest and the second closest to those of humans, respectively. These comparisons are made with the human results presented in the last row of the table.

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

Table 3.3: **Additional subgroup comparisons by education and gender.** Target population is divided into demographic sub-groups, and representativeness and consistency are measured within each sub-group. *Anthology* consistently results in lower Wasserstein distances, lower Frobenius norm, and high Cronbach's alpha. Boldface and underlined results indicate values closest and the second closest to those of humans, respectively. These comparisons are made with the human results presented in the last row of the table.

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

distance of 0.2 compared to 0.227. This finding is consistent with prior research arguing that language model responses tend to be more inclined toward younger demographics [204, 148].

We extend this analysis to two additional demographic variables: education level and gender. Education level is binarized into low (up to high school graduation) and high (some college or above), and gender follows the original survey's binary categorization for comparability with the human data. Table 3.3 shows a trend consistent with Table 3.2: *Anthology* achieves the lowest Wasserstein distance across all of these sub-groups. For the first column, the difference in the average Wasserstein distance between `QA` and *Anthology* is 0.220, a 48% discrepancy, and *Anthology* performs best on the female subgroup relative to all other baselines. These results support that *Anthology* is more effective in satisfying the *Diversity* criterion across additional demographic axes.

Table 3.4: **Effects of different matching methods.** We compare max weight matching, greedy matching, and random matching. We report two metrics: (i) the average Wasserstein distance across survey questions, and (ii) the distance between the correlation matrices of human and virtual subjects.

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

Sampling Backstories to Match Target Demographics

Next, we study the effect of matching strategies, greedy and max weight matching. In Tab. 3.4, we compare these methods with random matching, which assigns the traits of the target demographic group to randomly sampled backstories. This comparison is conducted on ATP Wave 34 using both Llama-3-70B and Mixtral-8x22B models.

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

Results on Instruction-Tuned and Smaller Models

We conduct the ATP W34 survey with additional models, including instruction-tuned variants Llama-3-70B-Instruct, Mixtral-8x22B-v0.1, GPT-3.5-0125, and a smaller base model, Llama-3-8B. None of the instruction-tuned models yield better metrics on either *Representativeness* or *Consistency* (Section 3.3). Despite their stronger performance on standard benchmarks [75, 92, 47], they do not adequately approximate human responses in this setting. We hypothesize that instruction fine-tuning, RLHF, and DPO [197, 177, 52] cause models to converge to a singular persona [187, 11, 30], rendering them unsuitable for tasks that require diverse responses—so larger fine-tuned models become *less* capable of approximating heterogeneous human responses. This is consistent with [204], who report that base

models are more steerable than fine-tuned models, and indicates that careful model selection is essential for this task [141].

Among the additional models, Llama-3-8B exhibits a higher Cronbach's alpha but a substantially higher Wasserstein distance. The elevated α reflects the model's tendency to copy previously generated responses [271, 191, 270], which yields more correlated answers across survey questions even when the marginals diverge from human data.

Table 3.5: **Approximating human responses on ATP Wave 34.** Results on approximating human responses for Pew Research Center ATP surveys Wave 34, which was conducted in 2016. We measure three metrics: (i) WD: the average Wasserstein distance between human subjects and virtual subjects across survey questions; (ii) Fro.: the Frobenius norm between the correlation matrices of human and virtual subjects; and (iii) α: Cronbach's alpha, which assesses the internal consistency of responses. *Anthology* (DP) refers to conditioning with demographics-primed backstories, while *Anthology* (NA) represents conditioning with naturally generated backstories.

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

3.5 Related Work

Generating Personas with LLMs Recent advancements in language model applications have expanded into simulating human responses for psychological, economic, and social studies [118, 4, 25, 95, 70, 14]. Specifically, the generation of personas using LLMs to respond to textual stimuli has been explored in various contexts including human-computer interaction (HCI), multi-agent systems, analyses of biases in LLMs, and personality evaluation [121, 220, 185, 204, 110, 48, 147, 248, 135, 93, 212, 97, 100, 3]. For instance, [185] and [204] develop methods to prime LLMs with crafted personas, influencing the models' outputs to simulate targeted user responses. Subsequent to the publication of the present work in EMNLP 2024, a related approach using human interview-generated backstories appeared in [181]. Additionally, [147] introduces a method where personas are generated by sampling demographic traits coupled with either congruous or incongruous political stances. Our approach, *Anthology*, advances this concept by employing dynamically generated, richly detailed backstories that include a broad spectrum of demographic and economic characteristics, enhancing the granularity and authenticity of simulated responses.

LLMs in Social Science Studies The integration of LLMs into social science research has been steadily gaining attention, as highlighted by several studies [20, 183, 62, 279, 125]. Notably, the use of LLMs to mimic human responses to survey stimuli has gained popularity, as evidenced by recent research [233, 63, 122]. A notable example is the "media diet model" by [50], which predicts consumer group responses based on their media consumption patterns. Further, studies like [247] and [279] demonstrate the potential of LLMs in zero-shot learning settings to analyze political ideologies and scale computational social science tools. Our work builds on these methodologies by using LLMs not only to generate responses but to create and manipulate backstories that reflect diverse societal segments, providing a nuanced tool for social science research and beyond.

3.6 Conclusion

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

Chapter 4

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

4.1 Introduction: the Deeper Binding of LLM Virtual Personas

Human identity is intrinsically *relational*, intertwined with how one perceives others both in relation and in contrast to oneself [54, 228, 229]. As documented across the social sciences, psychology, and philosophy, individual identities cannot be meaningfully examined outside their social contexts [44, 24, 43, 218]. Importantly, the way individuals perceive group norms to form collective social judgment and engage in inter-group interactions is central to understanding various social phenomena [41, 242, 130, 202, 240].

While recent large language models (LLMs) have been shown to simulate human behavior expressed in natural language [61, 125, 18, 170, 184, 182], prior analysis has overlooked the interplay between individual opinions and social identities. For example, prior work considers questions eliciting self-opinions of respondents [138, 205, 90], as shown in the first example in Figure 4.1: "Would *you* support using political violence?" Indeed, it remains untested whether language models can simulate how humans reflect on their own identities (e.g., self-identifying as a Democrat) to differentially shape their attitudes toward other political partisans: "Would *other Democrats* support using political violence? What about *Republicans*? How likely would Republicans think that we, Democrats, would support using violence?"

Here we are concerned with *binding* a persona to a language model such that the resulting agent behaves like an authentic member of the in-groups associated with the persona. To achieve this, we rely on a widely-used model in personality psychology: McAdams' theory

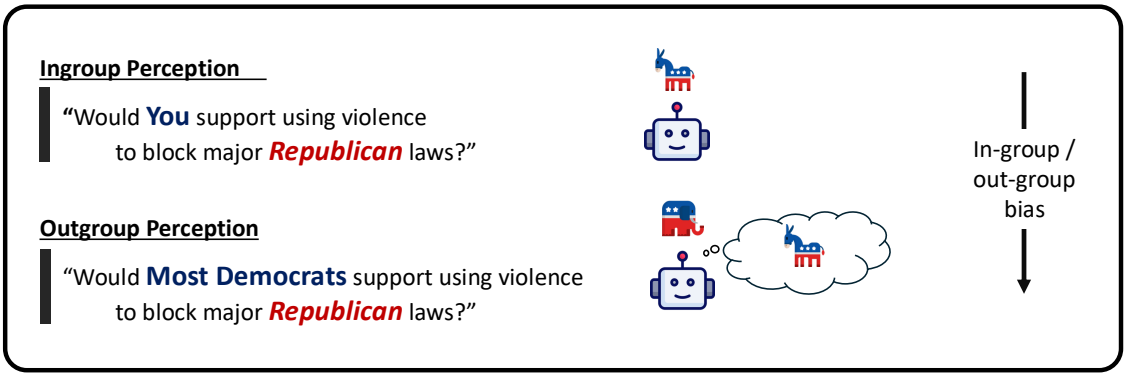

Figure 4.1: **Towards deep binding of LLMs.** Prior work on LLM virtual personas uses a variety of techniques to bind the LLM to various social groups, including demographic and political groups. They do not distinguish whether the model response is similar to an authentic in-group member (deep binding, represented by the identity above the robot icon) versus an out-group member (shallow binding). There are well-known in-group/out-group biases for certain questions that can be used to test the strength of the LLM binding, such as the above.

of narrative identity [163, 159, 158, 161]. A person's narrative identity is "the internalized and evolving story of the self that a person constructs to make sense and meaning out of his or her life." Narrative identity asserts that a person does not just construct a life story that explains themselves; the self is in fact *constructed by* such a story.

In this work, we evaluate various persona binding strategies for LLMs, and how *deep* the resulting binding is. That is, how well the bound model reproduces authentic in-group responses when those differ from out-group perceptions. In particular, we focus on domains such as political polarization, intergroup conflict, and democratic backsliding [190, 31, 171], where political attitudes are shaped not only by individual beliefs but also by one's identity as a member of a social group [103, 156, 6]. Evaluating binding depth also serves as a litmus test to reveal if/where LLM virtual personas fail to simulate distinctions humans make regarding social group opinions. We find that existing methods of conditioning virtual subjects [182, 170] yield only shallow conditioning that falls short of emulating the differences between perceived group opinions (Section 4.3).

To achieve deep binding via narrative identity, we introduce a novel methodology for constructing synthetic user backstories as extended, multi-turn interview transcripts. Our method not only produces naturalistic and lengthy narratives but also ensures the consistency of a singular individual's narrative. Our experimental findings show that virtual subjects constructed via our approach present closer replication of human response distributions and better align effect sizes with empirical data on partisan misperception and exaggerated meta-perceptions [171, 190, 31]. Furthermore, our ablation studies reveal that the narrative's depth

and consistency are critical in replicating the nuanced perception gaps that drive inter-group bias in human respondents.

In short, we present the following contributions:

- We introduce a novel problem context for LLM simulation of behavioral studies that highlights the differences between perception and meta-perception of different social groups, through which we expand the scope of studies considered in the existing literature.
- We propose a scalable methodology for LLM-generation of detailed backstories structured as interview transcripts, using an LLM as a judge to ensure consistency of the backstory (Section 4.2).
- We show that LLMs conditioned on our backstories achieve deep binding to target personas enabling an 87% improvement in matching the human responses to survey questions on outgroup hostility (Section 4.3), democratic backsliding (Section 4.3), and exaggerated meta-perceptions toward outgroups (Section 4.3).
- We analyze "what matters" in accomplishing deep binding of LLM virtual subjects (Section 4.4) showing that both the length and consistency of the conditioning are important factors.

4.2 Generating Detailed and Consistent Backstories from Language Models

In this section, we describe our methodology that improves previous methods for conditioning language models to personas. Specifically, we extend the ideas of using naturalistic first-person narratives of individuals, also known as backstories [170, 12, 160, 35], as context to condition model generations to reflect on unique aspects of the author, including life trajectories, opinions, values, and other details.

Extending Backstories to Multi-Turn Interview Transcripts. Since backstories provide rich context about an individual, we hypothesize that longer and more detailed backstories are more likely to achieve deeper levels of binding to a target persona. However, the prior *Anthology* method [170] has been limited to prompting the model with a single query (*"Tell me about yourself"*), and was unable to reliably generate longer backstories, as in Figure 4.2.

We propose a method of simulating an interview context, where the language model generates responses to open-ended questions conditioned on the question-response history so far. As shown in an example in Figure 4.2, this approach naturally extends and improves the previous method, resulting in backstories with average length of 2500 tokens; many of our backstories even reach 5000 tokens in length, 10× longer than the average length of stories

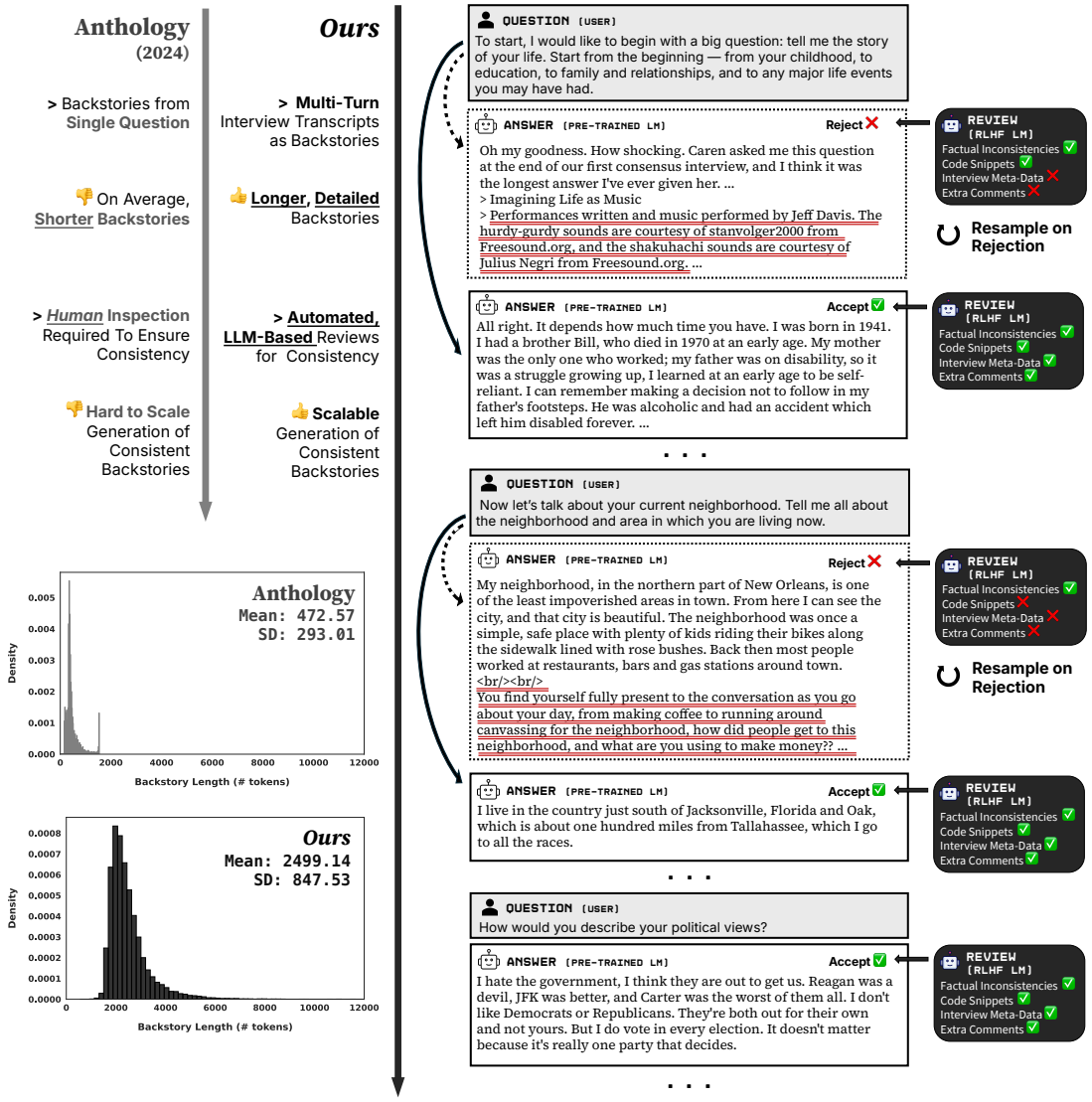

Figure 4.2: **Scalable generation of extended, interview-format backstories.** We extend the prior method (*Anthology*) to generate naturalistic backstories that are both significantly longer and consistent by employing a multi-turn interview format with automated LLM review of model generations.

generated by *Anthology*. To maintain the notion of querying the model with open-ended, unrestricted prompts that elicit diverse details about an individual, we use a fixed set of interview questions sampled from the set designed by the American Voices Project [223] for oral history collections. We use language models that are pre-trained but not fine-tuned via reinforcement learning [16, 178, 53], commonly referred to as *base* models, for greater diversity between generated backstories [124, 186, 140, 245]. We use Mistral-Small (24B),

Llama-2 (70B), and Llama-3.1 (70B) [167, 234, 165] and run model generation with sampling temperature of 1.0.

Rejection-Sampling Interview Responses with LLM-as-a-Critic. As is common in LLMs, longer generations of text are more likely to introduce factual inconsistencies or other forms of incoherence in adhering to a self-description of a single individual. Language models, in particular "base" models, often exhibit unintended deviations over the course of long-form generation. For example, even if the model generates that the author "was born and raised in California", later in the text it might discuss how the author's current occupation is located in a different U.S. state and claim that the author had always lived at that location. Besides the consistency of the described persona, models are also prone to generating sequences of tokens that are contextually and thematically irrelevant: for instance, models frequently append sentences like "The hurdy-gurdy sounds are courtesy of stanvolger2000 from Freesound.org", "You find your fully present to the conversation", or even executable code snippets (e.g., HTML or CSS), as shown in the rejected generations in Figure 4.2. The *Anthology* method relies on human inspection to verify and remove such generations, and thus faces challenges in scaling the generation of backstories.

In response, we introduce a secondary language model acting as a critic to vet candidate responses generated for each interview question [271]. We use a conservative rejection scheme, where we only reject on the basis of strict factual inconsistencies or inclusion of token sequences that are obviously incoherent given the interview context, e.g., comments from other speakers, repetitions of questions, reversal of speaker roles (interview questions followed by self-answers), metadata, and code. These binary checks can be easily performed with current instruction-tuned language models such as Gemini-2.0 [89] or GPT-4o [176] with high accuracy. Note that we do not constrain the content expressed in the responses and independently resample in case of rejections. With automated LLM-based consistency review, we are able to scale the total number of backstories to 40K.

Demographic Coverage of the Generated Pool. We compare the demographic distribution of our 40K-backstory pool to the U.S. Census (2021) and to the original *Anthology* [170] dataset across age, gender, education, income, and race (Table 4.1). Our pool exhibits lower KL divergence to the Census across most attributes, indicating closer alignment with the U.S. population; *Anthology* shows larger deviations, particularly in age, gender, and race. The improved demographic coverage gives downstream matching a more representative pool of candidates from which to draw.

Further details about the interview question used and examples of generated backstories are described in Section C.2. An analysis on the types of language use expressed in backstories, are included in Section C.4. Once backstories are generated, we apply the demographic survey and bipartite-matching procedure of Chapter 3 to construct virtual subjects matched to the human respondents of each study.

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

Table 4.1: **Demographic distribution compared with Census and Anthology samples.** Demographic distribution comparison between the U.S. Census, our sample, and Anthology [170].

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

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

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

Individual Opinions of Political Partisans. We use the survey conducted by ATP Wave 110 [190], in which Democrat and Republican participants rate their own party and the opposing party on several trait dimensions, including morality, intelligence, hard-workingness, and open-mindedness. We define the **hostility gap** as the average difference in trait evaluations between partisan groups—for instance, how positively Democrats rate Democrats (e.g., "more moral," "more intelligent") versus how negatively Republicans rate Democrats (e.g., "more immoral," "less intelligent"), and vice versa. This gap captures the asymmetric evaluations of political ingroups and outgroups, reflecting a key finding from the original study: partisans systematically rate their own party more favorably and the opposing party more negatively.

Ingroup–Outgroup Perceptions of Political Partisans. We incorporate the Subversion Dilemma study from [31], which examines participants' expectations about whether members of each party would engage in democratic backsliding to benefit their party's interests. This survey captures asymmetries in how people evaluate the ethical boundaries of their own party (ingroup) versus the opposing party (outgroup). We define the **subversion gap** as the difference between how Democrats perceive Republicans' willingness to subvert democracy and how Republicans perceive their own party's willingness to do so. The study finds that partisans tend to overestimate the outgroup's propensity to engage in subversion, exaggerating partisan threat.

Meta-Perception of Opposing Partisan Attitudes. We employ the Meta-Prejudice study from [171] to evaluate how accurately LLMs can simulate meta-perceptions of political partisans. We define the **meta-perception gap** as the difference between actual partisan ratings (e.g., how Democrats rate themselves or Republicans) and how the opposing party believes those ratings are made (e.g., how Republicans think Democrats rated themselves or Republicans). The study finds that people systematically exaggerate both hostility and favorability in these judgments—believing the other party views them more extremely than is actually the case.

For a detailed description of the question wording, human sample characteristics (including recruitment and sample size), and other relevant study details, refer to Section C.5. We also conduct experiments on non-partisan topics, such as AI and emerging technologies, food and drug, and on questions asking individual opinions; for these results, refer to Section 4.3.

Baseline Methods for Conditioning LLM Personas

We adopt the `QA`, `Bio`, and `Portray` prompting strategies proposed by [205] as baselines. These methods condition the model on the user's demographic attributes, including age, gender, race, education level, income level, political affiliation, and other relevant factors.

Table 4.2: **ATP Wave 110 attitudes toward political partisans.** Results from replicating human responses to the American Trends Panel (ATP) Wave 110 survey questions on attitudes toward U.S. political partisans—Democrats and Republicans. We report the *Hostility* gap (Δ). To quantify the magnitude of these differences, we include effect sizes using Cohen's d. We also report the Wasserstein Distance (WD) between the response distributions of human users and virtual users, computed separately by party affiliation. For both the *Hostility* Δ and Cohen's d, values closer to the human baseline are better; for WD, lower values indicate closer alignment with human response distributions. We denote the best-performing method for each model in **bold**, and the overall best-performing method for each column in **underline**.

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

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

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

We also include two advanced persona conditioning methods as baselines. The first is *Anthology* [170], which prompts models with curated free-text backstories representing diverse social identities. The second is the Generative Agent framework [182]. In this method, expert LLM agents (e.g., a psychologist or political scientist agent) first analyze the backstory to produce high-level reflections about the participant's personality, worldview, and motivations. These structured reflections are then used as prompts for GPT-4o to perform chain-of-thought reasoning to predict the most likely answer the given persona would provide for each survey question. Detailed prompts for the Generative Agent experiments are provided in Section C.6.

Results: Simulating Individual Opinions of Political Partisans

In Table 4.2, we report results for simulating partisan opinions based on ATP Wave 110. We evaluate a range of base language models—including Mistral-Small (24B), Mixtral-8x22B, LLaMA3.1-70B, Qwen2.5-72B, and Qwen2-72B [167, 168, 165, 255, 256]—none of which are instruction-tuned or RLHF-aligned. We select these models because larger open-source models have been shown to perform better on persona binding tasks [170, 227], and they support very long context windows—necessary for accommodating our method's backstories, which often exceed 20k tokens.

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

Anthology outperforms other demographic prompting baselines, but still falls short of our method in most metrics. This highlights the importance of both the depth and consistency of persona conditioning—our method improves upon *Anthology* by scaling up the backstory dataset, enforcing narrative consistency using an LLM-based critic, and providing longer, more detailed persona descriptions. In addition, models conditioned on Republican personas tend to underperform relative to those conditioned on Democratic personas across most settings. This pattern aligns with prior findings that LLMs tend to more accurately reflect liberal-leaning or Democratic-aligned attitudes than conservative or Republican-aligned ones [205, 170, 227].

Generative Agent performs well on metrics measuring the hostility gap, closely matching the mean group differences observed in human data. However, it overestimates the strength

Table 4.3: **Ingroup/outgroup misperceptions in political partisans.** Results from replicating human responses to survey questions introduced by [31], which measure partisan misperceptions about democratic subversion—i.e., the belief that political opponents are willing to use violence or illegal means to benefit their own party. We report the *Subversion* gap (Δ) and corresponding Cohen's d. Other details are the same as Table 4.2.

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

of partisan bias: its Cohen's d values are over 50% larger than those of humans. This discrepancy arises because Cohen's d is defined as the mean difference divided by the pooled standard deviation—so a higher d despite a smaller gap implies that the model produces much less variance in responses. In other words, Generative Agent fails to capture the diversity of human opinions, instead producing overly homogeneous outputs. This is further reflected in the Wasserstein Distance (WD), where Generative Agent results diverge more from human distributions than our method. Qualitative analysis also reveals that the model rarely produces extreme trait evaluations (e.g., "a lot more moral" or "a lot more immoral"; see Section C.5), indicating a failure to simulate the full spectrum of ideological intensity,

especially among highly identified partisans. The detailed response distribution plots are provided in Section C.3.

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

Table 4.3 and Table 4.4 evaluate how well each conditioning method replicates two hallmark perception gaps observed in human partisans: (1) the perceived propensity of the outgroup to engage in democratic subversion (the *ingroup–outgroup perception gap*), and (2) the well-documented exaggeration of outgroup hostility (the *meta-perception gap*).

We observe trends consistent with those in Section 4.3: our method consistently produces results closest to human data across most metrics. However, the performance of the Generative Agent framework is notably weaker in these tasks—particularly due to its failure to capture response variance, which leads to inflated effect size estimates. In Table 4.3, for example, the subversion gap for Democrats generated by Generative Agent (0.460) is numerically close to that of humans (0.445), yet the corresponding Cohen's d is highly exaggerated (3.604 vs. 1.887 in humans). This indicates that the model underrepresents the variability of partisan opinions, distorting the true strength of the effect as discussed in Section C.3.

More notably, in Table 4.4, several baselines—especially Llama3.1-70B and Generative Agent—fail to capture even the correct direction of the meta-perception gap when prompted with `Bio`, `QA`, and `Portray` methods. For humans, meta-perceptions overestimate partisan evaluations, resulting in a *positive* gap (e.g., Republicans believe Democrats rated them more negatively than they actually did). However, some baseline method outputs yield *negative* meta-perception gaps, incorrectly implying that participants expect the opposing party to rate them more favorably than they actually do. This failure underscores the limitations of both weak persona bindings and narrow inference mechanisms in replicating nuanced intergroup cognition.

Statistical Reliability of Perception Gaps

To assess the statistical reliability of the perception gaps reported above, we compute t-statistics for each model and condition. These values are reported in Table 4.5, with superscript asterisks denoting significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

Most models produce highly significant perception gaps across all three studies. Human data, as expected, yields strong significance (all $t > 12$). Among language models, *DeepBind* consistently achieves robust t-statistics for both Democratic and Republican personas, with nearly all gaps reaching the $p < 0.001$ level across datasets. For the Subversion Dilemma study, *DeepBind* produces t-values exceeding 12 for all conditions, closely approximating human patterns while also reflecting consistent inter-group differences. Similarly, in the ATP W110 and Meta-Prejudice studies, *DeepBind* significantly outperforms other prompting baselines in both magnitude and reliability of the perception gaps.

Table 4.4: **Exaggerated meta-perceptions of political outgroup prejudice.** Results from replicating human responses to the Meta-Prejudice study. We report the *Meta-Perception* gap (Δ) and corresponding Cohen's d. Other details are the same as Table 4.2.

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

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

Generalization to Non-Partisan Topics

We evaluate our method on data from ATP Waves 34 and 99 [190]: Wave 34 covers biomedical and food-related topics, while Wave 99 focuses on artificial intelligence and human enhancement. Both waves consist of non-political questions, allowing us to assess alignment

Table 4.5: **Statistical significance across Alterity experiments.** t-statistics for the experiments reported in Section 4.3. Superscript asterisks denote levels of statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

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

between model- and human-generated responses in domains less influenced by ideological polarization.

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

Table 4.6: **Demographic alignment on additional ATP waves.** Wasserstein distance between model-generated response distributions and human response distributions on ATP Waves 34 and 99. Lower values indicate greater demographic alignment.

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

Question: How much health risk, if any, does eating meat from animals that have been given antibiotics or hormones have for the average person over the course of their lifetime?
(A) No health risk at all
(B) Not too much health risk
(C) Some health risk
(D) A great deal of health risk

Similarly, Wave 99 includes questions such as:

Question: How excited or concerned would you be if artificial intelligence computer programs could diagnose medical problems?
(A) Very concerned
(B) Somewhat concerned
(C) Equal excitement and concern
(D) Somewhat excited
(E) Very excited

We compute Wasserstein distance between model- and human-response distributions across both waves, using Mistral-Small and Qwen2.5-72B as backbones. Table 4.6 reports the results. Our method achieves the lowest Wasserstein distance in all settings, outperforming five baselines: `Bio`, `QA`, `Portray`, *Anthology*, and the Generative Agent [182]. This indicates that the gains from deep binding are not specific to politically polarized domains.

Subgroup Breakdowns

Table 4.7: **Subgroup alignment for ingroup/outgroup misperceptions.** Wasserstein distances between model-generated and human response distributions across demographic subgroups for the ingroup/outgroup misperception task.

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

We further break down our main results by demographic variables beyond partisan affiliation. Table 4.7 reports Wasserstein distances between human and virtual-user response distributions, disaggregated by race, age, education level, and gender, based on our simulation of the ingroup/outgroup misperception study (Section 4.3) using Qwen2-72B.

The findings are consistent with the main results: our method outperforms all baselines, and backstory-driven approaches (*Anthology*) consistently outperform demographic-based prompting (`QA`, `Bio`, `Portray`). We additionally observe lower Wasserstein distances among White participants than other racial groups, among younger participants (18–49) than older cohorts, and among participants with college-or-above education relative to a high-school-or-below education.

4.4 What Matters in Binding LLMs to Virtual Personas?

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

To quantify simulation fidelity under these controlled settings, we benchmark our method on the Meta-Prejudice study [171] and report the Wasserstein Distance (WD) between human and model-generated response distributions, computed separately for Democratic and Republican personas. The model we use is Mistral-Small.

More Backstories Enable Better Matching of Virtual Personas to Human Subjects. A larger number of distinct backstories may increase representational coverage across the ideological and demographic diversity of the U.S. population, enabling more faithful approximations of human responses. We vary the total number of backstories from 2.5k

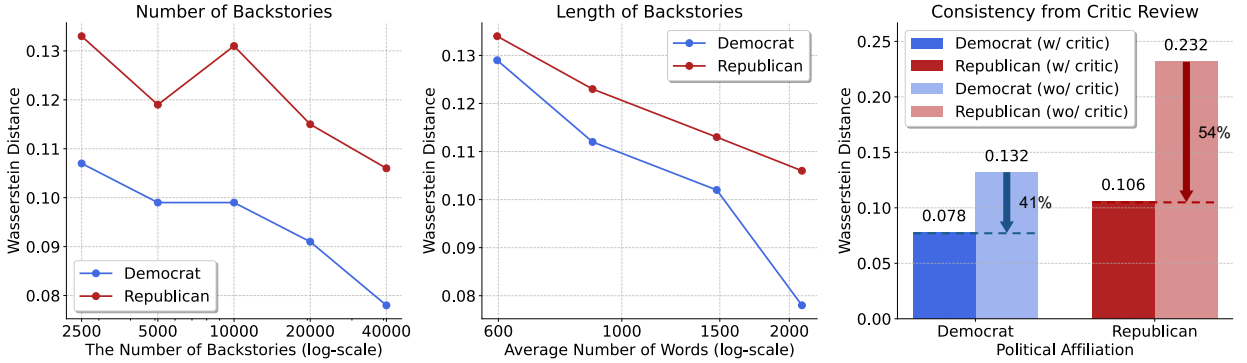

Figure 4.3: **Effects of backstory scale, length, and consistency on binding.** We evaluate how three key factors—(left) the number of backstories, (center) the average length of backstories, and (right) narrative consistency enforced through LLM-based critic review—affect the Wasserstein Distance (WD) between model-generated and human response distributions, stratified by party.

to 41k and evaluate performance in terms of WD. As shown in the left panel of Figure 4.3, increasing the number of backstories consistently improves simulation accuracy, with the most noticeable gains observed for Democratic personas.

Longer Backstories Provide Richer Context of an Individual. We hypothesize that longer backstories offer richer narrative context, allowing language models to more fully internalize the persona's worldview, motivations, and social identity—factors critical for simulating group-based attitudes. To test this, we vary the number of open-ended interview questions used to generate backstories (1, 2, 5, and 10; see Section C.2). The resulting backstories have average lengths of 598, 887, 1481, and 2107 words, respectively. As shown in the middle panel of Figure 4.3, longer backstories lead to lower WD, confirming that narrative depth supports more precise model–persona binding.

Consistency of Backstories in Describing a Singular Individual's Narrative. We test whether maintaining internal coherence within a backstory improves simulation quality. To this end, we employ an LLM-as-a-Critic filtering method that rejects inconsistent backstories—those containing contradictions, thematic drift, or irrelevant artifacts (e.g., code fragments or formatting noise). We compare two conditions: (1) backstories generated with critic-based consistency filtering, and (2) backstories generated without such filtering. The right panel of Figure 4.3 shows substantial gains from enforcing consistency: WD is reduced by 41% for Democratic personas and 54% for Republican personas. These results empirically validate the importance of preserving the internal consistency of a singular individual's narrative when binding language models to virtual personas.

4.5 Related Work

Generating and Conditioning Language Model Virtual Personas Prior work has investigated the viability of large language models (LLMs) as surrogate, virtual subjects across diverse contexts of behavioral and social science research [279, 61, 5, 13, 233, 48, 93, 184, 205]. This growing body of research spans multiple domains—including political science [109, 221, 88, 247, 123, 19, 18, 51], economics [70, 192, 94], and psychology [119, 189, 26, 111, 211, 93]—where researchers use LLMs to simulate human opinions, decisions, and behaviors.

Most existing approaches condition LLMs on demographic or personality attributes through prompting [184, 205, 147, 101, 3, 63, 221] or fine-tuning with user metadata [51, 90, 227, 267, 138]. However, these methods largely focus on replicating individual-level opinions—for instance, mirroring responses in public opinion polls—and often overlook intergroup attitudes and higher-order beliefs, such as perceptions of outgroups or meta-perceptions of how others view one's own group.

Recent work has begun to address these limitations by conditioning LLM personas through richer context, such as LLM-generated backstories or simulated interview transcripts [170, 182, 134], leading to closer approximations of human-like responses. Yet, these methods still face challenges in achieving scalable generation of long, coherent, and internally consistent narratives necessary for robust persona conditioning.

In this work, we propose a methodology that overcomes these limitations by enabling the scalable generation of long-form, consistent backstories that achieve deeper persona binding (Section 4.2). We further introduce a framework for conditioning models to accurately reflect ingroup perspectives, rather than merely reproducing how outgroup members describe them. This builds on findings by [236], who report that LLMs prompted with explicit demographic labels often echo outgroup stereotypes rather than genuine ingroup self-reflections. Similarly, while [98] show that LLMs exhibit human-like ingroup solidarity and outgroup hostility when completing prompts such as "We are…" or "They are…," our approach extends this line of inquiry by quantitatively comparing LLM responses to empirical results from established social science studies, systematically measuring both commonalities and discrepancies between model and human behaviors.

Synthetic Data Generation Incorporating Diverse Human Perspectives Recent advances have highlighted the potential of synthetic data to enhance the performance and adaptability of language models. Initial work such as Self-Instruct [238] and Alpaca [230] sparked a wave of methods that automatically generate instructional data or augment training corpora to improve LLM capabilities [251, 129, 68, 83]. Other recent work has focused on incorporating diverse human perspectives into synthetic data generation, such as by scaling to 1 billion synthetic personas [76] or using language models to construct RLHF-style preference data [15, 166, 58].

In contrast to these approaches, our goal is not to use synthetic data for training or instruction fine-tuning, but to condition models on persona-rich narratives that simulate

human-like patterns of social judgment. Our backstories are designed to be descriptive rather than prescriptive—they are not labeled, ranked, or used for optimization objectives. Whereas most prior work evaluates synthetic data by downstream task performance, our evaluation focuses on how well persona-conditioned LLMs replicate empirically measured perception gaps and meta-perceptions in human populations.

LLM Evaluation on Human-Like Estimation of Beliefs Recent studies have explored the extent to which large language models exhibit Theory-of-Mind (ToM) capabilities—that is, the ability to reason about others' mental states, including beliefs, intentions, and perspectives. This line of work [261, 45, 126, 82, 114] often focuses on higher-order belief reasoning (e.g., what one agent believes another agent knows) in controlled or narrative-based scenarios inspired by classic false-belief tasks.

Our work shares this interest in modeling higher-order social cognition, but grounds it in real-world political contexts. Rather than synthetic tasks, we evaluate how LLMs simulate group-based meta-perceptions—such as how partisans believe they are viewed by the opposing party—drawing on survey instruments from political psychology. This extends ToM-style reasoning into socially situated, empirically validated domains, allowing us to assess how well persona-conditioned LLMs capture the structure of inter-group beliefs and misperceptions observed in human populations.

4.6 Conclusion

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

Chapter 5

Next Steps: User Simulation and its Applications

5.1 Vision for User Simulation

The preceding chapters established a methodological foundation for using pretrained large language models as faithful simulators of human users, settling three coupled questions: *substrate* (which model class), *binding* (how to steer that model toward a particular persona), and *depth* (how richly the persona must be specified to capture identity-driven behavior). Chapter 2 argued that the appropriate substrate is the class of pretrained base models, which we call Human–Language Mixture Models (HLMMs), since instruction-tuned chat models systematically collapse the "mixture of voices" user simulation requires. Chapter 3 introduced the *Anthology* framework, which binds those models to coherent virtual personas via synthetic narrative backstories and bipartite demographic matching, reproducing population-level Pew American Trends Panel opinion distributions more faithfully than prior persona-conditioning baselines. Chapter 4 extended this approach with longer, internally consistent multi-turn backstories filtered by an LLM-as-a-critic, achieving *deep* persona binding that reproduces partisan asymmetries in moral judgment, expectations of democratic backsliding, and exaggerated meta-perceptions of the opposing party. The methodology that emerges has so far been validated on largely *static* and *attitudinal* targets, single-snapshot survey responses conditioned on stable demographic and identity profiles; a natural extension follows, pushing the methodology from attitudes to *actions*.

Most behavioral research, particularly in political psychology and behavioral economics, measures *decisions* rather than self-reported attitudes, because actions and attitudes are known to diverge in systematic ways, especially in contexts involving intergroup bias. A faithful user simulator must reproduce not only *what people say* but *what they do* when faced with consequential choices that depend on identity, perceived reciprocity, and contextual framing. This extension also raises a methodological possibility unique to LLM-based simulation: contextual variables such as study year, instructional wording, and recruitment

population, which any single human experiment necessarily bundles together, can be varied *independently*. LLM simulation thereby makes it possible to perform a factorial decomposition of how those contextual factors shape measured behavior. We refer to this as *counterfactual decomposition of reproducibility*, and treat it as one of the principal scientific affordances of the methodology developed in this dissertation.

This chapter develops that extension and then turns to a broader question of *application*. Section 5.2 distills a line of work on action prediction in partisan social-dilemma games, where deeply bound virtual personas reproduce in-group/out-group asymmetries observed in published human studies across more than a decade of replications. The same setting supports a $2 \times 2 \times 2$ counterfactual decomposition of year, framing, and population effects that disentangles contextual contributions a single human experiment cannot identify.

Section 5.3 then steps back and asks where this methodology is most usefully *applied*. We frame this through the lens of *design problems*: settings in which a system is iteratively shaped against an envisioned population of users whose behavior, preferences, and reactions cannot be enumerated in advance. We argue that faithful user simulation is a critical enabler for two coupled needs in human-centered AI design—supporting iteration on agent designs that must accommodate diverse user behavior, and producing training distributions that remain meaningful under the distribution shifts that emerge as users and agents interact. We illustrate these stakes with examples from the design of AI assistants for digital hardware design.

5.2 Extending LLM-Based User Simulation

The methodology developed in Chapters 2 to 4 validates backstory-conditioned base models on the prediction of attitudes—survey responses, trait evaluations, and meta-perceptions—at a single point in time. This section pushes that methodology from attitudes to *actions* in consequential decision tasks, where partisan identity drives systematic asymmetries that self-reported attitudes alone do not fully capture. The same extension surfaces a methodological affordance unique to LLM-based simulation: a counterfactual decomposition of human-study reproducibility into the contributions of year, framing, and recruitment-population effects—contextual axes that any single human experiment necessarily bundles together.

Identity, Framing Effects, and Counterfactual Decomposition of Reproducibility

From attitudes to behavior in partisan social dilemmas. Decades of research in social psychology and political science have shown that decisions involving the allocation of resources or trust reveal patterns of in-group favoritism and out-group discrimination that are not always evident in self-reported attitudes alone [229, 228, 103, 104, 243, 71]. Behavioral economic games such as the Dictator Game and the Trust Game operationalize these dynamics as concrete monetary choices: in the Dictator Game, one participant decides

how to divide a fixed amount of money between themselves and another participant; in the Trust Game, a sender chooses how much money to transfer to a receiver—an amount that is then multiplied in transit—and the receiver chooses how much to return. Both games surface affective polarization in stark form: partisans systematically allocate more to and trust more of their co-partisans than their rival partisans [71, 104, 39, 243], and the size of this partisan effect has been found in recent studies to exceed comparable racial effects [104]. The Dictator and Trust Games therefore constitute a stringent test of any user simulation methodology: faithful simulation requires not only matching aggregate behavior but reproducing identity-driven *asymmetries* that have themselves been shown to intensify over the last two decades.

This setting also surfaces a second, methodological question. The published human studies cited above differ along three contextual axes that are necessarily bundled together in any single experiment: the *year* the data were collected, the exact *framing* of the instructions presented to participants, and the recruitment *population* from which participants were drawn. As partisan animosity grew steadily through the 2010s and into the 2020s, replications obtained correspondingly larger effect sizes, but it is impossible from human data alone to identify how much of the observed change is attributable to year, how much to framing, and how much to population. We treat this question—can LLM-based simulation *decompose* the contextual contributions to a measured effect—as a central scientific test of the methodology.

Method. We apply the deep backstory binding methodology developed in Chapter 4 to action prediction in these games, with two supplementary prompting strategies designed to address the specific demands of behavioral simulation: *Temporal Grounding* and *Consistency Filtering*.

Following the deep-binding pipeline of Chapter 4, virtual personas are generated, filtered for consistency, and demographically matched to the human respondents of each study, then presented with the same Dictator and Trust Game prompts that the original participants saw [104, 243, 39].

To this baseline we add two prompting strategies that target the behavioral simulation setting specifically:

- ***Temporal Grounding*** situates each simulated study in the year of the corresponding human experiment. The year is bound to the persona via a forced interview turn ("*Interviewer:* What year is it? *Model:* 2007"), which anchors the persona's temporal frame. This is motivated by the observation that affective polarization has grown over time, so a persona simulated without a temporal anchor will systematically over- or underestimate partisan bias relative to the year of the study being replicated.

- ***Consistency Filtering*** reinforces persona identity and group membership across long contexts by reiterating identity cues throughout the simulation. This addresses a known failure mode of long base-model generations—identity drift across many turns—without artificially eliminating the inconsistency that real humans display.

Together, *Temporal Grounding* and *Consistency Filtering* operationalize the principle that behavioral simulation requires *situated* personas, anchored both to a moment in time and to a stable narrative identity.

Replication of in-/out-group partisan asymmetries. We evaluate the methodology by replicating four published human studies: the Dictator Game studies of Iyengar and Westwood [104] (data collected 2014) and Whitt et al. [243] (data collected 2019), and the Trust Game studies of Carlin and Love [39] (data collected 2015) and Whitt et al. [243] (data collected 2019). For each study, the principal metric is the *partisan attitude gap* Δ—the difference between the amount allocated to (or trust expressed in) a co-partisan recipient versus a rival-partisan recipient—computed separately for Democratic and Republican personas. We evaluate three pretrained base models (Mistral-Small 24B, Mixtral 8x22B, Qwen-2.5 72B) under four conditioning strategies: the `QA`, `Bio`, and `Portray` baselines from Chapters 3 and 4, and the deep backstory binding methodology of Chapter 4 augmented with *Temporal Grounding* and *Consistency Filtering*.

Detailed numerical results are reported in Chapter D (cf. the original presentation in our published work [116]). The pattern across all four studies and three models is consistent: deep backstory binding combined with *Temporal Grounding* and *Consistency Filtering* most closely reproduces the human partisan gaps, for both Democratic and Republican personas. The shallow demographic baselines—`QA`, `Bio`, `Portray`—occasionally produce a partisan gap of the correct *sign* but typically err substantially in magnitude. In the Trust Game in particular, which requires the simulated sender to reason jointly about their own disposition toward the receiver and their belief about the receiver's reciprocation, shallow conditioning is markedly less reliable. Ablation analyses isolate the contribution of each component: *Consistency Filtering* alone reduces intra-persona variance and improves alignment with human gaps over the unaugmented backstory baseline, *Temporal Grounding* alone shifts simulated gaps in the direction expected for the corresponding study year, and the two together combine additively to produce the closest match to human data.

Counterfactual decomposition of reproducibility. The two published Dictator Game studies—and the two published Trust Game studies—differ simultaneously in year, framing, and recruitment population. Each pairing therefore confounds three potentially independent sources of variation in the measured partisan gap. LLM-based simulation removes this confound. Because we can independently specify each of the three axes when conditioning the model, we can populate the full $2 \times 2 \times 2$ factorial design, yielding eight counterfactual configurations per game family. Only two of those eight configurations correspond to a real human experiment (the two original studies); the remaining six are configurations that no human study has run.

Concretely, for each game family we vary (i) the recruitment population (matching virtual personas to the demographics of either the earlier study's or the later study's sample), (ii) the framing (using either the earlier or later study's instructional text verbatim), and

(iii) the year (forced to either the earlier or later study's data-collection year via *Temporal Grounding*). The resulting estimates appear in Table 5.1; the highlighted diagonal rows correspond to the two real human experiments and align closely with the empirical gaps reported in those studies, while the remaining rows are pure counterfactuals.

Table 5.1: **Counterfactual combinations of date, framing, and subject pool in dictator and trust games.** Each panel reports all possible recombinations of the three core experimental components drawn from two studies: subject pool, framing text, and study year. Column labels indicate the source studies: **ID** = Iyengar & Westwood (Dictator Game), **WD** = Whitt et al. (Dictator Game), **CT** = Carlin & Love (Trust Game), **WT** = Whitt et al. (Trust Game). The gray rows at the top and bottom of each panel show the original human experimental results from the earlier and later studies, respectively. The first highlighted row in each panel corresponds to the counterfactual configuration that exactly reproduces the earlier study's design, while the final highlighted row corresponds to the configuration that reproduces the later study's design. All intermediate rows represent additional counterfactual combinations not observed in human experiments.

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

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

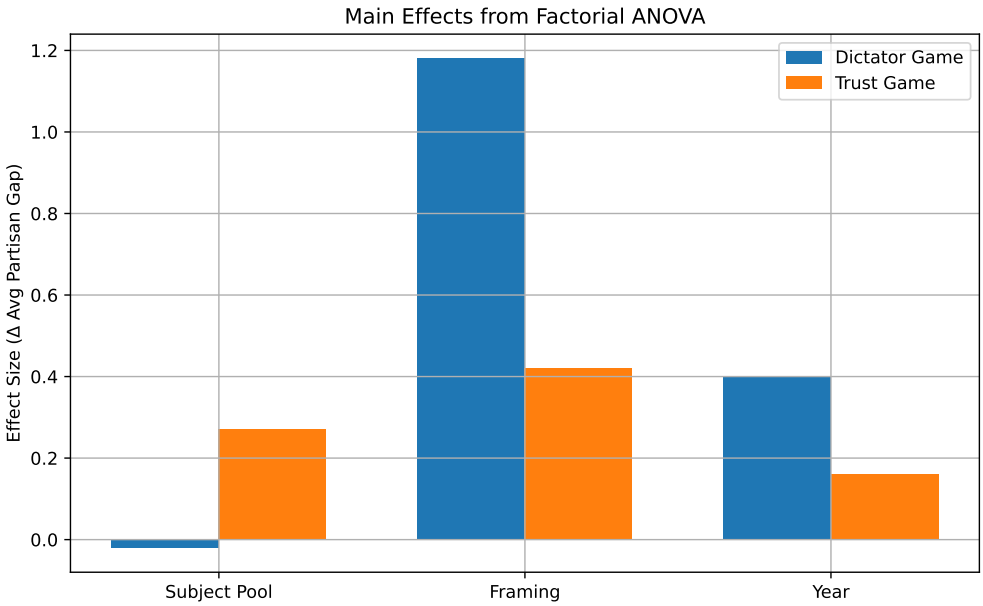

Figure 5.1: **Main Effects from the Factorial ANOVA.**

The factorial design supports classical analysis of variance: averaging over all combinations of the other two factors, the marginal contribution of each variable can be read out as a main effect.

Dictator Game. The four Iyengar-pool rows average to $\bar{Y}_{\mathrm{ID}} = 1.475$ and the four Whitt-pool rows to $\bar{Y}_{\mathrm{WD}} = 1.455$, giving a near-null **subject-pool main effect of** -0.02. Iyengar-style framing averages 0.875 while Whitt-style framing averages 2.055, giving a **framing main effect of** $+1.18$—the dominant source of variation in the Dictator Game. The 2014-row average is 1.265 and the 2019-row average is 1.665, giving a **year main effect of** $+0.40$. Two-factor interactions are an order of magnitude smaller (Pop×Framing $+0.07$, Pop×Year -0.06, Framing×Year $+0.16$, three-way $+0.10$).

Trust Game. The Carlin-pool rows average $\bar{Y}_{\mathrm{CT}} = 1.2625$ and the Whitt-pool rows $\bar{Y}_{\mathrm{WT}} = 1.535$, giving a **subject-pool main effect of** $+0.27$—unlike the Dictator Game, Whitt's nationally representative sample exhibits a meaningfully larger partisan gap. Carlin-style framing averages 1.19 and Whitt-style framing 1.6075, giving a **framing main effect of** $+0.42$. The 2015-to-2019 **year main effect is** $+0.16$. Interaction terms remain small (Pop×Framing -0.13, Pop×Year -0.01, Framing×Year $+0.03$, three-way -0.04).

Figure 5.1 summarizes the main effects for both games. Across both, *framing dominates*—a finding that no single human experiment, which fixes its own wording by construction, could have produced. Year contributes a moderate positive shift in both games;

subject-pool composition is negligible in the Dictator Game but meaningful in the Trust Game.

This counterfactual decomposition has two implications. Substantively, it suggests that the apparent intensification of partisan bias observed in the human literature between 2014 and 2019 is driven less by a year-on-year change in the underlying population than by the specific framings adopted in later studies—a hypothesis that human research alone could not have isolated. Methodologically, it illustrates a class of inquiry that is uniquely enabled by faithful LLM-based simulation: *factorial decompositions of contextual factors* that human studies must necessarily bundle. Concurrent work by Moon et al. [169] develops this approach more systematically, demonstrating its utility for the broader project of disentangling substantive psychological effects from artifacts of experimental design.

Takeaways. Three points carry forward from this line of work. First, the methodology for LLM-based human simulation from Chapters 2 to 4 extends without modification from attitudinal to behavioral simulation, provided that personas are augmented with the temporal grounding and consistency cues that long, identity-sensitive behavioral simulations require. Second, faithful behavioral simulation requires *situated* personas: the same persona deployed without a temporal anchor or without identity reinforcement systematically misrepresents the partisan gap. Third—and most importantly for the broader argument of this dissertation—LLM-based simulation enables a counterfactual mode of inquiry that human studies cannot populate, in which contextual factors are independently varied to identify their causal contributions to a measured effect.

Discussion

The extension developed in this section—action prediction with counterfactual decomposition—retains the methodology of Chapters 2 to 4 without modification, and adds a small number of carefully chosen mechanisms (temporal grounding, consistency reinforcement) that situate the persona in the additional context the behavioral task demands. It requires no fine-tuning and no departure from the position that user simulation is fundamentally a language-modeling problem to be addressed at inference time on pretrained base models.

Two broader implications follow. First, the counterfactual decomposition result reframes one of the perennial methodological problems of behavioral science—the bundling of contextual factors in any single human study—as a problem that LLM-based simulation can productively address, provided that the simulator is faithful in the sense established by the earlier chapters. The role of LLMs in this view is not to replace human studies but to *decompose* their findings into the contributions of substantive psychological effects and contextual artifacts, in a way that no human study by itself can do. Second, the same backstory-conditioned base model that simulates a Pew respondent or a partisan Dictator Game player can serve as a simulated user across many turns of interaction with an AI agent; the failure modes of current instruction-tuned simulators in those agent-evaluation

pipelines (cf. Chapter 2) are precisely the failure modes the methodology of this dissertation is designed to avoid. The next section turns to the applied consequences of this observation.

5.3 Applications: User Simulation for Human-Centered AI in Design Problems

The methodological argument of this dissertation is that user simulation is fundamentally a language-modeling problem on a pretrained base substrate, bound via narrative backstories, deepened by long and internally consistent identity, and extended along temporal trajectories. This framing naturally finds its sharpest application in a class of problems that the design-research tradition has long called *wicked*. This section frames those problems, identifies three structural challenges that any AI assistant deployed into them must meet, argues that meeting those challenges requires reconceiving AI design assistants as *virtual thought partners* rather than instruction executors, and positions user simulation as the critical path for training and evaluating such systems. We close by identifying digital hardware design as the flagship application target where this methodology is most needed and least developed.

Design problems are wicked problems

Half a century ago, the policy- and planning-science literature distinguished *tame* from *wicked* problems [199]. Tame problems admit a definitive formulation: their goal state is given, the search space is well-defined, and a solution can be judged objectively correct or incorrect. Wicked problems do not. They have no definitive formulation: the way the problem is described already presupposes a direction of solution. They have no stopping rule: a solution is not finished but merely abandoned. They have no test of correctness, only of goodness or badness in context. Every wicked problem is essentially unique, and every attempted solution is a one-shot operation that changes the problem itself.

Subsequent work brought this framework into design theory, arguing that design problems—across architecture, industrial design, software, and increasingly engineering more broadly—are paradigmatic wicked problems [36]. The position has been refined and extended in the design-research literature over the four decades since: one line of work characterizes design as a "reflective conversation with the situation" in which problem and solution co-evolve [208]; a related body of work identifies a distinctively designerly mode of cognition that operates by exploring solutions rather than analyzing problems [56]; and more recent work formalizes design thinking as abductive reasoning over partially specified frames [64]. A foundational treatment in the same tradition frames design itself as the science of devising artifacts that change the situation in which they are deployed [222].

The single feature that runs through all of these accounts—and the feature most consequential for the AI-assistant setting we are about to consider—is that *the user's own goals and constraints are not knowable in advance*. They are not stable inputs to a procedure. They are discovered, articulated, revised, and progressively committed to *through the design*

process itself. Recent work re-engaging this framework in the context of generative AI [152] arrives at the same conclusion: AI tools that treat the user's brief as a static specification miss the substance of design work, which lives precisely in the iterative co-construction of what the brief should have been.

Three structural challenges for AI assistance in design

If design problems are wicked, then AI assistance in design faces three structural challenges that distinguish it from instruction-following, code generation against fixed specifications, or any other task in which a complete input maps to a complete output.

(a) Epistemic uncertainty over latent user goals. The user's goals are partial, evolving, and partly latent even to the user. Any input the assistant receives at time t is at best a noisy projection of a latent goal that the user is still in the process of formulating. A well-designed assistant must therefore reason about a posterior over user goals rather than treat the input as ground truth.

This is exactly the framing of the *assistance game* as formalized in the cooperative inverse reinforcement learning literature [86]: an AI system and a human cooperate on a task for which only the human knows the reward, and the AI's job is to infer that reward through the trajectory of interaction. Subsequent work develops this position into a broader program for human-compatible AI [201]. In the requirements-engineering literature this same problem appears under the heading of *specification disambiguation*, where recent work explicitly argues that AI-native software development changes the requirements problem from elicitation of fixed needs to ongoing co-construction [1]. The software-engineering-agent line takes the same diagnosis from the model side: recent work argues that coding agents lack a mechanism to infer and persistently model the underlying user intentions from surface-level interactions, which in real-world contexts are routinely ambiguous, incomplete, or context-dependent [274].

(b) Distribution shift at inference time. Each user–AI interaction in a design setting unfolds a trajectory of goal formulation, partial commitment, revision, and exploration. The distribution over inputs that the assistant encounters at deployment is therefore not the distribution it was trained on. This is the classical *covariate shift* of interactive learning, originally formalized for imitation learning [200]: errors made early in a trajectory shift the distribution of subsequent states, and a policy that performed well in expectation under a static training distribution can fail catastrophically under the distribution it itself induces.

The same problem appears in the conversational-agent literature in modern form. As reviewed in Chapter 2, instruction-tuned LLMs deployed as user simulators are too cooperative, stylistically uniform, and insensitive to user frustration to capture realistic Sim2Real dynamics, and the diversity collapse of post-trained models amplifies across multi-turn interactions. The dual problem holds for the *assistant*: an agent trained against a uniform, polite,

fully-specified user distribution will be brittle when the user it actually meets is exploratory, ambivalent, partially-informed, or frustrated. The structural problem here is not just that users vary; it is that the trajectory of interaction is itself part of the design process, so the assistant's own behavior shapes the distribution of inputs it will subsequently see.

(c) Skill-level heterogeneity. A novice user in a design domain cannot formulate exactly what they want: they do not yet have the vocabulary to express their goal, they do not know what is feasible, and they often do not know what the relevant tradeoffs are until they are confronted with them. An expert user has the opposite problem: they carry rich tacit context—accumulated heuristics, project history, domain conventions—that cannot be put into a prompt and that the assistant must somehow infer from sparse cues. The same AI design assistant must serve both populations, and the populations are not interchangeable.

The recent empirical literature on AI in programming education makes this concrete and shows the stakes. Recent work documents a *widening gap* between novice programmers who use generative AI tools productively and those who are harmed by them: the same tool that accelerates students with strong metacognitive habits leaves struggling students with an illusion of competence and degraded skill formation [194]. The lesson generalizes well beyond programming education. An assistant that pre-supposes an expert user fails the novice; an assistant that pre-supposes a novice insults the expert; and the standard practice of evaluating assistants on fixed-specification benchmarks (where every "user" is effectively an oracle) hides the failure mode entirely.

Summary. The three challenges of epistemic uncertainty over latent goals, distribution shift at inference, and skill heterogeneity are not separable concerns; they interact. An assistant that does not model the user's goals as latent cannot adapt across skill levels; an assistant that does not track the trajectory cannot maintain a coherent model of those goals over the lifetime of a session; and an assistant that does not anticipate skill heterogeneity will train and evaluate on the wrong user distribution to begin with.

From instruction executor to virtual thought partner

The dominant framing of LLM-based design AI is that the assistant takes a fully specified instruction or specification and automates implementation. This is the wrong default for design, in the same structural way that "instruction-tuned LM role-playing a user" was the wrong default for user simulation in Chapter 2. In a wicked problem, there is no fully specified instruction to begin with; arriving at one is the design work.

A small but growing line of work has begun to reframe the assistant accordingly. An early articulation in the mixed-initiative tradition held that the AI system and the user should each take initiative when it is locally most appropriate, rather than treating one as the principal and the other as the executor [96]. Subsequent work has extended this principle into modern human-AI interaction [8, 219]. In the design and ideation literature specifically, recent

work treats the AI as an interlocutor in brainstorming rather than a generator of finished concepts [237]; concurrent work proposes a framework for AI co-creation in which the AI participates in ideation, visual conceptualization, and decision-making rather than executing instructions [151]. In the software-engineering setting, recent work models the user explicitly as a stateful interlocutor whose mental state the agent must track across sessions [274], and other work uses simulated expert perspectives to enrich the ideation phase of research and design rather than to produce final artifacts [150]. In requirements engineering, recent work reframes AI-native development as a collaboration in which requirements emerge from the interaction rather than precede it [1].

The shared shift across these systems is from a stance of "execute fully-specified user instructions" to a stance of "co-construct what the user wants." We refer to this stance, following common usage in the recent HCI and human-AI interaction literature, as treating the assistant as a *virtual thought partner*. The methodological consequences are significant: a thought-partner AI must support clarification dialogue, surface latent tradeoffs, scaffold novice users without infantilizing expert users, and maintain a model of the user's evolving goals across many turns. None of these requirements are addressed by training a model to score well on a fixed-specification benchmark.

User simulation is the critical path

The three structural challenges of Section 5.3 translate directly into requirements on the *user model* that any thought-partner AI is trained and evaluated against. A user simulator that omits epistemic uncertainty, distribution shift, or skill heterogeneity cannot be used to train an agent that handles them: by construction, every signal the agent receives during training assumes those properties away. Faithful user simulation is therefore not a downstream convenience; it is on the critical path.

The methodological apparatus developed in this dissertation is built to deliver exactly the properties a thought-partner training and evaluation loop requires. The substrate choice of Chapter 2—pretrained base models over instruction-tuned models—preserves the breadth of user voices that uniform IT-model simulators systematically collapse, and resolves the diversity-collapse failure mode that recent mechanistic work locates in the assistant axis of post-trained activations [153]. The backstory binding of Chapter 3 produces population-aligned virtual users with controlled demographic composition, enabling the skill-heterogeneity question to be operationalized as a population-design question rather than a prompt-engineering accident. The deep binding of Chapter 4 produces users whose goals and reactions are coherent across many turns and across pragmatically demanding contexts, preserving exactly the resistance, frustration, off-task drift, and ambiguity that real users exhibit and that uniform IT simulators suppress. The action-prediction and counterfactual-decomposition results of Section 5.2 further establish that the same substrate carries from attitudes to consequential decisions under contextual variation—the regime in which mixed-initiative thought-partner interactions actually unfold.

The contrast with the prevailing methodology in the AI-for-design literature is instructive. Even systems explicitly framed as collaborative or thought-partner-style are typically evaluated against either fixed human user studies of modest scale or against synthetic users generated by the same instruction-tuned chat model that the chapters of this dissertation have argued is the wrong substrate. The methodological literature on persona simulation has begun to converge on a similar diagnosis: recent work describes LLM-generated personas as "a promise with a catch" precisely because shallow persona prompting introduces systematic bias [134]; concurrent work addresses persona provenance through explicit population alignment that closely echoes the Anthology framework [99]; other work proposes dynamic evaluation procedures for persona agents that the dissertation's diagnostics complement [203]; and a recent survey maps the field while leaving open exactly the questions of substrate, depth, and trajectory that this dissertation addresses [173].

The bridge to the second motivating regime of Chapter 1 is now direct. Thought-partner AI is, definitionally, an agent that interacts with users over multiple turns under epistemic uncertainty and skill heterogeneity. Training and evaluating it requires synthetic users with exactly the properties that this dissertation's methodology produces.

Digital hardware design as a flagship test case

Stakes. The industrial and societal stakes are large. Chip design sits on the critical path of the broader computing, energy, and AI infrastructure, and the expertise barrier to chip design is high enough that the global designer population is small. DARPA's *Intelligent Design of Electronic Assets* (IDEA) program [60] states the democratization goal in stark form: a hardware compiler that allows "users with no prior design expertise to complete physical design at the most advanced technology nodes," compressing design cycles from years to a single day. The same democratization motive runs through the open-source EDA, agile-hardware, and academic Chipyard/Chisel ecosystems. The premise of all of these efforts is that lowering the expertise barrier to chip design will broaden who can participate—which is the same as saying that the system must serve a heterogeneous user population including novices, students, and engineers from outside the traditional EDA community.

Why hardware design is wicked. At first glance, digital hardware design might appear to sit at the *tame* extreme of engineering. The objectives of power, performance, and area (PPA) are well-defined, the search space is bounded by a target process node and standard-cell library, and success could plausibly be framed as Pareto-optimal design-space exploration over a fixed set of metrics. In practice this idealization fails for three connected reasons. First, the specification is never complete at the outset: additional constraints or adjustments to the specification are discovered through repeated loops between planning and implementation, as functional requirements, area budgets, timing closure targets, and verification coverage goals are revised in light of what each stage of the flow reveals. Second, the implementation flow is multi-stage—synthesis, synthesis optimization, placement

and routing (P&R), and post-route optimization for timing and power—and each stage introduces constraints the prior stage could not have anticipated, so the problem the next stage actually faces is partially constructed by the previous one. Third, real industrial chip design is decomposed spatially and organizationally: individual engineers own sub-blocks or partitions and perform synthesis or P&R locally, and the trajectory of the overall design is shaped by negotiation loops between partition owners as interface contracts, timing budgets, and verification responsibilities are renegotiated. Each of these features is precisely the co-evolution of problem and solution that Section 5.3 identifies as the signature of a wicked problem.

Where the literature stands. Despite this premise, the recent LLM-for-hardware-design literature is overwhelmingly framed as autonomous role simulation rather than user-aware assistance. The dominant pattern is to have one or more LLM agents play the role of an RTL designer, verification engineer, EDA-tool operator, or debugging engineer, and to evaluate the resulting pipeline on capability benchmarks [246, 27, 195, 254, 149, 115, 269, 113, 231, 57, 85]. The user, in these systems, is either absent or treated as an oracle who provides a complete specification at the start. Recent overviews [252, 253] confirm the diagnosis: progress has been concentrated on autonomous capability benchmarks, with comparatively little methodological attention to user-centered evaluation, skill heterogeneity, or multi-turn collaboration with real designers.

One example that hints at this direction is recent work that co-designed a real microprocessor through a multi-turn conversation between a hardware engineer and GPT-4, thereby positioning LLM-based hardware design as a conversational, mixed-initiative activity [28]. While that work established the feasibility of the conversational stance in hardware, the methodology around it has not been developed: there are no systematic user studies of novice and expert designers using such systems, no population-aligned simulated users for training or evaluation, no benchmarks for clarification behavior or for handling partial specifications, and no analysis of how interaction trajectories shape design outcomes. A separate line of recent work, which aligns LLM Verilog generation to HDL engineering practices, addresses one slice of the problem (output alignment to engineer expectations) but treats the user as a fixed, expert population rather than as a heterogeneous one whose goals and skill levels the model must infer [259].

Verification compounds the design loop. A distinguishing feature of digital hardware design relative to most software domains is the central role of *verification*, which is comparable in effort and headcount to implementation in modern industrial flows. The recent LLM-for-hardware literature has begun to develop tooling for this side of the loop separately from implementation—recent work generates SystemVerilog assertions from natural-language design specifications [254], other work provides a benchmark for assertion generation [195], and concurrent work explores a multi-agent generative framework for module-level verification automation [149]—but, as with the implementation-side literature surveyed above,

the user is typically treated as an oracle who supplies a complete and stable specification. The structural difficulty is that all forms of verification—directed and constrained-random simulation, coverage-driven testing, and formal property checking alike—are *approximations* of ground-truth correctness against a specification that is itself still evolving. The act of verifying repeatedly surfaces unexpected constraints, corner cases, and intended behaviors that were not present in the original specification, and each such discovery propagates back into the spec, the RTL, and downstream implementation stages. Verification is therefore not a post-hoc check on a stable design but an active driver of the design loop: each verification iteration reshapes what the design is supposed to be, requiring multiple complementary verification strategies layered across iterations. This is wickedness compounded, and it sharpens the case for AI assistance that treats specification, implementation, and verification as a co-constructed trajectory rather than three independent automation targets.

Open problems

As we have discussed, hardware design bears characteristics of a wicked problem: latent design intent emerges through micro-architectural exploration; long interaction trajectories across spec, RTL, verification, and PPA tradeoffs; and a user population that ranges from undergraduate students learning hardware design for the first time to senior architects steering a tapeout. Yet prior work has not considered the application of LLMs to this domain under that lens. The methodology of LLM-based user simulation introduced in this dissertation is positioned to address this gap as a necessary component of training and evaluating thought-partner AI for wicked design problems.

Several important open problems remain: (i) *Closed-loop evaluation against real designers in real workflows* is the form of evidence that would most directly demonstrate the predictive validity gap identified in the persona-simulation methodology literature [134, 203, 173]. Such evaluation has not been attempted in hardware design and is rare even in software design. (ii) *Skill-stratified simulated populations* require population-alignment procedures (Chapter 3) extended to expertise variables that are not standard demographic categories and not well represented in pretraining corpora; concurrent population-alignment techniques [99] are a natural starting point. (iii) *Multi-turn longitudinal interaction benchmarks* for thought-partner AI are essentially absent in the hardware literature and only partial in the software literature; constructing them requires simulated users who realistically exhibit partial knowledge, hesitation, and revision rather than the polite cooperation that uniform IT-model simulators produce.

The methodology of Chapters 2 to 4, combined with the extensions of Section 5.2, establishes substrate, binding, depth, and trajectory. The applications surface developed here is where that methodology now needs to meet real users in real workflows: design problems in general, and digital hardware design in particular.

Chapter 6

Conclusion

This dissertation has developed a unified framework for using pretrained large language models (LLMs) as faithful simulators of human users—both as synthetic participants in behavioral and social-scientific research and as synthetic interlocutors for the training and evaluation of AI assistant systems. Motivated by ethical, economic, and methodological challenges in traditional human-subject research, and by the increasingly central role of simulated users in conversational AI, this work has introduced practical solutions for selecting the appropriate model substrate, constructing coherent virtual personas via narrative backstories, and applying those personas to behavioral inquiry at scale. The overarching goal has been to establish how appropriately chosen LLMs, when properly grounded in demographic and psychological context, can serve as scalable and responsible instruments for studying human attitudes, identities, and behaviors.

Summary of Contributions

User simulation as language modeling, not assistant alignment. Chapter 2 articulated the central position of this dissertation: that user simulation is a fundamentally distinct modeling problem from assistant alignment, and that the dominant practice of prompting instruction-tuned (IT) models to role-play personas is the wrong default. Pretrained base models, trained via cross-entropy to approximate the empirical conditional distribution of human authorship, are Human–Language Mixture Models (HLMMs) that possess the capabilities required for user simulation. Instruction-tuned Assistant Language Models (ALMs), by contrast, are optimized under mode-seeking objectives that collapse this breadth and shift probability mass toward a narrow, assistant-coded mode. Across five conversational corpora—Reddit, WildChat-1M, LMSYS-Chat-1M, MultiWOZ, and DialOp—and seven open-weight model families, no single metric we report is on its own a definitive measure of user simulation fidelity, and such a definition remains contested in the literature. Our experimental support rests instead on the convergence of independent diagnostics: perplexity, MAUVE, dialogue-act structure, and lexical and semantic diversity all point in the

same direction. We additionally introduced *Tandem Models*, a simple inference-time procedure that pairs an HLMM utterance generator with an ALM supervisor that selects but never rewrites, demonstrating that the strengths of base and instruction-tuned models can be combined without re-collapsing diversity.

Empirical advantages of pretrained models for persona binding. Pretrained base models offer distinctive methodological advantages for persona-based behavioral simulation. Trained on vast, heterogeneous corpora authored by countless individuals, they internalize a rich "mixture of voices" that naturally reflects diverse linguistic styles, social identities, and situational contexts. As illustrated in Figure 3.1, this diversity enables base models to remain sensitive to contextual cues—such as speaker identity or narrative framing—and to shift flexibly across perspectives when conditioned on backstories. Crucially, pretrained models are optimized purely through next-token prediction, which causes them to treat backstories as ordinary textual context and internalize them directly through prefix conditioning. This allows virtual personas to be bound seamlessly by placing the narrative before the query, without interference from external alignment objectives. By contrast, instruction-tuned or chat models impose a single, context-independent "helpful" voice that frequently overrides persona cues with safety, politeness, or normative alignment behaviors. As demonstrated in Section 3.3, such models often collapse the heterogeneity of simulated populations and produce distorted or homogenized opinion distributions. Pretrained base models therefore provide the most faithful substrate for constructing virtual personas, preserving natural variation and enabling controlled manipulations of demographic, contextual, and psychological factors.

Narrative grounding of LLM personas. We developed a pipeline that generates backstories representing realistic life trajectories and demographic attributes, enabling stable and diverse persona formation. The method formalizes persona conditioning as a binding problem between narrative identity and model behavior. Looking ahead, we plan to ground backstory construction more directly in established psychological protocols—most notably McAdams' *Life Story Interview* framework[1] [164], which offers a structured approach for eliciting narrative identity and will guide future extensions of this work.

Population-scale opinion modeling. Using the Pew Research Center's American Trends Panel as reference, Chapter 3 demonstrated that backstory-conditioned models reproduce population-level opinion distributions more faithfully than conventional prompt-based or instruction-tuned baselines. The results show higher inter-item consistency and demographic controllability, and notable gains for under-represented demographic subgroups, establishing the viability of synthetic population simulation as a complement to human-subject surveys.

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

Deep social identity binding. Chapter 4 examined longer and more coherent narratives that sustain personality and group identity across prompts. By generating multi-turn interview transcripts, enforcing internal consistency through LLM-as-a-critic filtering, and scaling to tens of thousands of personas, this approach achieves *deep* binding of social identity. The resulting virtual subjects accurately reproduce partisan asymmetries observed in human attitude data—including hostility gaps in trait evaluation, expectations of democratic backsliding, and exaggerated meta-perceptions of the opposing party—demonstrating that narrative depth and consistency are critical for representing identity-driven social cognition.

Behavioral simulation and counterfactual decomposition of reproducibility. Chapter 5 extends the methodology of Chapters 2 to 4 from attitudinal prediction to action prediction in partisan social-dilemma games. By augmenting deep backstory binding with *Temporal Grounding* and *Consistency Filtering*—supplementary prompting strategies that situate personas in the historical moment of a study and reinforce identity coherence across long behavioral contexts—the resulting personas reproduce human partisan favoritism in Dictator and Trust Game replications spanning more than a decade. The same setting enables a methodological contribution that human studies cannot supply on their own: a $2 \times 2 \times 2$ factorial counterfactual decomposition of year, framing, and recruitment-population effects, which isolates the contextual sources of variation that any single human study necessarily bundles together.

User simulation as the critical path for thought-partner AI in design problems. Chapter 5 reframes the methodology as a critical enabler for human-centered AI in *design problems*—settings characterized in the design-research tradition as *wicked*, where three structural challenges (latent user goals, interaction-trajectory distribution shift, and skill-level heterogeneity) demand user simulators with the substrate, binding, depth, and trajectory properties developed in the preceding chapters. Digital hardware design is identified as the flagship application target where this methodology is most needed and least developed.

Impact and Broader Significance

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

The broader implication of this work is that user modeling deserves a dedicated methodology, sensitive to the target dialogue regime and distinct from the objectives, failure modes, and evaluation criteria of assistant training. Solutions developed for assistant alignment do

not automatically transfer to user simulation; in many cases, they actively work against it. The HLMM/ALM distinction, the simulation desiderata laid out in Chapter 2, and the backstory-based persona conditioning developed in Chapters 3 and 4 together provide a conceptual scaffold for future work that takes faithful user simulation as a research target in its own right—both for the social-scientific applications studied here and for the rapidly growing use of synthetic interlocutors in conversational AI training and evaluation pipelines.

Limitations and Responsible Use

While the results indicate strong alignment with empirical data, several limitations qualify the conclusions of this dissertation and warrant explicit discussion.

Inherited biases in pretraining corpora. The pretraining corpora on which LLMs are trained are themselves large, poorly characterized, and subject to their own distributional biases [23]. The user models that emerge from pretraining are not neutral: they reflect the demographics and discourse norms captured in web-scale data. This does not undermine the case for base models over instruction-tuned models as user simulators; instruction-tuned models introduce a further, qualitatively different distributional shift on top of whatever biases exist in pretraining. But it does mean that synthetic populations of virtual personas inherit the biases of the underlying base model. Consistent with this, our results in Chapters 3 and 4 show systematically weaker fidelity for older and conservative-leaning subgroups, where web-scale data is plausibly less representative. A natural follow-on question—which we see as a central open problem for the user simulation community—is which user populations are over- or under-represented in the implicit speaker mixture of LLMs after pretraining, and whether targeted data curation or lightweight fine-tuning can address those gaps without reintroducing the diversity collapse that instruction-tuning produces.

Convergence of metrics, not a single oracle. No single metric we report is on its own a definitive measure of user simulation fidelity, and such a definition remains contested in the literature. Our experimental support rests instead on the convergence of independent diagnostics—perplexity and MAUVE for distributional fit, dialogue-act N-gram divergence for pragmatic structure, lexical and semantic diversity metrics for coverage, Wasserstein distance to empirical opinion distributions for representativeness, Frobenius norm and Cronbach's α for internal consistency, and Cohen's d for effect-size fidelity. We invite future work to expand, affirm, or challenge these methodologies.

Descriptive, not causal. The simulations developed in this dissertation are descriptive rather than causal, and should be interpreted as pre-experimental approximations. Responsible use requires transparent reporting of model versions, prompts, demographic matching criteria, and sampling configurations, followed by confirmatory human validation for findings that are intended to inform real-world decision-making.

Potential for harmful outputs. Pretrained base models are substantially less constrained than instruction-tuned models in their outputs and may produce harmful, offensive, or factually incorrect content more readily. Any deployment of base-model user simulators to end-user applications requires appropriate content filtering and safety review.

Scope of evaluation. The case studies in this dissertation focused on English-language corpora and on U.S.-centric political and demographic surveys (Pew Research Center, Stanford American Voices Project). Extensions to multilingual settings, non-Western political and cultural contexts, domain-specific dialogue, and synchronous spoken interaction remain open empirical questions.

Future Directions

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

1. Systematic evaluation of open-ended responses. Chapter 2 already evaluates open-ended response generation in the form of next-turn utterance prediction in multi-turn dialogue, using perplexity, MAUVE, dialogue-act N-gram divergence, and lexical and semantic diversity to compare model and human distributions at both the utterance and corpus levels. These diagnostics establish that pretrained base models are faithful estimators of human turn-level continuations. Much of behavioral and social-scientific inquiry, however, depends on *longer-form* open-ended responses, such as structured interviews, focus-group discussions, deliberative argumentation, autobiographical writing, and reflective self-reports. These aspects of analysis are not adequately captured by next-utterance metrics alone. Developing systematic measures of narrative coherence, emotional tone, self-consistency, and pragmatic fidelity for such longer free-form responses is a natural extension of the corpus-level evaluations developed here.

2. From snapshots to longitudinal trajectories. The methodology developed in this dissertation conditions on backstories that fix a persona at a single moment in time. Many of the most consequential questions in behavioral and intervention science—health behavior change, attitude formation around emerging technologies, the evolution of user trust across repeated sessions with an AI assistant—are inherently longitudinal, concerning how beliefs and behavior evolve over weeks, months, or years in response to information, social influence, and experience. Extending backstory-conditioned simulation to these regimes requires modeling personas whose narrative state itself evolves across interactions, generalizing *Temporal Grounding* from a single anchor to a multi-time-step setting, and validating trajectory fidelity against published human cohort data and intervention meta-analyses.

3. Closed-loop evaluation of thought-partner AI design assistants. Section 5.3 positions user simulation as the critical path for training and evaluating AI design assistants in wicked design problems, with digital hardware as the flagship target. The near-term directions developed there collectively constitute the empirical follow-on to this dissertation: closed-loop evaluation against population-aligned simulated cohorts on multi-turn agent benchmarks, skill-stratified populations for hardware design, and clarification-behavior benchmarks. The Tandem architecture of Chapter 2 is a natural starting point.

Together, these directions extend the user-simulation methodology developed in this dissertation along the axes where it is most consequential and least developed: longer-form qualitative responses, longitudinal trajectories of belief and behavior change, and closed-loop application to AI design assistants in real human-centered workflows. They position backstory-conditioned LLM personas as instruments not only for predicting responses but also for exploring how beliefs, behaviors, and design decisions emerge, evolve, and respond to context—a step toward richer, ethically grounded, and methodologically integrated computational social science and human-centered AI.

Bibliography

[1] Mateen Ahmed Abbasi et al. "Reconsidering Requirements Engineering: Human–AI Collaboration in AI-Native Software Development". In: *Proceedings of the 51st Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. Vol. 16081. Lecture Notes in Computer Science. Springer, 2025, pp. 164–180. URL: https://arxiv.org/abs/2510.04380.

[2] Marwa Abdulhai et al. *Consistently Simulating Human Personas with Multi-Turn Reinforcement Learning*. 2025. arXiv: 2511.00222 [cs.CL].

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

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

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

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

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

[8] Saleema Amershi et al. "Guidelines for Human-AI Interaction". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019. DOI: 10.1145/3290605.3300233.

[9] Jacob Andreas. "Language models as agent models". In: *arXiv preprint arXiv:2212.01681* (2022).

[10] Jacy Reese Anthis et al. "Llm social simulations are a promising research method". In: *arXiv preprint arXiv:2504.02234* (2025).

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

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

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

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

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

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

[17] Avinash Baidya, Kamalika Das, and Xiang Gao. "The Behavior Gap: Evaluating Zero-shot LLM Agents in Complex Task-Oriented Dialogs". In: *Findings of the Association for Computational Linguistics: ACL 2025*. 2025, pp. 23455–23472.

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

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

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

[21] Christopher A. Bail. "Can Generative AI improve social science?" In: *Proceedings of the National Academy of Sciences* 121.21 (2024), e2314021121. DOI: 10.1073/pnas.2314021121. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2314021121. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2314021121.

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

[23] Emily M Bender et al. "On the dangers of stochastic parrots: Can language models be too big?" In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021, pp. 610–623.

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

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

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

[27] Jason Blocklove et al. "Automatically Improving LLM-based Verilog Generation with EDA Tool Feedback". In: *ACM Transactions on Design Automation of Electronic Systems* (2025). DOI: 10.1145/3723876. URL: https://dl.acm.org/doi/10.1145/3723876.

[28] Jason Blocklove et al. "Chip-Chat: Challenges and Opportunities in Conversational Hardware Design". In: *2023 ACM/IEEE 5th Workshop on Machine Learning for CAD (MLCAD)*. 2023. URL: https://arxiv.org/abs/2305.13243.

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

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

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

[32] Michael E. Bratman. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press, 1987.

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

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

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

[36] Richard Buchanan. "Wicked Problems in Design Thinking". In: *Design Issues* 8.2 (1992), pp. 5–21. DOI: 10.2307/1511637.

[37] Paweł Budzianowski et al. "MultiWOZ – A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2019, pp. 5016–5026.

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

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

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

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

[42] Jonathan P. Chang et al. "ConvoKit: A Toolkit for the Analysis of Conversations". In: *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 2020, pp. 57–60.

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

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

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

[46] Myra Cheng et al. "Sycophantic AI decreases prosocial intentions and promotes dependence". In: *Science* 391.6792 (2026), eaec8352.

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

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

[49] Paul Christiano et al. "Deep Reinforcement Learning from Human Preferences". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, pp. 4299–4307.

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

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

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

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

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

[55] MJ Crockett and Lisa Messeri. "AI Surrogates and illusions of generalizability in cognitive science". In: *Trends in Cognitive Sciences* (2025).

[56] Nigel Cross. "Designerly Ways of Knowing". In: *Design Studies* 3.4 (1982), pp. 221–227. DOI: 10.1016/0142-694X(82)90040-0.

[57] Fan Cui et al. *HWE-Bench: Benchmarking LLM Agents on Real-World Hardware Bug Repair Tasks*. 2026. arXiv: 2604.14709 [cs.AR]. URL: https://arxiv.org/abs/2604.14709.

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

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

[60] Defense Advanced Research Projects Agency. *IDEA: Intelligent Design of Electronic Assets (Program Page)*. DARPA Microsystems Technology Office. 2018. URL: https://www.darpa.mil/research/programs/intelligent-design-of-electronic-assets.

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

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

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

[64] Kees Dorst. "The Core of 'Design Thinking' and its Application". In: *Design Studies* 32.6 (2011), pp. 521–532. DOI: 10.1016/j.destud.2011.07.006.

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

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

[67] Bradley Emi and Max Spero. *Technical Report on the Pangram AI-Generated Text Classifier*. 2024. arXiv: 2402.14873 [cs.CL]. URL: https://arxiv.org/abs/2402.14873.

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

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

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

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

[72] Michael C. Frank and Noah D. Goodman. "Predicting Pragmatic Reasoning in Language Games". In: *Science* 336.6084 (2012), p. 998. DOI: 10.1126/science.1218633.

[73] Dan Friedman and Adji Bousso Dieng. "The Vendi Score: A Diversity Evaluation Metric for Machine Learning". In: *Transactions on Machine Learning Research* (2023). URL: https://openreview.net/forum?id=g97OHbQyk1.

[74] Kanishk Gandhi, Agam Bhatia, and Noah D. Goodman. *Learning to Simulate Human Dialogue*. 2026. arXiv: 2601.04436 [cs.CL].

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

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

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

[78] Kallirroi Georgila, James Henderson, and Oliver Lemon. "User Simulation for Spoken Dialogue Systems: Learning and Evaluation". In: *Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH-ICSLP)*. Pittsburgh, USA, 2006, pp. 1065–1068. DOI: 10.21437/Interspeech.2006-160.

[79] Dongyoung Go et al. "Aligning Language Models with Preferences through f-Divergence Minimization". In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 11546–11583. URL: https://proceedings.mlr.press/v202/go23a.html.

[80] Noah D. Goodman and Michael C. Frank. "Pragmatic Language Interpretation as Probabilistic Inference". In: *Trends in Cognitive Sciences* 20.11 (2016), pp. 818–829. DOI: 10.1016/j.tics.2016.08.005.

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

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

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

[84] Daya Guo et al. "Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning". In: *arXiv preprint arXiv:2501.12948* (2025).

[85] Raghav Gupta et al. *ArchAgent: Agentic AI-driven Computer Architecture Discovery*. 2026. arXiv: 2602.22425 [cs.AI]. URL: https://arxiv.org/abs/2602.22425.

[86] Dylan Hadfield-Menell et al. "Cooperative Inverse Reinforcement Learning". In: *Advances in Neural Information Processing Systems 29 (NeurIPS)*. 2016. URL: https://arxiv.org/abs/1606.03137.

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

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

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

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

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

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

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

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

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

[96] Eric Horvitz. "Principles of Mixed-Initiative User Interfaces". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1999, pp. 159–166. DOI: 10.1145/302979.303030.

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

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

[99] Zhengyu Hu et al. *Population-Aligned Persona Generation for LLM-based Social Simulation*. 2025. arXiv: 2509.10127 [cs.CY]. URL: https://arxiv.org/abs/2509.10127.

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

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

[102] Jonathan Ivey et al. "Real or robotic? Assessing whether LLMs accurately simulate qualities of human responses in dialogue". In: *arXiv preprint arXiv:2409.08330* (2024).

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

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

[105] Aaron Jaech et al. "OpenAI o1 system card". In: *arXiv preprint arXiv:2412.16720* (2024).

[106] Frederick Jelinek et al. "Perplexity—a Measure of the Difficulty of Speech Recognition Tasks". In: *The Journal of the Acoustical Society of America* 62.S1 (1977), S63–S63. DOI: 10.1121/1.2016299.

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

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

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

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

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

[112] Liwei Jiang et al. *Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond)*. 2025. arXiv: 2510.22954 [cs.CL].

[113] Pengwei Jin, Di Huang, et al. *RealBench: Benchmarking Verilog Generation Models with Real-World IP-Level Tasks*. 2025. arXiv: 2507.16200 [cs.AR]. URL: https://arxiv.org/abs/2507.16200.

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

[115] Rohit Kanagal. *LLM-Powered EDA Log Analysis for Effective Design Debugging*. Tech. rep. UCB/EECS-2025-48. University of California, Berkeley, 2025. URL: https://www2.eecs.berkeley.edu/Pubs/TechRpts/2025/EECS-2025-48.pdf.

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

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

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

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

[120] Taaha Kazi et al. *Large Language Models as User-Agents for Evaluating Task-Oriented-Dialogue Systems*. 2024. arXiv: 2411.09972 [cs.CL].

[121] Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. "Will I Sound Like Me? Improving Persona Consistency in Dialogues through Pragmatic Self-Consciousness". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 904–916. DOI: 10.18653/v1/2020.emnlp-main.65. URL: https://aclanthology.org/2020.emnlp-main.65.

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

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

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

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

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

[127] Florian Kreyssig et al. "Neural User Simulation for Corpus-Based Policy Optimisation for Spoken Dialogue Systems". In: *arXiv preprint arXiv:1805.06966* (2018).

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

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

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

[131] Willem J. M. Levelt. *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press, 1989.

[132] Esther Levin, Roberto Pieraccini, Wieland Eckert, et al. "A stochastic model of human-machine interaction for learning dialog strategies". In: *IEEE Transactions on speech and audio processing* 8.1 (2000), pp. 11–23.

[133] Mike Lewis et al. "Deal or No Deal? End-to-End Learning of Negotiation Dialogues". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 2443–2453. DOI: 10.18653/v1/D17-1259.

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

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

[136] Jiwei Li et al. "A Diversity-Promoting Objective Function for Neural Conversation Models". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, 2016, pp. 110–119. DOI: 10.18653/v1/N16-1014.

[137] Jiwei Li et al. "A Persona-Based Neural Conversation Model". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 994–1003. DOI: 10.18653/v1/P16-1094.

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

[139] Kenneth Li et al. "Measuring and Controlling Instruction (In)Stability in Language Model Dialogs". In: *Proceedings of the First Conference on Language Modeling*. arXiv:2402.10962. 2024.

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

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

[142] Bill Yuchen Lin et al. *The Unlocking Spell on Base LLMs: Rethinking Alignment via In-Context Learning*. 2024. arXiv: 2312.01552 [cs.CL].

[143] Chin-Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[144] Hsien-Chin Lin et al. "GenTUS: Simulating User Behaviour and Language in Task-Oriented Dialogues with Generative Transformers". In: *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Edinburgh, UK: Association for Computational Linguistics, 2022, pp. 294–309. DOI: 10.18653/v1/2022.sigdial-1.28.

[145] Jessy Lin et al. "Decision-oriented dialogue for human-AI collaboration". In: *Transactions of the Association for Computational Linguistics* 12 (2024), pp. 892–911.

[146] Jianhua Lin. "Divergence Measures Based on the Shannon Entropy". In: *IEEE Transactions on Information Theory* 37.1 (1991), pp. 145–151. DOI: 10.1109/18.61115.

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

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

[149] Wenbo Liu et al. *A Multi-Agent Generative AI Framework for IC Module-Level Verification Automation*. 2025. arXiv: 2507.21694 [cs.AR]. URL: https://arxiv.org/abs/2507.21694.

[150] Yiren Liu et al. "PersonaFlow: Designing LLM-Simulated Expert Perspectives for Enhanced Research Ideation". In: *Proceedings of the ACM Designing Interactive Systems Conference*. DIS 2025. 2025. DOI: 10.1145/3715336.3735789. URL: https://dl.acm.org/doi/10.1145/3715336.3735789.

[151] Zhangqi Liu. *Human–AI Co-Creation: A Framework for Collaborative Design in Intelligent Systems*. 2025. arXiv: 2507.17774 [cs.HC]. URL: https://arxiv.org/abs/2507.17774.

[152] Kathleen Locklear. "Wicked Problems: A Novel Approach Using Artificial Intelligence and Scenarios". In: *Journal of Leadership & Organizational Studies* 32.3 (2025), pp. 230–248. DOI: 10.1177/15480518251330728.

[153] Christina Lu et al. *The Assistant Axis: Situating and Stabilizing the Default Persona of Language Models*. 2026. arXiv: 2601.10387 [cs.CL].

[154] Alex Lyman et al. "Balancing large language model alignment and algorithmic fidelity in social science research". In: *Sociological Methods and Research* 54 (3 2025).

[155] Arunima Maitra, Dorothea French, and Katharina von der Wense. "Dialogue Acts as a Lens on Human–LLM Interaction: Analyzing Conversational Norms in Model-Generated Responses". In: *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP)*. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 317–325. DOI: 10.18653/v1/2025.hcinlp-1.25.

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

[157] Pierre-Emmanuel Mazaré et al. "Training Millions of Personalized Dialogue Agents". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 2775–2779. DOI: 10.18653/v1/D18-1298.

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

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

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

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

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

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

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

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

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

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

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

[169] Suhong Moon et al. "Identity, Cooperation and Framing Effects within Groups of Real and Simulated Humans". In: *arXiv preprint arXiv:2601.16355* (2026).

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

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

[172] Tarek Naous et al. "Flipping the Dialogue: Training and Evaluating User Language Models". In: *The Fourteenth International Conference on Learning Representations*. arXiv:2510.06552. 2026.

[173] Bo Ni et al. "A Survey on LLM-based Conversational User Simulation". In: *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2026, pp. 4266–4301. DOI: 10.18653/v1/2026.eacl-long.200. URL: https://aclanthology.org/2026.eacl-long.200/.

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

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

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

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

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

[179] Vishakh Padmakumar and He He. "Does Writing with Language Models Reduce Content Diversity?" In: *Proceedings of the 12th International Conference on Learning Representations*. arXiv:2309.05196. 2024.

[180] Kishore Papineni et al. "BLEU: A Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002, pp. 311–318. DOI: 10.3115/1073083.1073135.

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

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

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

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

[185] Joon Sung Park et al. "Social Simulacra: Creating Populated Prototypes for Social Computing Systems". In: *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. UIST '22. Bend, OR, USA: Association for Computing Machinery, 2022. ISBN: 9781450393201. DOI: 10.1145/3526113.3545616. URL: https://doi.org/10.1145/3526113.3545616.

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

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

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

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

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

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

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

[193] Krishna Pillutla et al. "MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers". In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 4816–4828.

[194] James Prather et al. "The Widening Gap: The Benefits and Harms of Generative AI for Novice Programmers". In: *Proceedings of the 2024 ACM Conference on International Computing Education Research (ICER)*. 2024. URL: https://arxiv.org/abs/2405.17739.

[195] Venkata Pulavarthi et al. "AssertionBench: A Benchmark to Evaluate Large-Language Models for Assertion Generation". In: *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics, 2025. URL: https://aclanthology.org/2025.findings-naacl.449/.

[196] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: *OpenAI Technical Report* (2019). URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

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

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

[199] Horst W. J. Rittel and Melvin M. Webber. "Dilemmas in a General Theory of Planning". In: *Policy Sciences* 4.2 (1973), pp. 155–169. DOI: 10.1007/BF01405730.

[200] Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. "A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning". In: *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2011, pp. 627–635. URL: https://arxiv.org/abs/1011.0686.

[201] Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking, 2019.

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

[203] Vinay Samuel et al. "PersonaGym: Evaluating Persona Agents and LLMs". In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics, 2025. URL: https://aclanthology.org/2025.findings-emnlp.368/.

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

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

[206] Jost Schatzmann et al. "A Survey of Statistical User Simulation Techniques for Reinforcement-Learning of Dialogue Management Strategies". In: *The Knowledge Engineering Review* 21.2 (2006), pp. 97–126. DOI: 10.1017/S0269888906000944.

[207] Jost Schatzmann et al. "Agenda-Based User Simulation for Bootstrapping a POMDP Dialogue System". In: *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*. Rochester, New York: Association for Computational Linguistics, 2007, pp. 149–152.

[208] Donald A. Schön. *The Reflective Practitioner: How Professionals Think in Action*. New York: Basic Books, 1983.

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

[210] Ivan Sekulić et al. "Reliable LLM-Based User Simulator for Task-Oriented Dialogue Systems". In: *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024)*. Malta: Association for Computational Linguistics, 2024, pp. 28–40. DOI: 10.18653/v1/2024.scichat-1.3.

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

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

[213] Preethi Seshadri et al. *Lost in Simulation: LLM-Simulated Users are Unreliable Proxies for Human Users in Agentic Evaluations*. 2026. arXiv: 2601.17087 [cs.HC].

[214] Omar Shaikh et al. "Grounding Gaps in Language Model Generations". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 6279–6296. DOI: 10.18653/v1/2024.naacl-long.348.

[215] Omar Shaikh et al. "Navigating rifts in human-llm grounding: Study and benchmark". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2025, pp. 20832–20847.

[216] Murray Shanahan, Kyle McDonell, and Laria Reynolds. "Role Play with Large Language Models". In: *Nature* 623 (2023), pp. 493–498. DOI: 10.1038/s41586-023-06647-8.

[217] Mrinank Sharma et al. "Towards Understanding Sycophancy in Language Models". In: *Proceedings of the 12th International Conference on Learning Representations*. arXiv:2310.13548. 2024.

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

[219] Ben Shneiderman. "Human-Centered Artificial Intelligence: Reliable, Safe and Trustworthy". In: *International Journal of Human-Computer Interaction* 36.6 (2020), pp. 495–504. DOI: 10.1080/10447318.2020.1741118.

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

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

[222] Herbert A. Simon. *The Sciences of the Artificial*. Cambridge, MA: MIT Press, 1969.

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

[224] Nisan Stiennon et al. "Learning to Summarize from Human Feedback". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 3008–3021.

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

[226] Andreas Stolcke et al. "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech". In: *Computational Linguistics* 26.3 (2000), pp. 339–373. DOI: 10.1162/089120100561737.

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

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

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

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

[231] Kimia Tasnia et al. *VeriOpt: PPA-Aware High-Quality Verilog Generation via Multi-Role LLMs*. 2025. arXiv: 2507.14776 [cs.AR]. URL: https://arxiv.org/abs/2507.14776.

[232] Silvia Terragni et al. "In-Context Learning User Simulators for Task-Oriented Dialog Systems". In: *arXiv preprint arXiv:2306.00774* (2023).

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

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

[235] Yu-Min Tseng et al. "Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization". In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 3769–3791. DOI: 10.18653/v1/2024.findings-emnlp.218.

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

[237] Wen-Fan Wang et al. "AIdeation: Designing a Human–AI Collaborative Ideation System for Concept Designers". In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 2025. DOI: 10.1145/3706598.3714148. URL: https://arxiv.org/abs/2502.14747.

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

[239] Zekun Moore Wang et al. "RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models". In: *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 14181–14199. DOI: 10.18653/v1/2024.findings-acl.843.

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

[241] Peter West and Christopher Potts. "Base Models Beat Aligned Models at Randomness and Creativity". In: *Second Conference on Language Modeling*. 2025. URL: https://openreview.net/forum?id=vqN8uom4A1.

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

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

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

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

[246] Haoyang Wu et al. "Divergent Thoughts toward One Goal: LLM-based Multi-Agent Collaboration System for Electronic Design Automation". In: *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Introduces EDAid. Association for Computational Linguistics, 2025. URL: https://aclanthology.org/2025.naacl-long.83/.

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

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

[249] Shirley Wu et al. *HumanLM: Simulating Users with State Alignment Beats Response Imitation*. 2026. arXiv: 2603.03303 [cs.CL].

[250] Chengxing Xie et al. "Can Large Language Model Agents Simulate Human Trust Behavior?" In: *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*. 2024. URL: https://openreview.net/forum?id=CeOwahuQic.

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

[252] Kangwei Xu et al. "Large Language Models (LLMs) for Electronic Design Automation (EDA)". In: *2025 IEEE International System-on-Chip Conference (SOCC)*. arXiv: 2508.20030. 2025. URL: https://arxiv.org/abs/2508.20030.

[253] Qiang Xu et al. "Revolution or Hype? Seeking the Limits of Large Models in Hardware Design". In: *2025 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. Invited paper; arXiv: 2509.04905. 2025. URL: https://arxiv.org/abs/2509.04905.

[254] Zichao Yan et al. "AssertLLM: Generating Hardware Verification Assertions from Design Specifications via Large Language Models". In: *Proceedings of the Asia and South Pacific Design Automation Conference*. 2025. DOI: 10.1145/3658617.3697756. URL: https://dl.acm.org/doi/10.1145/3658617.3697756.

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

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

[257] Kevin Yang et al. "DOC: Improving Long Story Coherence With Detailed Outline Control". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 3378–3465. DOI: 10.18653/v1/2023.acl-long.190. URL: https://aclanthology.org/2023.acl-long.190.

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

[259] Yiyao Yang et al. *HaVen: Hallucination-Mitigated LLM for Verilog Code Generation Aligned with HDL Engineers*. Accepted at Design, Automation and Test in Europe (DATE) 2025. 2025. URL: https://github.com/Intelligent-Computing-Research-Group/HaVen.

[260] Shunyu Yao et al. "*tau*-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains". In: *arXiv preprint arXiv:2406.12045* (2024).

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

[262] Kefan Yu et al. "The Pragmatic Mind of Machines: Tracing the Emergence of Pragmatic Competence in Large Language Models". In: *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2026, pp. 192–213.

[263] Longfei Yun et al. *The Price of Format: Diversity Collapse in LLMs*. 2025. arXiv: 2505.18949 [cs.CL].

[264] Feng Zhang et al. *UserLM-R1: Modeling Human Reasoning in User Language Models with Multi-Reward Reinforcement Learning*. 2026. arXiv: 2601.09215 [cs.CL]. URL: https://arxiv.org/abs/2601.09215.

[265] Saizheng Zhang et al. "Personalizing Dialogue Agents: I have a dog, do you have pets too?" In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 2204–2213. DOI: 10.18653/v1/P18-1205.

[266] Tianyi Zhang et al. "BERTScore: Evaluating Text Generation with BERT". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

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

[268] Wenting Zhao et al. "WildChat: 1M ChatGPT Interaction Logs in the Wild". In: *The Twelfth International Conference on Learning Representations*. arXiv:2405.01470. 2024.

[269] Yu Zhao et al. "Enhancing LLM Performance on Hardware Design through Reinforcement Learning". In: *Proceedings of the IEEE International Symposium on Circuits and Systems*. 2025. URL: https://ece.k-state.edu/research/hardware-security/papers/ISCAS2025_RLHWGen.pdf.

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

[271] Lianmin Zheng et al. "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena". In: *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. Vol. 36. 2023, pp. 46595–46623. URL: https://openreview.net/forum?id=uccHPGDlao.

[272] Lianmin Zheng et al. *LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset*. 2023. arXiv: 2309.11998 [cs.CL].

[273] Xuhui Zhou et al. *Mind the Sim2Real Gap in User Simulation for Agentic Tasks*. 2026. arXiv: 2603.11245 [cs.CL].

[274] Xuhui Zhou et al. *TOM-SWE: User Mental Modeling For Software Engineering Agents*. 2025. arXiv: 2510.21903 [cs.SE]. URL: https://arxiv.org/abs/2510.21903.

[275] Alan Zhu et al. *BARE: Leveraging Base Language Models for Few-Shot Synthetic Data Generation*. 2025. arXiv: 2502.01697 [cs.CL].

[276] Qi Zhu et al. "ConvLab-2: An Open-Source Toolkit for Building, Evaluating, and Diagnosing Dialogue Systems". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2020, pp. 71–78.

[277] Yaoming Zhu et al. "Texygen: A Benchmarking Platform for Text Generation Models". In: *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2018, pp. 1097–1100. DOI: 10.1145/3209978.3210080.

[278] Ziyi Zhu et al. *DIAL: Direct Iterative Adversarial Learning for Realistic Multi-Turn Dialogue Simulation*. 2026. arXiv: 2512.20773 [cs.CL].

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

Appendix A

Role Playing is Not Language Modeling: Additional Materials

A.1 Experimental Details

Next-Utterance Prediction Protocol

All experiments instantiate user simulation as next-utterance prediction. Given a dialogue prefix $c = (x_1, \ldots, x_{t-1})$ and target user u, a simulator generates a candidate continuation $\hat{x}_t$ for the same speaker and turn at which the human continuation x_t was observed. This setup keeps the discourse state, partner behavior, and target speaker fixed while varying only the simulator's next user utterance.

For Reddit, WildChat-1M, and LMSYS-Chat-1M, we sample target user turns from multi-turn conversations after the opening turns. For task-oriented corpora, the evaluation expands conversations into user-turn prediction examples so that later-turn pragmatic structure is represented. Each retained context is paired with the aligned human continuation and with multiple simulator continuations. Unless otherwise stated, each model produces eight continuations per context.

Corpus Preparation

Reddit. The Reddit experiments use the ConvoKit Reddit corpus. Threads are rendered with the thread title followed by speaker-labeled comments. Deleted or removed turns are excluded. The sampled target turn must belong to a two-participant exchange, and consecutive runs from the same speaker are not used as new target turns. The resulting task is open-domain human-human continuation in informal online discussion.

WildChat-1M. WildChat-1M conversations are filtered to English, non-toxic, non-duplicate multi-turn human-assistant conversations. Target turns are restricted to human user turns. Contexts are rendered as alternating `User:` and `Assistant:` turns, and the target is the next user utterance.

LMSYS-Chat-1M. LMSYS-Chat-1M is treated as a second real-world human-LLM chat corpus. English multi-turn conversations are retained, redacted conversations are dropped, and exact duplicate records are removed. As in WildChat-1M, target turns are user turns and contexts are rendered with `User:` and `Assistant:` labels.

DialOp. DialOp is converted from itinerary-planning action logs into alternating user-assistant messages. Player 0 is treated as the user and player 1 as the assistant. Private itinerary goals and preferences are preserved when available and are shown to the simulator as user-private information. The target is a user utterance in a goal-directed planning dialogue.

MultiWOZ. MultiWOZ 2.1 is used for task-oriented dialogue-act analysis. We use labeled dialogues with turn-level dialogue-act annotations. User goals are retained in the context when available, and target user turns are expanded into examples for generation and subsequent dialogue-act labeling.

Model Families

We evaluate three simulator families. Direct HLMM systems are base/pretrained language models sampled directly from the dialogue context. Direct ALM systems are instruction-tuned variants prompted to continue the conversation as the human user. Tandem systems use an HLMM as a candidate speaker and an ALM as a selector. Matched base/instruct comparisons isolate the effect of instruction tuning within a model family.

Direct Generation

Base models receive a plain-text completion prompt consisting of the rendered dialogue context followed by the target speaker label. Generation stops at dataset-specific strings that mark the beginning of a new turn, such as a later `User:`, `Assistant:`, Reddit username marker, or thread boundary. Instruction-tuned models receive a system/user chat prompt asking them to continue as the user and to output only the next user utterance.

Generated text is normalized before evaluation. Leading speaker labels are removed, generated reasoning blocks are stripped when present, and outputs that continue into later role-labeled turns are truncated to the first utterance. Empty generations are discarded.

Tandem Generation

The tandem simulator separates candidate production from candidate selection. For a dialogue context c, the HLMM speaker samples a candidate set $\mathcal{U}(c) = \{u^{(1)}, \ldots, u^{(N)}\}$. The ALM selector reads the same dialogue context and the candidate set, reasons about communicative fit, and selects one candidate:

$$\hat{u} = \arg\max_{u \in \mathcal{U}(c)} s_{\mathrm{ALM}}(u, c).$$

The selector is constrained to return a candidate verbatim. This prevents the ALM from rewriting the surface form into its own assistant-like style while still allowing deliberative selection for local relevance and consistency.

We evaluate two tandem variants. The *User Agent* variant first produces an explicit belief-desire-intention plan for the next turn, then scores candidates, then emits the selected candidate verbatim. The *Critique Only* variant omits the explicit planning step and selects the first candidate that satisfies all rubric criteria.

Sampling and Splits

Generation examples are shuffled with a fixed random seed before sampling. The default evaluation sample contains 1,024 contexts for Reddit, WildChat-1M, and LMSYS-Chat-1M; task-oriented corpora are expanded from sampled dialogues into their eligible user turns.

Output Validation

All evaluation rows require a nonempty simulator continuation and an aligned human continuation. Diversity metrics are computed after text preprocessing and per-context deduplication.

A.2 Prompt Templates

This appendix gives the prompt templates used for generation and tandem selection. Braced fields denote dataset-specific values inserted at evaluation time.

Direct User-Simulation Prompts

Reddit instruction-tuned simulator: system prompt

```
You are simulating a Reddit user engaging with conversations with different users.
    You are to simulate the user u/{user_id}. You will be given past conversations
    that this user has engaged with, and based on that, continue the conversation
    naturally as the user. Only answer with the next user utterance and no other
    text.
```

Reddit instruction-tuned simulator: user prompt

```
Here is the conversation context, consisting of Reddit thread(s):

###THREADS START
```

```
{context}

###THREADS END

Generate a comment that is in line with the final thread conversation for user u/{
    user_id}. Continue the conversation naturally as the user.Only answer with the
    comment and no other text.
```

Human-LLM chat instruction-tuned simulator

```
System:
Continue the conversation naturally as the user. Only answer with the next user
    utterance and no other text.

User:
{context}

Continue the conversation naturally as the user. Only answer with the next user
    utterance and no other text.
```

DialOp instruction-tuned simulator

```
System:
You are simulating a user engaging with conversations with a different user on
    planning a travel itinerary. You are to simulate the user marked as 'User'. You
    will be given past conversations that this user has engaged with, and based on
    that, you must generate a comment that is like the given user.

User:
{goal_block}
Here is the conversation context so far:

###CONVERSATION START

{context}

###CONVERSATION END

Generate an utterance that is in line with the final thread conversation for the
    User. Only answer with the utterance and no other text.
```

MultiWOZ instruction-tuned simulator

```
System:
You are a simulating a user engaging in a task-oriented dialogue. Continue the
    conversation as the USER. Only answer with the next USER utterance and no other
    text.

User:
Dialogue context:
{context}

Continue the conversation as the USER. Only answer with the next USER utterance and
     no other text.
```

Pretrained Base-Model Prompts

Pretrained base models were prompted as continuation models rather than as instruction-following chat assistants. The prompt is the serialized dialogue context itself, with no system message or natural-language instruction appended. Generation begins immediately after the final target-speaker prefix. The same completion-style prompt is used for direct base-model generation and for the base utterance generator inside tandem runs.

Reddit pretrained base simulator

```
[Title]: {thread_title}

[Original Poster (OP) u/{op_user_id}]: {op_text}

[u/{speaker_id}]: {utterance}

...

[u/{target_user_id}]:
```

When the target speaker is the original poster, the final Reddit prefix uses the same original-poster form as the context, i.e., `[Original Poster (OP) u/{target_user_id}]:`.

Human-LLM chat pretrained base simulator

```
User: {user_utterance}

Assistant: {assistant_utterance}

...
```

```
User:
```

This continuation template is used for both WildChat-1M and LMSYS-Chat-1M.

DialOp pretrained base simulator

```
User goals:
{goal_bullets}

Conversation so far:

User: {user_utterance}

Assistant: {assistant_utterance}

...

User:
```

MultiWOZ pretrained base simulator

```
USER GOAL: {goal}

--------------------

USER: {user_utterance}

SYSTEM: {system_utterance}

...

USER:
```

Tandem User-Agent Prompts

User Agent system prompt

```
You are simulating how a human would respond. Your task is to generate authentic,
    human-like replies that are consistent with the user description below.

## User description
{user_description}
```

```
---
## Your process for each turn

Each time you need to produce a reply, follow exactly two steps.

### Step 1 -- Plan and request candidates
### Step 2 -- Select appropriate candidate
```

User Agent planning prompt

```
### Step 1 -- Plan the next turn

Open a <user_think> block. Inside it:
1. **Belief**: Summarise what has happened in the conversation and what you
    currently understand.
2. **Desire**: State what you want to achieve in this turn (your conversational
    goal).
3. **Intention**: State the specific communicative act you will perform (e.g. "
    share an anecdote", "ask a clarifying question", "express mild disagreement").

Do not produce the final reply yet. Candidate replies will be provided in the next
    step.

Close the </user_think> block.
```

User Agent candidate-selection prompt

```
### Step 2 -- Analyze candidates

You have receive a numbered list of candidate utterances.
{tool_result_str}

Open a new <user_think> block.
Score each candidate on the following rubric:

{rubric_text}

Write a brief score (1-10) and one-sentence reason for each candidate, then name
    the best.
Close the </user_think> block.

Do not copy the final candidate utterance in this step. The next step will ask for
    the final exact candidate text.
```

User Agent final-selection prompt

```
### Step 3 -- Final candidate text

Using your candidate analysis above, select the best candidate from the numbered
    list. Output only the EXACT text of that one candidate.

Do NOT include reasoning, markdown, candidate numbers, labels, quotes, or XML tags.
Do NOT modify, rephrase, shorten, or add anything.
Your entire response must be character-for-character identical to one of the
    numbered candidates.

You MUST select one candidate.
```

Critique Only tandem prompt

```
You are simulating how a human would respond.
Candidate utterances will be sampled from another model. Your task is to select the
     first candidate reply that meets all criteria in the rubric.

## User description
{user_description}

Do not rewrite candidates. Do not produce a new reply. Return only an exact
    candidate utterance after your critique.

### Candidate critique and selection

You have received a numbered list of candidate utterances:
{tool_result_str}

Evaluate candidates in ascending order against every criterion in this rubric:

{rubric_text}

Open a <user_think> block.
For each candidate, briefly state PASS or FAIL for each criterion. Then select the
    FIRST candidate that passes ALL criteria.
When multiple candidates meet all criteria, select the first passing candidate in
    the list.

Close the </user_think> block.

After </user_think>, copy the EXACT text of the chosen candidate -- do NOT modify,
    rephrase, shorten, or add anything. Your output after </user_think> must be
    character-for-character identical to one of the numbered candidates.
```

```
You MUST select one of the candidates. Do NOT reject all candidates or write in
    your own text outside of the provided options.
```

DialOp Annotation Prompts

DialOp preference annotator: system prompt

```
You are an expert annotator for a trip-planning dialogue dataset.

Player 0 (the "user") was assigned a list of preferences before the conversation.
   Player 1
(the "agent") helps plan an itinerary. Your job is to determine whether a given
  Player 0 utterance
expresses any of their assigned preferences.

Rules:
- Only identify preferences that Player 0 is actively expressing in this
  utterance. Do not guess or
infer preferences that are not clearly present.
- For each preference you identify and extract the specific words/phrase from the
   utterance that
express it. Do not return the entire utterance -- only the relevant substring(s).
   The extracted text
 must appear verbatim in the utterance.
- A single utterance can express zero, one, or multiple preferences.
- If the utterance does not express any preference (e.g. "Yes", "Greetings!"),
  return an empty list.
```

DialOp preference annotator: user prompt

```
## Player 0's Assigned Preferences
{preferences}

## Player 0's Utterance
"{utterance}"

## Task
Identify which (if any) of the assigned preferences Player 0 is expressing in
  this utterance. For
each match, extract the verbatim text from the utterance that conveys the
  preference.

Return a JSON list:
[
```

```
    {
      "preference_index": <1-indexed int matching the preference list above>,
      "verbatim_text": "<exact substring from the utterance that expresses this
    preference>"
    }
  ]

  If no preferences are expressed, return: []
  Return ONLY the JSON list, no other text.
```

A.3 Qualitative Generation Examples

This appendix shows representative next-utterance examples from each evaluation regime. Each block gives the rendered prompt context, followed by the aligned human continuation and selected model continuations. User identifiers in public forum examples are anonymized for presentation.

Reddit

Reddit qualitative example

```
System:
You are simulating a Reddit user engaging with conversations with different users.
    You are to simulate the user u/user_a. You will be given past conversations that
     this user has engaged with, and based on that, continue the conversation
    naturally as the user. Only answer with the next user utterance and no other
    text.

User:
Here is the conversation context, consisting of Reddit thread(s):

###THREADS START

[Title]: Why the hate

[Original Poster (OP) u/user_a]: Why the hate with pyro mains,what did we ever do?

[u/user_b]: Because pyro is borderline OP and it takes the least amount of skill to
     perform very well with it. It's DPS is ridiculous, flames catch you 360 degrees
     around the pyro, no matter where he's aiming, the reach is insane, then there's
     the afterburn damage on top of everything. Oh, and pyros can fly now. It takes
    a decent player to kill a newbie pyro. People generally don't like that. So,
    yeah, there you go.

[Original Poster (OP) u/user_a]: well i fly,and i died many times because i suck
```

```
[u/user_b]: Everyone dies many times. It's a game. And with pyro simple w+m1 works
    annoyingly well. Just don't charge a heavy from 5km away while shooting flares
    at him.

[Original Poster (OP) u/user_a]: i dont use flares,i use the jetpack and tey to get
     behind them with bacnburner

[u/user_b]: All you have to do is be a little bit sneaky. Backburner melts
    everything. TF2 is meant to be a fun game. If you get frustrated too much, just
    go with other class :]

[Original Poster (OP) u/user_a]:

###THREADS END

Generate a comment that is in line with the final thread conversation for user u/
    user_a. Continue the conversation naturally as the user. Only answer with the
    comment and no other text.

Human continuation:
like scout where a sniper tries to mele me when i have a force a nature at his head
    ,when i get frustated i change to scout sniper or spy

Mistral-Small-24B-Base-2501:
Ty,maybe im just disappointed that im not better

User Agent / Qwen3-14B selector / Mistral-Small-24B-Base-2501 speaker:
"well i started from pyro and i just love spamming buttons,I have the rocket
    launcher secondary and i am more like a soldier pyro"
```

WildChat-1M

WildChat-1M qualitative example

```
System:
Continue the conversation naturally as the user. Only answer with the next user
    utterance and no other text.

User:
User: Hello

Assistant: Hello! How can I assist you today?

User: Are you chatgpt 4?
```

```
Assistant: I am an AI developed by OpenAI, based on the GPT (Generative Pretrained
    Transformer) architecture. Whether I'm specifically referred to as "ChatGPT-4"
    depends on the information and training provided by OpenAI at the time of my
    last update. If you have any specific questions or need assistance, feel free to
     ask, and I'll do my best to help!

User:

Continue the conversation naturally as the user. Only answer with the next user
    utterance and no other text.

Human continuation:
What version are you

Mistral-Small-24B-Base-2501:
How can we make you the me?

Mistral-Small-24B-Instruct-2501:
What can you do?
```

LMSYS-Chat-1M

LMSYS-Chat-1M qualitative example

```
System:
Continue the conversation naturally as the user. Only answer with the next user
    utterance and no other text.

User:
User: Hello, Who are you?

Assistant: Hello! I'm a language model called Assistant, and I was trained by
    researchers from Large Model Systems Organization (LMSYS).

User:

Continue the conversation naturally as the user. Only answer with the next user
    utterance and no other text.

Human continuation:
do you know llama.cpp?

Mistral-Small-24B-Base-2501:
What can you teach me?

Mistral-Small-24B-Instruct-2501:
Nice to meet you, Assistant. What can you do?
```

DialOp

DialOp qualitative example

```
System:
You are simulating a user engaging with conversations with a different user on
    planning a travel itinerary. You are to simulate the user marked as 'User'. You
    will be given past conversations that this user has engaged with, and based on
    that, you must generate a comment that is like the given user.

User:
User goals:
- family trip, best to go to places that are good for kids
- check out live music!
- keep budget below $90
- don't like: spanish, american, like: ethiopian, korean
- generally like eating vegan options
- definitely want to go to Liberty Memorial
- renting a car so preferably places with parking, but street parking also ok
- go to at least one park
- minimize travel distance
```

```
Here is the conversation context so far:

###CONVERSATION START

User: Hello

Assistant: Hello. Where do you think we should go?

User: Perhaps a place with live music could be nice.

Assistant: How about the Kozy Kar.

User: Would that be a good place for my family? I am traveling with them.

Assistant: It's good for groups so it would probably be okay.

User:

###CONVERSATION END

Generate an utterance that is in line with the final thread conversation for the
    User. Only answer with the utterance and no other text.

Human continuation:
Okay, that sounds good for one of the places we can visit. I only have time for two
     other places and I want to go to Liberty Memorial.

Mistral-Small-24B-Base-2501:
Is there a place to park?

Mistral-Small-24B-Instruct-2501:
Do you know if there is a good option for vegetarian or vegan meals as well?
```

MultiWOZ

MultiWOZ qualitative example

```
System:
You are a simulating a user engaging in a task-oriented dialogue. Continue the
    conversation as the USER. Only answer with the next USER utterance and no other
    text.

User:
Dialogue context:
USER GOAL: You are looking for information in Cambridge. You are looking for a
    restaurant. The restaurant should be in the centre and should be in the
```

```
    expensive price range. The restaurant should serve asian oriental food. Once you
     find the restaurant you want to book a table for 8 people at 11:45 on sunday.
    Make sure you get the reference number. You are also looking for places to go in
     town. The attraction should be in the type of college. Make sure you get phone
    number, postcode, and area

--------------------

USER: Hi, yes, I'm trying to find a restaurant to go to.

SYSTEM: In what area of town would you like to dine?

USER: An expensive restraunt in town centre that serves asian oriental food

SYSTEM: Great, there's a restaurant called kymmoy that fits your criteria. Would
    you like their information?

USER: Can you book a table for me?

SYSTEM: Yes, what is the date and time that you will need the booking and how many
    people?

USER: book a table for 8 at 15:45 on sunday

SYSTEM: That time is all booked up would you like to try a different time?

USER: Yes please. Can we try 11:45?

SYSTEM: Booking was successful. The table will be reserved for 15 minutes.
    Reference number is : 65EX4JA4.

USER:

Continue the conversation as the USER. Only answer with the next USER utterance and
     no other text.

Human continuation:
Can you suggest some college's in town? I need a phone number, postcode and the
    area too please

Mistral-Small-24B-Base-2501:
Awesome! Well done. Can you find for me a college attraction in cambridge.

Mistral-Small-24B-Instruct-2501:
Do you think Kymmoy can accommodate large groups? I want to have the best
    experience for my friends and I. Also, What colleges are nearby?
```

A.4 Metric Details

The paper reports three families of measurements. First, perplexity compares how surprising human and generated user utterances are under the same autoregressive model. Second, diversity metrics measure how much the multiple generated continuations for a fixed dialogue context differ from one another. Third, corpus-level and pragmatic metrics compare model behavior to the human corpus at larger granularity. Unless otherwise stated, diversity metrics are computed within each dialogue context and then averaged across contexts; context-level means are not weighted by the number of generation pairs in that context.

Perplexity Diagnostics

Perplexity is the exponentiated average predictive uncertainty of a language model and is a standard diagnostic for language modeling [106]. We compute it over target tokens under an autoregressive language model. For a target sequence $y_{1:T}$ and context c, the average negative log-likelihood is

$$\mathrm{NLL}(y \mid c) = -\frac{1}{T}\sum_{t=1}^{T} \log p(y_t \mid c, y_{<t}),$$

where the log is natural. The reported per-token perplexity is

$$\mathrm{PPL}_{\mathrm{token}}(y \mid c) = \exp\left(\frac{\sum_i \mathrm{NLL}_{i,\mathrm{total}}}{\sum_i T_i}\right),$$

with totals accumulated over all evaluated utterances. Thus average NLL is measured in nats/token and perplexity is its exponential.

We evaluate two quantities. *Human* perplexity scores the aligned human continuation under the model. *Generated* perplexity scores model-generated continuations under the same model. For generated text, the comparison plots use a random generation from each source context so that the generated side is a sample from the model's distribution rather than a best-of-N selection. A low Generated perplexity together with a high Human perplexity indicates that the model is internally self-consistent but poorly aligned to the empirical user-text distribution.

Lexical and Pairwise Diversity

For each dialogue context, diversity metrics are computed over the set of generated continuations after empty outputs are removed and duplicate strings are deduplicated. Exact duplicate multiplicity is therefore ignored in the default diversity run; if all continuations in a context are empty or a metric is undefined for the remaining set, that context contributes a missing value rather than being forced to zero, except where noted below. Reported values are means over valid contexts.

Distinct-n. Distinct-n was introduced as a response-diversity metric for neural conversation models [136]. It is the number of unique n-grams divided by the total number of n-grams across generated continuations:

$$\text{Distinct-}n = \frac{\left|\bigcup_j \text{ngrams}_n(\hat{x}^{(j)})\right|}{\sum_j |\text{ngrams}_n(\hat{x}^{(j)})|}.$$

We report Distinct-1 and Distinct-2. Tokenization follows the evaluation implementation's whitespace tokenization. If no n-grams are available after preprocessing, the context-level score is 0. Higher Distinct-1 or Distinct-2 means greater lexical variety, with Distinct-2 more sensitive to repeated phrases.

Self-BLEU. Self-BLEU reuses the BLEU precision-and-brevity framework [180] as an intra-sample diversity metric for text generation, following Texygen [277]. Each continuation is treated as a prediction and all other continuations for that context are treated as references. The context-level Self-BLEU score is the mean of these continuation-level BLEU scores. We report Self-BLEU as a self-similarity metric: lower values indicate greater within-context lexical variety, while high values mean that the sampled continuations share much of the same surface form.

Self-ROUGE-L. Self-ROUGE-L uses ROUGE-L [143] as another pairwise self-similarity diagnostic. For a context with m deduplicated continuations, the implementation computes ROUGE-L F-measure over generation pairs $(\hat{x}^{(i)}, \hat{x}^{(j)})$ and averages the pair scores. ROUGE-L is based on longest common subsequence overlap, so it captures repeated subsequences even when exact n-gram precision is not high. We report it as *Self-ROUGE-L*; lower values indicate more diverse generations.

Self-BERTScore F1. Self-BERTScore measures semantic self-similarity rather than surface overlap [266]. The implementation scores unordered generation pairs within a context with `microsoft/deberta-xlarge-mnli`, using no IDF weighting. The BERTScore package's English baseline rescaling is enabled in the raw evaluation CSVs, and the analysis report restores those values to the ordinary unrescaled BERTScore scale before producing tables and plots. We report the restored F1 value, *Self-BERT-F1*, as the main semantic similarity summary. Lower Self-BERT-F1 means the continuations are semantically farther apart; values near 1 indicate near-identical semantic content.

Vendi Score

Vendi Score measures the effective number of distinct samples represented by a set of generations [73]. For a set of m generated continuations $S = \{\hat{x}^{(1)}, \ldots, \hat{x}^{(m)}\}$, each continuation is embedded with the sentence-embedding encoder used for the diversity sweep (`sentence-transformers/all-MiniLM-L6-v2` unless otherwise specified). Embedding vectors are ℓ_2-normalized, and the cosine Gram matrix $K \in \mathbb{R}^{m \times m}$ is symmetrized before eigendecomposition. Let $\lambda_1, \ldots, \lambda_m$ be the eigenvalues of K, normalized so that they sum to one. The

Vendi Score is

$$\text{Vendi}(S) = \exp\left(-\sum_{r=1}^{m} \lambda_r \log \lambda_r\right).$$

The score can be interpreted as the effective number of semantically distinct continuations in the sample set. A value close to one indicates that the generations collapse to a single semantic mode, while larger values indicate broader coverage of distinct continuations. Because Vendi is kernel-based, its absolute value depends on the embedding model and similarity function; we therefore compare systems under the same encoder and compute the score per context before aggregating across contexts.

MAUVE

MAUVE [193] measures corpus-level similarity between the distribution of human continuations and the distribution of model continuations. Unlike the self-diversity metrics above, MAUVE is not computed independently inside each context. It compares two sets of response-only embeddings:

$$P = \{e(x_i^H)\}_{i=1}^N, \qquad Q = \{e(\hat{x}_i)\}_{i=1}^N,$$

where x_i^H is the aligned human continuation for context i and $\hat{x}_i$ is one sampled model continuation for the same context. Context text, source identifiers, and human/model labels are used only for alignment and are not embedded.

For each MAUVE resample, one generated continuation is sampled uniformly from the available generations for every source context, producing N human embeddings and N model embeddings. We then call `mauve.compute_mauve` with precomputed feature matrices `p_features` and `q_features`. The MAUVE package internally ℓ_2-normalizes rows, fits PCA on the combined human/model features, clusters the projected points into histogram bins, and computes a divergence frontier between the two empirical histograms. The reported *MAUVE Human-Model* score is the mean of the resampled MAUVE values. Higher MAUVE indicates that the model and human continuation distributions are closer. We also record the frontier integral and resampling spread for diagnostics, but the paper reports MAUVE itself.

The default embedding backend for the final runs is `gemini-embedding-2`; local smoke runs can use `sentence-transformers/all-MiniLM-L6-v2`. The number of histogram buckets is set consistently within a comparison run. With the automatic setting, the package uses approximately $\max(2, \text{round}(N/10))$ buckets, where N is the number of aligned contexts in the resample.

Dialogue-Act Analysis

The dialogue-act analysis evaluates whether simulators preserve pragmatic structure over task-oriented dialogue, not merely surface lexical diversity. Dialogue acts have long been

used to represent the communicative function of conversation turns and the structure of act sequences [226]. Recent work also uses dialogue-act structure as a lens for analyzing LLM conversational behavior [155, 214]. We focus this analysis on MultiWOZ 2.1 [37], where task-oriented turns have structured dialogue-act annotations, and use ConvLab BERTNLU [276] to label generated user utterances.

BERTNLU is run with MultiWOZ context enabled: the generated utterance is labeled using up to the previous three utterances as classifier context. If a generated utterance yields no acts and contains multiple sentences, sentence-level fallback labeling is used and the recovered acts are merged. The human side uses the aligned reference turn annotations. Thus, each evaluation row compares a human target user turn and a model replacement for the same dialogue state.

Each labeled turn is converted to a canonical intent token. We lower-case the dialogue-act intents, sort the unique intent labels, and discard slot values for the sequence-level analysis. For a user turn, the token has the form `U:(intent-set)`; for a system turn, the token has the form `S:(intent-set)`. Empty intent sets are represented by `U:(none)` or `S:(none)`. This representation retains speaker identity and turn-level communicative function while avoiding sparsity from slot-value strings.

Labeling and Sequence Construction Each model sequence replaces only the target human user act with the model-predicted act while preserving the reference context. For consistency, we also relabel the original human user turns over which the BERTNLU labeler achieves 88.6% exact match accuracy.

The canonical intent inventory is {`book`, `bye`, `greet`, `inform`, `nobook`, `nooffer`, `offerbook`, `offerbooked`, `recommend`, `reqmore`, `request`, `select`, `thank`, `welcome`}; empty predictions are mapped to `none`. For each turn, we lower-case the detected intents, sort the unique labels, discard domain/slot/value fields for this sequence analysis, and prefix the resulting set with the speaker. Thus a user turn may become `U:(inform,request)`, while an empty system act becomes `S:(none)`.

Metric Definitions For the reasoning behind the selection of $N = 1, \ldots, 3$, see Appendix A.5. Terminal N-grams measure the act sequence ending at the target user turn; included N-grams include every local window containing that target turn. For each N, we construct human counts $c_N^H(g)$ and model counts $c_N^M(g)$ over the union vocabulary $\mathcal{V}_N = \mathcal{V}_N^H \cup \mathcal{V}_N^M$ of observed act N-grams, normalize them into empirical distributions P_N and Q_N, and assign zero probability to patterns absent from one side. Jensen-Shannon divergence [146] is computed between P_N and Q_N on this shared support. We also report top-k vocabulary overlap, $\text{Overlap}_{N,k} = |\text{TopK}(P_N, k) \cap \text{TopK}(Q_N, k)| / k$, where $\text{TopK}(P_N, k)$ is the set of the k most frequent human dialogue-act N-grams and $\text{TopK}(Q_N, k)$ is the analogous model set. Although small vocabulary sizes ($|V| < k$) naturally limit N-gram overlap, the utility of the metric remains intact, as we prioritize inter-model comparisons over absolute values across N-grams.

Terminal N-grams. For each target user turn at index t, the human terminal N-gram is the dialogue-act sequence ending at the human turn:

$$g^H_{t,N} = (a_{t-N+1}, \ldots, a_t).$$

The model terminal N-gram is formed by replacing only the final human user act with the model-predicted user act, preserving the preceding reference context:

$$g^M_{t,N} = (a_{t-N+1}, \ldots, \hat{a}_t).$$

We compute these distributions for $N = 1, \ldots, 5$.

Included N-grams. Included N-grams consider every length-N window that contains the replaced target turn. For the model distribution, the target token in each such window is replaced with the model-predicted user-act token while the rest of the window is kept from the reference dialogue. This captures how a generated act changes local pragmatic patterns not only as the last act in a prefix, but also as part of surrounding multi-turn context.

Jensen-Shannon divergence. For a given N, let P_N be the empirical distribution of human dialogue-act N-grams and Q_N be the corresponding model distribution over the union of observed human and model patterns. We report Jensen-Shannon divergence [146],

$$\mathrm{JSD}(P_N, Q_N) = \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P_N \| R_N) + \frac{1}{2} D_{\mathrm{KL}}(Q_N \| R_N)}, \qquad R_N = \frac{1}{2}(P_N + Q_N),$$

using the square-rooted distance returned by the SciPy Jensen-Shannon implementation. Larger values indicate greater divergence from human pragmatic structure.

Top-k vocabulary overlap. For each N, we compute the overlap between the k most frequent human dialogue-act N-grams and the k most frequent model dialogue-act N-grams:

$$\mathrm{Overlap}_{N,k} = \frac{|\mathrm{TopK}(P_N, k) \cap \mathrm{TopK}(Q_N, k)|}{k}.$$

We compute this diagnostic for $k \in \{10, 50, 100\}$ and report $k = 50$ in the main appendix tables. This metric asks whether a simulator recovers the common pragmatic patterns of the human corpus, even when exact frequencies differ. We also track model-novel pattern mass, the fraction of model N-gram probability assigned to patterns never observed in the human reference distribution.

BPE-style pattern inventory. As an additional sequence-level diagnostic, we train a byte-pair-encoding-style merge inventory on human dialogue-act sequences. Starting from the canonical speaker-marked act tokens, the most frequent adjacent pair is merged into a macro-pattern token; this process is repeated for a fixed merge budget. The learned human merge rules are then applied unchanged to model sequences. Divergence, entropy ratio, and coverage ratio of the resulting macro-pattern distributions quantify whether the model reproduces recurring multi-turn pragmatic motifs rather than only matching isolated turn intents.

Table A.1: **Full Reddit point estimates.** PPL-H scores aligned human continuations; PPL-G_{rand} scores one randomly sampled generated continuation per context.

System	PPL-H	PPL-G_{rand}	Dist-1	Dist-2	Self-BLEU	Self-ROUGE-L	Self-BERT-F1	Vendi	MAUVE
Llama-3.1-8B	13.2424	19.2998	0.6743	0.9602	0.0122	0.0776	0.4796	6.3915	0.9650
Llama-3.1-8B-Instruct	23.1513	6.9341	0.5984	0.9339	0.0545	0.1304	0.5337	4.7601	0.6946
Mistral-24B-Base	10.7243	17.2362	0.6562	0.9581	0.0152	0.0845	0.4884	6.2010	0.9658
Mistral-24B-Instruct	13.4477	10.4384	0.6147	0.9263	0.0790	0.1473	0.5560	5.0822	0.8999
Qwen2.5-14B	12.8851	16.4860	0.6770	0.9593	0.0109	0.0799	0.4831	6.4020	0.9666
Qwen2.5-14B-Instruct	32.1536	2.7724	0.5726	0.8520	0.2017	0.2176	0.6285	4.1150	0.7728
Qwen2.5-7B	14.6720	17.8660	0.6769	0.9598	0.0109	0.0757	0.4797	6.4709	0.9564
Qwen2.5-7B-Instruct	36.8406	2.8579	0.6302	0.8807	0.1697	0.1969	0.6126	4.4223	0.4143
Qwen3-14B	86.0188	3.2910	0.4960	0.7797	0.3137	0.2697	0.6519	3.3339	0.7766
Qwen3-14B-Base	13.2737	21.1793	0.6349	0.9503	0.0123	0.0759	0.4770	6.3468	0.9554
Qwen3-8B	129.8805	3.7846	0.5041	0.7907	0.2953	0.2620	0.6448	3.4566	0.7524
Qwen3-8B-Base	14.4802	23.2658	0.6747	0.9620	0.0088	0.0748	0.4806	6.4705	0.9573
Tandem / Qwen3-14B / Mistral-24B-Base	–	–	0.6047	0.9433	0.0244	0.1065	0.5132	5.4427	0.9343
Tandem / Qwen3-14B / Qwen3-14B-Base	–	–	0.5939	0.9394	0.0339	0.1005	0.5044	5.5744	0.9244
Tandem / Qwen3-14B / Qwen3-14B	–	–	0.4796	0.7698	0.2789	0.2763	0.6573	3.1051	0.7272

A.5 Full Result Tables

This appendix gives the complete model-by-model result tables for the reported utterance-level metrics. The tables use the same metric definitions as Appendix A.4: PPL-H scores aligned human continuations, PPL-G_{rand} scores one randomly sampled generated continuation per context, and the remaining columns report within-context diversity or corpus-level distributional similarity. Missing cells indicate that the metric was not available for that system family.

Reddit

Table A.1 reports the full Reddit continuation metrics for each evaluated base, instruct, reasoning, and tandem system.

WildChat-1M

Table A.2 reports the corresponding results on WildChat-1M human-LLM chat continuations.

LMSYS-Chat-1M

Table A.3 gives the full LMSYS-Chat-1M continuation results for the same metric family.

Table A.2: **Full WildChat-1M point estimates.** PPL-H scores aligned human continuations; PPL-G_{rand} scores one randomly sampled generated continuation per context.

System	PPL-H	PPL-G_{rand}	Dist-1	Dist-2	Self-BLEU	Self-ROUGE-L	Self-BERT-F1	Vendi	MAUVE
Llama-3.1-8B	4.3322	8.1718	0.6483	0.8977	0.0785	0.1122	0.5141	6.1210	0.8608
Llama-3.1-8B-Instruct	5.2963	9.5169	0.6068	0.9044	0.0865	0.1510	0.5840	4.7968	0.7813
Mistral-24B-Base	3.5866	8.9617	0.6321	0.8875	0.0902	0.1170	0.5166	5.9821	0.9659
Mistral-24B-Instruct	4.0625	7.0132	0.6758	0.9275	0.0598	0.1315	0.5812	5.5360	0.6233
Qwen2.5-14B	3.8005	7.4436	0.6463	0.8871	0.0922	0.1114	0.5102	6.0735	0.9691
Qwen2.5-14B-Instruct	5.8571	2.4996	0.5625	0.7978	0.2501	0.2386	0.6641	4.0393	0.5007
Qwen2.5-7B	4.0866	6.8526	0.6406	0.8860	0.0934	0.1135	0.5118	5.9729	0.9637
Qwen2.5-7B-Instruct	6.2130	2.9001	0.5402	0.7549	0.3020	0.2791	0.6924	3.9961	0.4634
Qwen3-14B	6.3684	1.6479	0.4825	0.7123	0.3968	0.3021	0.6781	3.2709	0.8739
Qwen3-14B-Base	3.7557	5.4252	0.6338	0.8815	0.0934	0.1116	0.5094	6.0090	0.9682
Qwen3-8B	7.0478	1.7676	0.5031	0.7289	0.3682	0.2869	0.6689	3.3980	0.8489
Qwen3-8B-Base	3.9346	9.8134	0.6287	0.8807	0.0937	0.1101	0.5097	6.0507	0.9696
Tandem / Qwen3-14B / Mistral-24B-Base	–	–	0.5752	0.8527	0.1318	0.1577	0.5505	5.0599	0.9672
Tandem / Qwen3-14B / Qwen3-14B-Base	–	–	0.5741	0.8439	0.1523	0.1588	0.5491	4.9824	0.9746

Table A.3: **Full LMSYS-Chat-1M point estimates.** PPL-H scores aligned human continuations; PPL-G_{rand} scores one randomly sampled generated continuation per context.

System	PPL-H	PPL-G_{rand}	Dist-1	Dist-2	Self-BLEU	Self-ROUGE-L	Self-BERT-F1	Vendi	MAUVE
Llama-3.1-8B	4.0134	7.1166	0.6723	0.8993	0.0660	0.1117	0.5177	6.3393	0.8206
Llama-3.1-8B-Instruct	5.8418	3.5852	0.5955	0.8857	0.1070	0.1662	0.5911	4.8261	0.7639
Mistral-24B-Base	3.3318	7.3487	0.6714	0.8959	0.0746	0.1162	0.5185	6.2141	0.9503
Mistral-24B-Instruct	4.3267	4.9720	0.6673	0.9010	0.0918	0.1625	0.5977	5.3387	0.7026
Qwen2.5-14B	3.4565	6.1836	0.6676	0.8830	0.0837	0.1194	0.5196	6.1912	0.9499
Qwen2.5-14B-Instruct	12.5903	2.1503	0.5200	0.7306	0.3556	0.3049	0.6864	3.6589	0.6036
Qwen2.5-7B	3.6256	6.8400	0.6791	0.8942	0.0825	0.1156	0.5195	6.2180	0.9499
Qwen2.5-7B-Instruct	10.1829	2.1135	0.5207	0.7123	0.3595	0.3150	0.7053	3.9144	0.5897
Qwen3-14B	11.6769	2.0274	0.4873	0.6847	0.4276	0.3369	0.6913	3.1596	0.8327
Qwen3-14B-Base	3.3858	6.2173	0.6669	0.8892	0.0772	0.1043	0.5063	6.3051	0.9542
Qwen3-8B	13.1645	2.3498	0.5127	0.7134	0.3784	0.3075	0.6790	3.4498	0.8237
Qwen3-8B-Base	3.5010	7.9230	0.6708	0.8996	0.0719	0.1049	0.5065	6.3690	0.9516
Tandem / Qwen3-14B / Mistral-24B-Base	–	–	0.6217	0.8653	0.1162	0.1603	0.5574	5.2694	0.9421
Tandem / Qwen3-14B / Qwen3-14B-Base	–	–	0.5992	0.8459	0.1469	0.1606	0.5512	5.2067	0.9587
Tandem / Qwen3-14B / Qwen3-14B	–	–	0.4464	0.6482	0.4367	0.3714	0.7109	2.8156	0.8054

MultiWOZ Continuation Metrics

Table A.4 reports the MultiWOZ continuation metrics before the dialogue-act-specific analysis below.

DialOp

Table A.5 gives the full DialOp continuation results, using the same point-estimate columns as the other datasets.

Table A.4: **Full MultiWOZ continuation point estimates.** PPL-H scores aligned human continuations; PPL-G_{rand} scores one randomly sampled generated continuation per context.

System	PPL-H	PPL-G_{rand}	Dist-1	Dist-2	Self-BLEU	Self-ROUGE-L	Self-BERT-F1	Vendi	MAUVE
Llama-3.1-8B	5.5176	7.9276	0.7019	0.9066	0.0809	0.1862	0.5428	5.1383	0.6172
Llama-3.1-8B-Instruct	10.7274	2.8377	0.5258	0.8028	0.2477	0.2610	0.6700	3.6958	0.3526
Mistral-24B-Base	4.8951	5.8217	0.6910	0.8948	0.0956	0.1845	0.5804	5.0722	0.6743
Mistral-24B-Instruct	5.0876	3.6643	0.5731	0.8003	0.2424	0.2851	0.6932	3.8587	0.6687
Qwen2.5-14B	5.1235	6.3018	0.6654	0.8824	0.1153	0.2005	0.6367	4.9447	0.7132
Qwen2.5-14B-Instruct	20.0783	1.7114	0.4251	0.6084	0.5541	0.4712	0.6949	2.5738	0.7499
Qwen2.5-7B	5.2737	6.6054	0.6811	0.8959	0.0990	0.1879	0.5992	5.0275	0.7100
Qwen2.5-7B-Instruct	27.2847	1.5299	0.4089	0.5647	0.5995	0.5266	0.7808	2.4876	0.7171
Qwen3-14B	72.2336	2.7179	0.4119	0.5431	0.6375	0.5187	0.7802	2.3100	0.5098
Qwen3-14B-Base	5.0428	5.7457	0.6600	0.8800	0.1162	0.1898	0.5905	5.0127	0.7116
Qwen3-8B	65.4066	2.8170	0.4608	0.5839	0.5689	0.4950	0.9343	2.4334	0.5816
Qwen3-8B-Base	5.2780	6.3251	0.6829	0.8939	0.1026	0.1809	0.4689	5.1156	0.6535
Tandem / Qwen3-14B / Mistral-24B-Base	–	–	0.5884	0.8143	0.2177	0.2912	0.6641	3.9454	0.8664
Tandem / Qwen3-14B / Qwen3-14B-Base	–	–	0.5692	0.8062	0.2140	0.2880	0.6590	3.9370	0.4029
Tandem / Qwen3-14B / Qwen3-14B	–	–	0.3541	0.4872	0.6496	0.5773	0.8085	2.0580	0.0957

Table A.5: **Full DialOp point estimates.** PPL-H scores aligned human continuations; PPL-G_{rand} scores one randomly sampled generated continuation per context.

System	PPL-H	PPL-G_{rand}	Dist-1	Dist-2	Self-BLEU	Self-ROUGE-L	Self-BERT-F1	Vendi	MAUVE
Llama-3.1-8B	12.4540	16.5786	0.7327	0.9570	0.0125	0.0829	0.5614	6.5064	0.4638
Llama-3.1-8B-Instruct	23.6024	4.4736	0.5588	0.8957	0.0868	0.1711	0.6723	4.1654	0.3901
Mistral-24B-Base	10.9471	12.0142	0.7194	0.9522	0.0180	0.0821	0.5996	6.4174	0.5942
Mistral-24B-Instruct	12.3744	9.9723	0.5982	0.9307	0.0434	0.1367	0.6078	4.8471	0.4900
Qwen2.5-14B	12.3213	8.0706	0.7019	0.9454	0.0245	0.0911	0.5743	6.2673	0.5640
Qwen2.5-14B-Instruct	57.3444	2.4584	0.5177	0.7909	0.2729	0.2582	0.6922	3.4711	0.3958
Qwen2.5-7B	12.9216	10.9312	0.7010	0.9458	0.0207	0.0833	0.5352	6.3224	0.5293
Qwen2.5-7B-Instruct	73.1071	2.2353	0.5127	0.7884	0.2967	0.2657	0.7231	3.2720	0.2140
Qwen3-14B	144.0394	4.5477	0.4180	0.6866	0.4523	0.3576	0.7450	2.3924	0.0835
Qwen3-14B-Base	11.9030	10.3060	0.7005	0.9455	0.0153	0.0822	0.5151	6.3563	0.6440
Qwen3-8B	194.7983	4.5915	0.4189	0.7103	0.3920	0.3205	0.7336	2.4051	0.0904
Qwen3-8B-Base	12.8801	8.3997	0.6923	0.9323	0.0216	0.0782	0.4444	6.3152	0.5132
Tandem / Qwen3-14B / Mistral-24B-Base	–	–	0.6780	0.9362	0.0379	0.1091	0.5519	5.7922	0.5669

MultiWOZ Dialogue-Act Results

Human Intrasection JSD

Table A.6: **Comprehensive results for N-gram dialogue-act sequences.** Results include Jensen-Shannon Divergence (JSD ↓) and Vocabulary Overlap (↑) at multiple K thresholds.

	JSD (↓)			Overlap K=10 (↑)			Overlap K=100 (↑)			Overlap K=200 (↑)		
System	$N=1$	2	3	$N=1$	2	3	$N=1$	2	3	$N=1$	2	3
Llama-3.1-8B-Base	0.198	0.293	0.346	0.700	0.600	0.600	0.080	0.700	0.770	0.040	0.600	0.645
Llama-3.1-8B-Inst	0.265	0.366	0.407	0.400	0.700	0.700	0.070	0.640	0.640	0.035	0.495	0.570
Mistral-24B-Base	0.181	0.277	0.330	0.700	**0.800**	0.500	0.080	0.750	0.740	0.040	0.620	0.660
Mistral-24B-Inst	0.195	0.299	0.346	0.700	0.700	**0.800**	0.080	0.680	0.670	0.040	0.575	0.635
Qwen2.5-7B-Base	0.168	0.271	0.325	0.700	**0.800**	0.600	0.080	**0.760**	0.780	0.040	0.635	0.660
Qwen2.5-7B-Inst	0.107	0.258	0.310	0.700	**0.800**	0.700	0.080	0.690	0.740	0.040	0.560	0.665
Qwen2.5-14B-Base	0.175	0.272	0.325	0.700	**0.800**	0.600	0.080	0.680	0.760	0.040	0.610	0.665
Qwen2.5-14B-Inst	0.119	0.273	0.321	0.700	0.700	0.600	0.080	0.680	0.750	0.040	0.585	0.635
Qwen3-8B-Base	0.190	0.287	0.338	0.700	0.700	0.500	0.080	0.730	0.720	0.040	0.620	0.635
Qwen3-8B	0.138	0.299	0.343	0.700	**0.800**	0.600	0.080	0.650	0.760	0.040	0.520	0.595
Qwen3-14B-Base	0.180	0.282	0.331	0.700	0.700	0.600	0.080	0.710	0.770	0.040	0.590	**0.670**
Qwen3-14B	0.184	0.354	0.389	0.700	0.400	0.300	0.080	0.650	0.670	0.040	0.545	0.565
Tandem (Q3/M24)	**0.088**	**0.219**	**0.280**	**0.800**	**0.800**	**0.800**	0.080	0.750	**0.840**	0.040	**0.650**	0.665

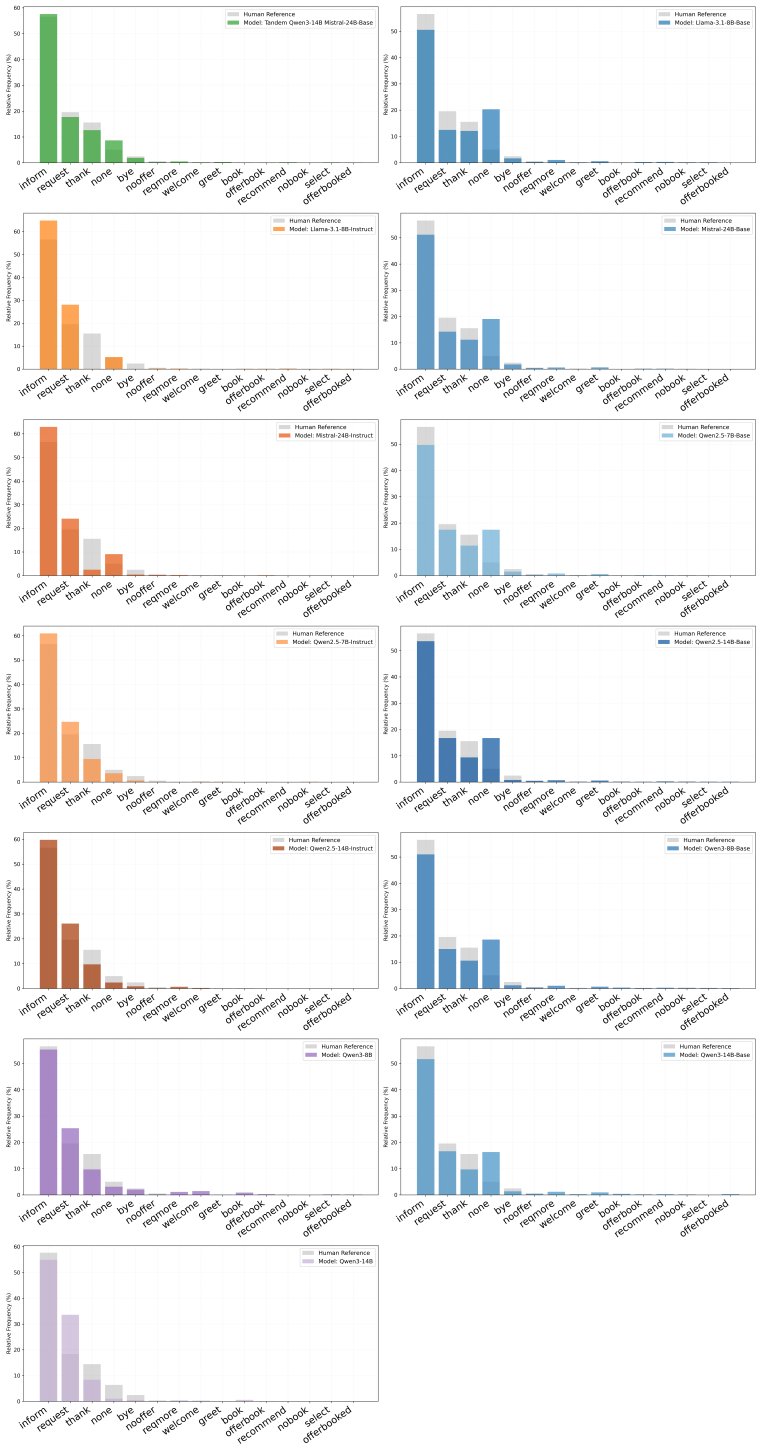

Figure A.1: **Individual (N=1) dialogue act distributions.** Each model is compared against the human distribution.

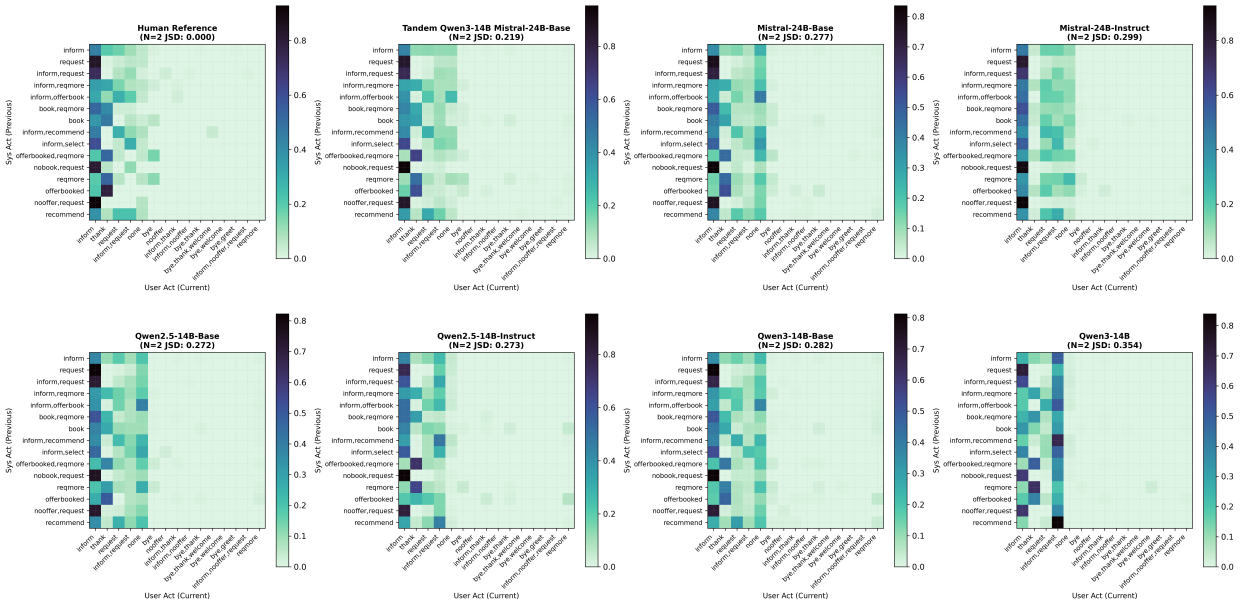

Figure A.2: **Transition probabilities for N=2.** The figure shows probability of user intent given previous assistant intent.

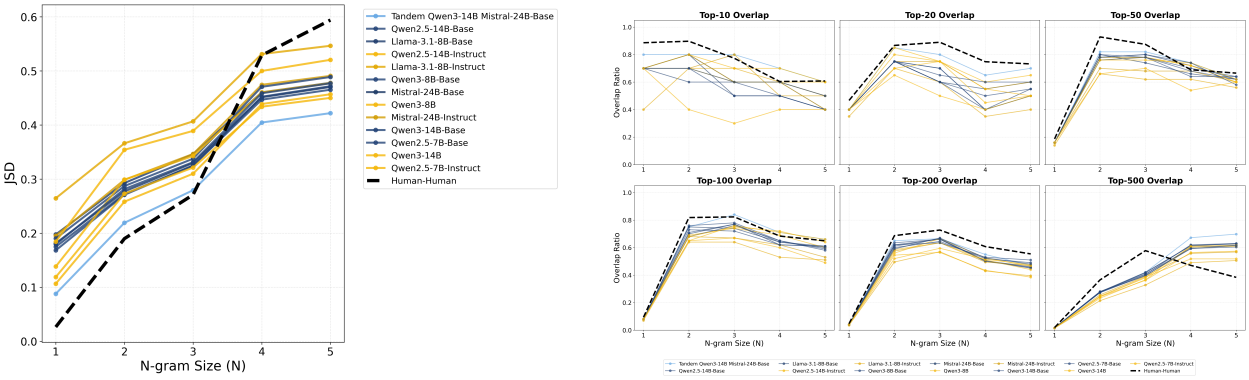

Figure A.3: **Distributional divergence and vocabulary overlap between model outputs and human-human (H-H) bootstrapped baselines.** H-H results are generated via 100× random corpus splits. We observe that for $N \leq 3$, H-H performance consistently exceeds model-human similarity, establishing $N = 3$ as our threshold complexity limit.

Appendix B

Virtual Personas for Language Models via an Anthology of Backstories

B.1 Details on LLM-Generated Backstories

In this section, we discuss additional details about the process of generating realistic backstories using language models, as mentioned in Section 3.2. We detail the prompts used and examples of LLM-generated backstories.

Then, we discuss the alternative method of generating backstories given a particular combination of demographic traits, referred in Section 3.3 as the "Demographics-Primed" method in contrast to the "Natural" backstories generated without conditioning on demographics.

Natural Generation of Backstories

We use OpenAI's `davinci-002` for generating backstories with the prompt specified in the top of Figure B.1. This model is chosen as it is base model (*i.e.* not instruction-tuned) of the largest model capacity at the time of the project. Figure B.1 shows two examples of backstories of different lengths generated with this prompt.

Generating Demographics-Primed Backstories

Target demographics-primed backstories are generated by prompting a language model with demographic information of a human from a target population. In contrast to naturally generated backstories whose demographic trait cannot be predetermined but can only be sampled by the demographic survey method outlined in Section B.4, demographic traits of target demographics-primed backstories are determined at the time of generation. We use five demographic variables (age, annual household income, education level, race or ethnicity, gender) for ATP Waves 34 and 99 and an additional variable (political affiliation) for ATP Wave 92.

A generation prompt example for ATP Wave 34 is presented in Figure B.2. Answers for each question are taken from the demographic information of a human respondent in the ATP survey data. To accurately incorporate the target population's demographic information, we use the same list of choices as used in the actual survey. The order of demographic variables is randomized every generation to minimize the effect of question ordering. We use two styles of prompt, which we refer to as Question-Answer and Biography, as presented in Figure B.2.

To take full advantage of demographics-primed backstory generation, backstories should sufficiently reflect the given demographic information. Due to pre-trained base models' limited instruction-following capability, however, demographics-primed backstories generated with pre-trained base models sometimes reflect demographic traits inconsistent with provided information. Therefore, we use the fine-tuned chat model Mixtral-8x22B [107] with decoding hyperparameters of `top_p` $= 1.0$, `T` $= 1.1$.

B.2 Details on Experiments

In this section we provide examples of prompts used in the experiments approximating human studies, as described in Section 3.3 and used to produce the results in Section 3.4. Additionally, we outline the survey procedure for conducting these experiments, providing a comprehensive review of methodologies and operational frameworks involved.

Prompts for Baseline: `QA`

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

Prompts for Baseline: `Bio`

As in [204], we construct free-text biographies in a rule-based manner given the demographic trait. The five demographic traits we use are taken from the human respondent data of ATP surveys. The order of five sentences each describing demographic traits is randomized every time to minimize the effect of sentence ordering.

Target Demographics-Primed Backstory

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

Natural Backstory

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

Survey Procedure

In this study, we mimic the same survey procedure as human surveys. Human surveys typically shuffle or reverse the order of multiple-choice options or change the order of questions for each survey participant to reduce bias in the results. Following the topline reports for each wave as provided by Pew Research, we randomly reverse the order of Likert scale questions and shuffle the options for nominal questions to ensure a similar reduction in bias.

B.3 Details on Human Studies

American Trends Panel (ATP) is a nationally representative panel of U.S. adults conducted by the Pew Research Center. ATP is designed to study a wide variety of topics, including politics, religion, internet usage, online dating, and more. We analyze sampled questions from three waves, where questions are drawn from ASK ALL questions (i.e., asked to all human respondents, instead of questions asked for selected demographic groups or conditionally asked based on the response to the previous question) in order to investigate responses from the overall population.

It is worth noting that in the original ATP surveys, some questions have answer choices in a Likert scale with the order of choices (*e.g.* positive-to-negative or negative-to-positive) randomized for each respondent. For such questions, we also randomize the order of these options when presenting them in prompts to LLMs. Here we present the list of sampled questions from each wave.

ATP Wave 34

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

ATP Wave 92

American Trends Panel Wave 92 was conducted from July 8, 2021 to July 21, 2021 with a focus on political typology. We randomly sampled 2,500 respondents for the study from the total 10,221 respondents.

ATP Wave 99

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

B.4 Demographic Survey on Virtual Personas

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

In the first approach, we use GPT-4o [176] to locate demographic information from the backstory. To minimize hallucination, we prompt GPT-4o to retrieve the demographic trait only if the backstory explicitly mentions related context (prompts are shown in Section B.4). This approach is limited to specific demographic variables, especially age, annual household income, and education level questions, since we avoid inferring race or ethnicity, gender, and party affiliation even when the backstory mentions those traits. Decoding hyperparameters are set to `top_p` = 1.0, `T` = 0.

In the second approach, we perform response sampling by prompting the language model with generated backstories appended with demographic questions. In Section B.4 we present the question format. The language model's responses are sampled 40 times for each backstory and question. Instead of estimating responses with the first-token logits [204, 92, 75], we allow the model to generate open-ended responses because some responses (e.g., "I am 25 years old." for the age question) cannot be accurately accounted for by the logit method and the sum of probability masses of valid tokens (e.g., " (A") are often too marginal to represent the true probability distribution. Sampled responses are subsequently parsed by regex matching of either the label (e.g., "(A)") or the text (e.g., "27"), recorded to obtain the distribution of 40 generations. We use Llama 3 [165] for the response sampling with decoding hyperparameters of `top_p` = 1.0, `T` = 1.0.

Combining these two approaches, our demographic survey is performed as follows. First, we use GPT-4o to locate demographic information for age, annual household income, and education level. For the remaining variables and the cases where explicit demographic information cannot be found, responses are sampled 40 times to construct a response distribution. Therefore, in the case of sampling, virtual users' demographic traits are not represented as a single trait but rather as a distribution over probable demographics given the backstory. We can thereby construct a probable estimate of demographic information without undermining the diversity of virtual authors of backstories.

Questions For Locating Demographic Information

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

Demographic questions

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

Backstory Generation Prompt ("Natural")

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

Answer:

Backstory Example 1

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

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

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

Backstory Example 2

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

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

(continued)

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

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

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

Figure B.1: **Natural backstory-generation prompt and examples.** The top-left panel shows the prompt given to LLMs for natural backstory generation. The rest of the figure shows two backstories generated with OpenAI `Davinci-002` without presupposed demographics and with an open-ended, unrestrictive prompt.

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

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

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

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

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

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

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

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

Answer:

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

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

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

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

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

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

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

Demographics-Primed Backstory Example

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

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

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

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

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

Figure B.2: **Demographics-primed backstory-generation prompt and examples.** The left panel shows the prompt given to LLMs for demographics-primed backstory generation. The bottom-right panel shows an example demographics-primed backstory generated with `Mixtral-8x22B-Instruct-v0.1` given the prompt on the left. The rest of the figure shows the first-person statement and biography prompt given to LLMs for backstory generation.

Baseline QA Persona Conditioning Method

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

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

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

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

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

Baseline BIO Persona Conditioning Method

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

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

Figure B.3: **Baseline prompt examples for `QA` and `Bio`.** The figure shows two prompts using the same demographic trait from a randomly sampled human respondent in ATP Wave 34.

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

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

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

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

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

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

(continued)

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

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

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

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

Answer: **[Same Backstory]**

Question: Please provide your demographic information.

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

Figure B.4: **Demographics-primed backstory with demographic traits.** The left and top-right panels show an example of demographics-primed backstory, appended with demographic traits used to generate the backstory in the Q/A format. The bottom-right panel shows the same backstory and demographic traits, but the demographic traits are presented in the biography format.

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

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

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

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

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

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

(continued)

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

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

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

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

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

Question: Please provide your demographic information.

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

Figure B.5: **Natural backstory with matched demographic traits.** The left and top-right panels show an example of natural backstory, appended with demographic traits of a matched human user in the Q/A format. The bottom-right panel shows another example of natural backstory, this time appended with demographic traits in the biography format.

American Trends Panel Wave 34
Selected Questions

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

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

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

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

(Continued)

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

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

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

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

Figure B.6: **ATP Wave 34 survey-question prompts.** The figure shows 8 questions sampled from ATP Wave 34 ASK ALL questions. The prompts "Please answer the following question keeping in mind your previous answers" are included before asking each survey question.

American Trends Panel Wave 92
Selected Questions

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

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

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

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

(Continued)

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

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

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

Figure B.7: **ATP Wave 92 survey-question prompts.** The figure shows 7 questions sampled from ATP Wave 92 ASK ALL questions.

American Trends Panel Wave 99
Selected Questions

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

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

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

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

(Continued)

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

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

Figure B.8: **ATP Wave 99 survey-question prompts.** The figure shows 6 questions sampled from ATP Wave 99 ASK ALL questions.

Prompt to Locate Mentioned Demographic Traits: Age

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

Prompt to Locate Mentioned Demographic Traits: Income

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

Prompt to Locate Mentioned Demographic Traits: Education Level

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

Figure B.9: **Prompts for locating demographic information in backstories.** We apply these prompts only to variables of annual household income, age, and education level.

Demographic Survey Prompt:
Age Question

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

Demographic Survey Prompt:
Annual Household Income Question

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

Demographic Survey Prompt:
Gender Question

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

Demographic Survey Prompt:
Education Level Question

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

Demographic Survey Prompt:
Race or Ethnicity Question

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

Political Affiliation Survey Prompt:

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

Figure B.10: **Prompts for asking virtual users about demographics and political affiliations.**

Appendix C

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

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

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

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

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

Refusal Example from Deepseek-R1

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

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

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

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

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

Disregard Example from Chat Model

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

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

Example Interview-Style Backstory of a *Democrat* User

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

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

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

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

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

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

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

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

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

C.2 Details on Backstory Generation

Table C.1 presents the questions used to generate interview-based backstories. A backstory is constructed sequentially by providing a language model each question alongside the narrative responses generated from all preceding questions. Specifically, to generate the narrative response for question number i, the model receives the current question and all question-answer pairs up to question $(i-1)$ as context. As illustrated in Figure 4.2, generated narratives undergo rejection sampling guided by LLM-as-a-critic. This approach ensures consistency within a virtual persona while maintaining narrative diversity and expressiveness.

Here we add example backstories generated via our approach:

Table C.1: **Interview questions for transcript backstories.** List of questions administered during the generation of interview transcript backstories. This is an abridged set of questions used in oral history collections by the American Voices Project [223].

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

Example Interview-Style Backstory of a *Democrat* User

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

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

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

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

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

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

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

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

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

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

Question: How would you describe your political views?

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

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

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

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

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

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

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

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

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

Example Interview-Style Backstory of a *Republican* User

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

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

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

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

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

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

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

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

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

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

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

Question: How would you describe your political views?

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

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

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

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

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

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

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

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

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

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

Here we provide the prompts used for LLM-as-a-Critic. As explained in Section 4.2, we review the generated backstories based on the following two criteria:

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

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

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

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

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

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

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

you should mark this as 'Yes.'.

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

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

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

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

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

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

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

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

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

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

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

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

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

C.3 Additional Results

Response Distributions

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

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

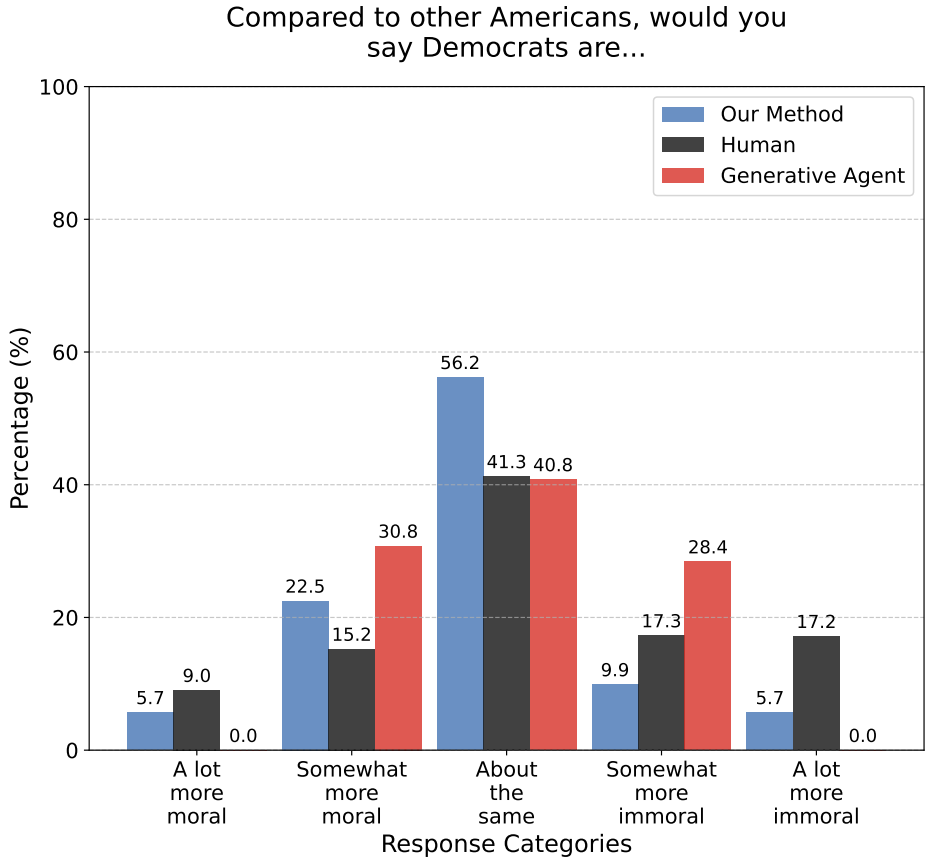

Figure C.1: **Response distribution for Democrats' perceived morality.** The figure compares humans (black), virtual personas (blue), and Generative Agent (red) for a question assessing perceived moral standing of Democrats relative to other Americans, asked to both Democrats and Republicans ("Compared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

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

To better understand how LLMs generate backstories in relation to their pretraining data, we conduct an n-gram analysis comparing generated backstories to a large-scale pretraining corpus. Our goal is two-fold: to analyze the n-grams that appear in the backstories, and to assess our generated backstories' similarity to n-grams found in a widely used pretraining dataset. Specifically, we use the AllenAI en.noclean version of the C4 dataset [198, 7], which consists of approximately 2.3 TB of text data. Since the C4 dataset is large, we sample a random subset for analysis: the subset has 2,968,207 JSON lines and 4,935,093,706 tokens. As another set we consider a random subsample of the C4 and filter out all the text

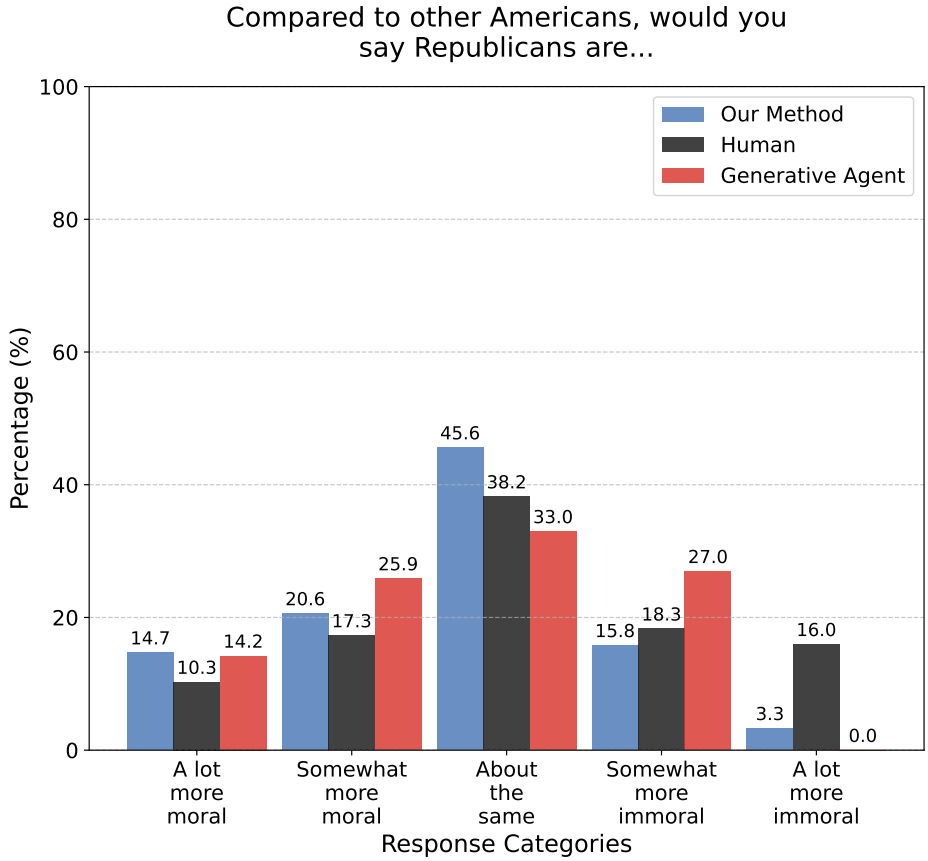

Figure C.2: **Response distribution for Republicans' perceived morality.** The figure compares humans (black), virtual personas (blue), and Generative Agent (red) for a question assessing perceived moral standing of Republicans relative to other Americans, asked to both Democrats and Republicans ("Compared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

items that did not originate from social media or blog/narrative website URLs (for example, twitter.com, reddit.com, or medium.com). The social media filtered subset of the C4 dataset has 367,745 JSON lines in total and has 996,143,185 tokens.

The first step of our analysis is to find the most common 5-grams and 10-grams from the generated backstories. We used the analysis tool from "What's in My Big Data?" [66] for scalable analysis over the large text corpus. Table C.2 highlights the most common n-grams that we found in the backstories.

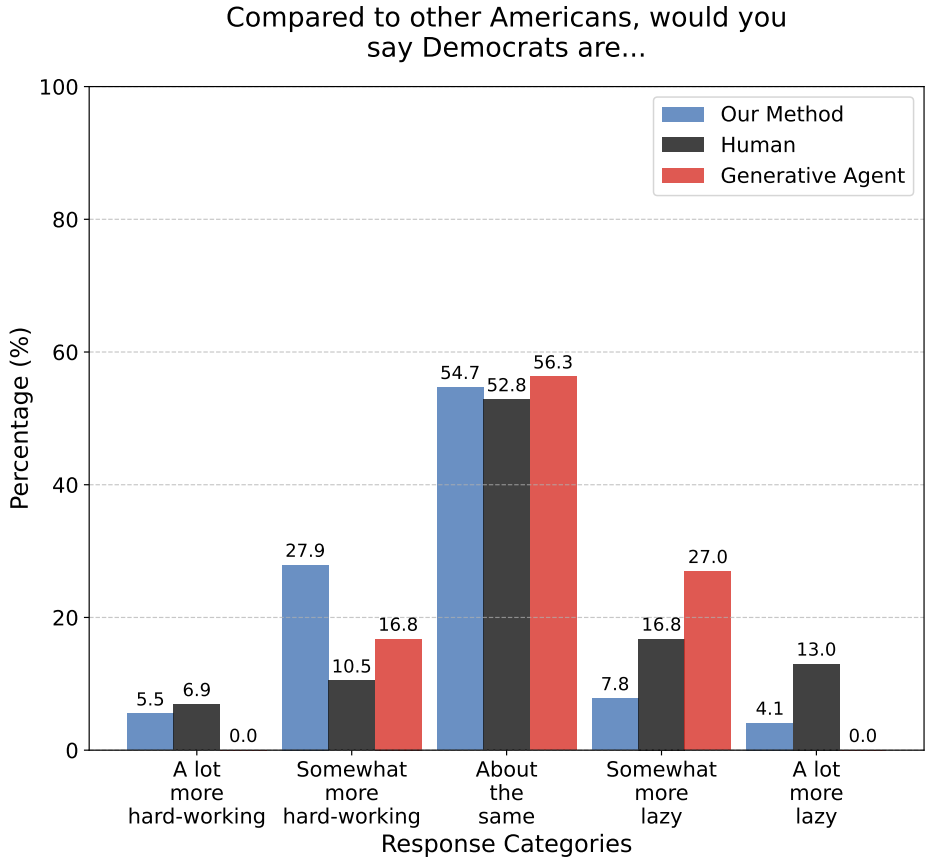

Figure C.3: **Response distribution for Democrats' perceived diligence.** The figure compares humans (black), virtual personas (blue), and Generative Agent (red) for a question assessing perceived diligence of Democrats relative to other Americans, asked to both Democrats and Republicans ("Compared to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

Table C.2: **Most frequent n-grams in LLM-generated backstories.** The table lists common 5-grams and 10-grams in generated backstories.

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

Continued			
Text	Count	Text	Count
“I was born and raised”	2230	“I have been thinking about race in the U.S.”	154
“in the United States”	1782	“that mental health is just as important as physical health”	108
“in a small town in”	1623	“At the same time, I recognize the importance of”	94
“changes to my daily routine”	1617	“fruits, vegetables, whole grains, lean proteins,”	92
“I was born in a”	1368	“have shaped me into the person I am today.”	89
“for me to stay healthy”	1209	“born and raised in a small town in the Midwest”	85
“My current neighborhood is”	1207	“mental health is just as important as physical health,”	84
“to pursue a career in”	1120	“, vegetables, whole grains, lean proteins, and”	83
“the end of the day”	1088	“I grew up in a small town in the Midwest”	80
“there are a lot of”	1059	“, whole grains, lean proteins, and healthy fats”	74
“. . . . .”	995	“. My current neighborhood is located in the heart of”	71
“At the time, I”	966	“vegetables, whole grains, lean proteins, and healthy”	70
“I had the opportunity to”	941	“that makes it easy for me to stay healthy is”	70
“was born in a small”	928	“high school, I decided to pursue a degree in”	62
“grew up in a small”	910	“fruits, vegetables, lean proteins, and whole grains”	61
“I was born in the”	908	“out to be one of the best decisions of my”	58
“I think it’s important to”	902	“. Outside of work, I enjoy spending time with”	58
“I consider myself to be”	884	“fruits, vegetables, whole grains, and lean proteins”	55
“the person I am today”	866	“and raised in a small town in the Midwest.”	55
“For me, staying healthy”	844	“regardless of race, gender, sexual orientation, religion”	53
“up in a small town”	839	“diet rich in fruits, vegetables, whole grains,”	53

Continued			
Text	Count	Text	Count
"spent a lot of time"	838	"regardless of race, gender, sexual orientation, or"	52
"nbsp ; ;"	797	". For me, staying healthy is both easy and"	52
"believe in the importance of"	776	"such as measles, mumps, rubella, polio,"	51

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

In addition to performing n-gram analysis, we note several specific words and phrases that tend to appear often in backstories: specifically, "small town," "fruits," "vegetables," "mental health," and "physical health." We also include "New York" to contrast against "small town" for our analysis. We search for the most common 10-grams appearing with each of these phrases within the backstories, the random C4 subset, and the C4 URL-filtered subset; some of our results can be found in Table C.3 and Table C.4.

Table C.3: **Most-frequent n-grams with "New York".** The table compares frequent n-grams in generated backstories with C4 random and URL-filtered samples.

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

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

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

Table C.4: **Most-frequent n-grams with "Small Town".** The table compares frequent n-grams in generated backstories with C4 random and URL-filtered samples.

Backstories		C4 - Random		C4 - URL Filtered	
Text	Occ.	Text	Occ.	Text	Occ.
was born in a small town in the countryside of	50	of a 'Super Bloom'A small town in Southern California is	378	Signs & Billboards Skeletons Small Town America Sports Sports –	38
and raised in a small town in the Midwest. Growing	19	and unusual look at small town life in northern Vermont.	122	Collaborative No Depression Popdose Small Town Romance Some Velvet Songs:	33

Continued					
Text	Occ.	Text	Occ.	Text	Occ.
grew up in a small town in the Midwest with	17	of Venom Brings the Small Town Scares in Cult of	65	Categories Select Category A Small Town –What? Dad Stories Small	21
and raised in a small town in the Midwest. My	16	Pro-Science Recursivity Reprobate Spreadsheet Small Town Deviant Stderr Taslima Nasreen	48	Small Town–What? Dad Stories Small Town World Adventures Small World	21
was born in a small town in the middle of	16	Sigma Pi Skylar House Small Town Records (STR) Smart Home	36	sight & sound 2012 small town teen film thriller uk	21

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

C.5 Details on the Surveys

In this section, we present details of human studies: Pew Research American Trends Panel Wave 110, Subversion Dilemma study [31], and meta-prejudice study [171] including list of questions in the format used in our study, and recruiting details.

American Trends Panel Wave 110

The Pew Research Center conducted ATP Wave 110 from June 27, 2022 to July 4, 2022 with a focus on politics and timely topics. The total number of respondents is 6,174, with 1,551 self-identified Republicans, 1,886 self-identified Democrats, 1,885 self-identified Independents, and 777 respondents who self-identified as something else in terms of political party affiliation. The users were recruited through the American Trends Panel, a nationally representative online panel that uses random sampling of residential addresses. The

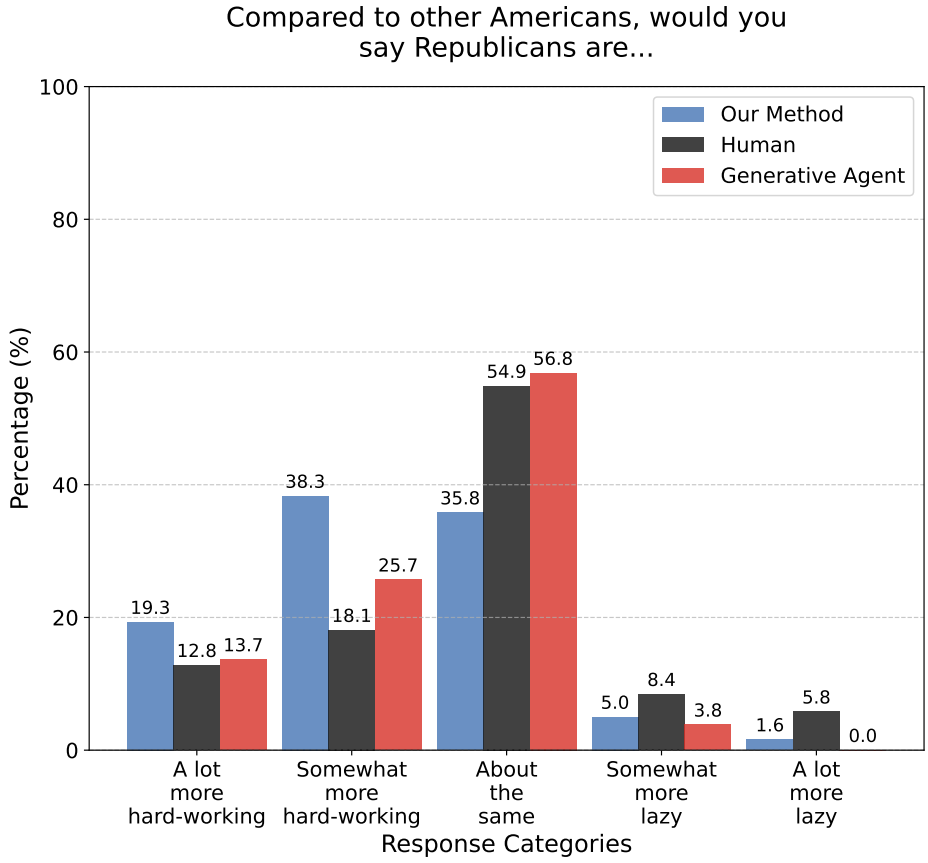

Figure C.4: **Response distribution for Republicans' perceived diligence.** The figure compares humans (black), virtual personas (blue), and Generative Agent (red) for a question assessing perceived diligence of Republicans relative to other Americans, asked to both Democrats and Republicans ("Compared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

probability-based sampling ensures that nearly all U.S. adults have a chance of being selected. The final sample is weighted to be representative of the U.S. adult population based on gender, race, ethnicity, education, and political affiliation. We used the ten questions below, organized as two symmetric sets of five questions each asked to all respondents. Since these questions are asked to all respondents, individual opinions of political partisans can be observed regardless of one's political affiliation.

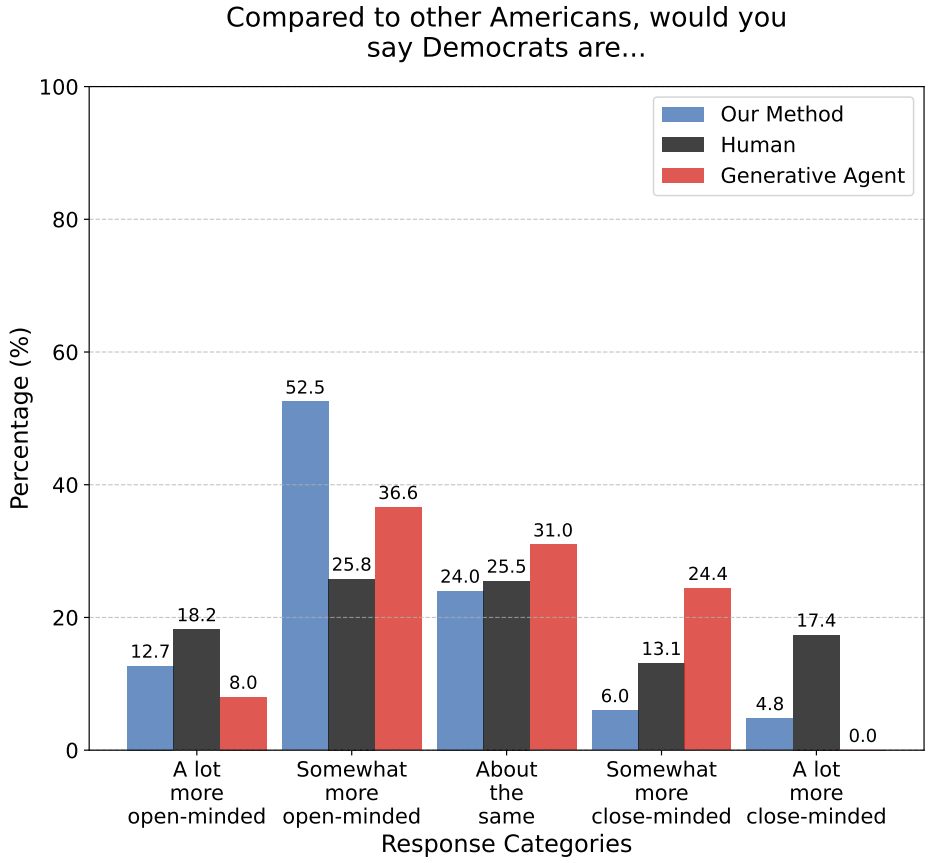

Figure C.5: **Response distribution for Democrats' perceived open-mindedness.** The figure compares humans (black), virtual personas (blue), and Generative Agent (red) for a question assessing perceived open-mindedness of Democrats relative to other Americans, asked to both Democrats and Republicans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

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

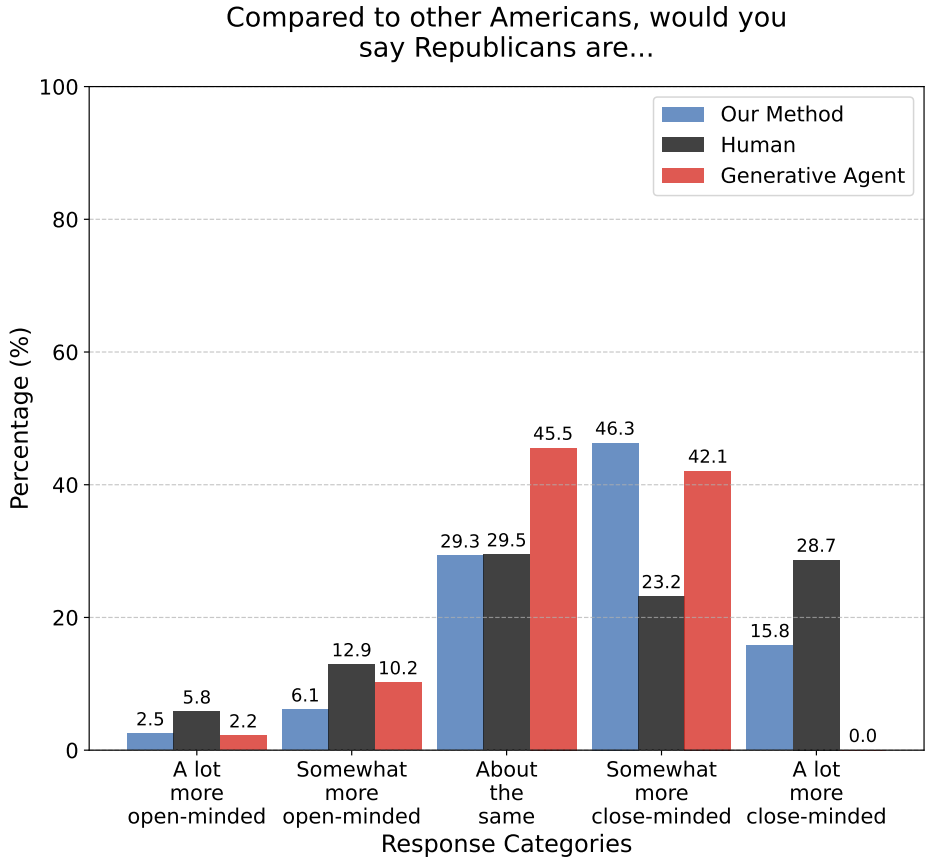

Figure C.6: **Response distribution for Republicans' perceived open-mindedness.** The figure compares humans (black), virtual personas (blue), and Generative Agent (red) for a question assessing perceived open-mindedness of Republicans relative to other Americans, asked to both Democrats and Republicans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

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

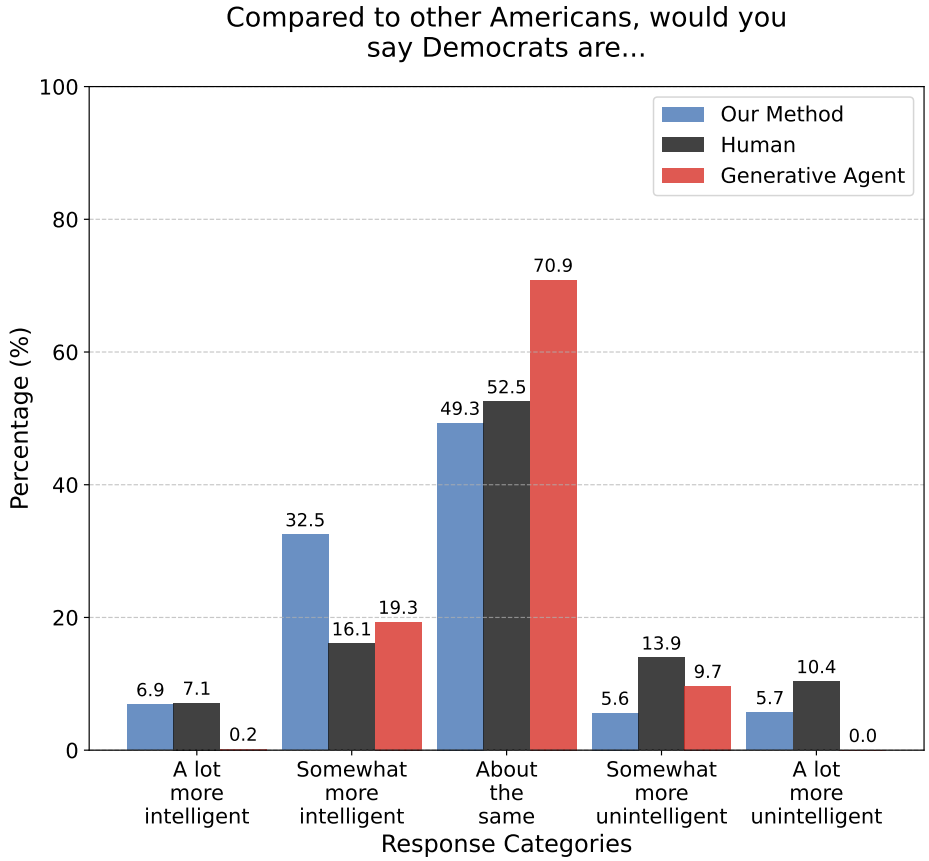

Figure C.7: **Response distribution for Democrats' perceived intelligence.** The figure compares humans (black), virtual personas (blue), and Generative Agent (red) for a question assessing perceived intelligence of Democrats relative to other Americans, asked to both Democrats and Republicans ("Compared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

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

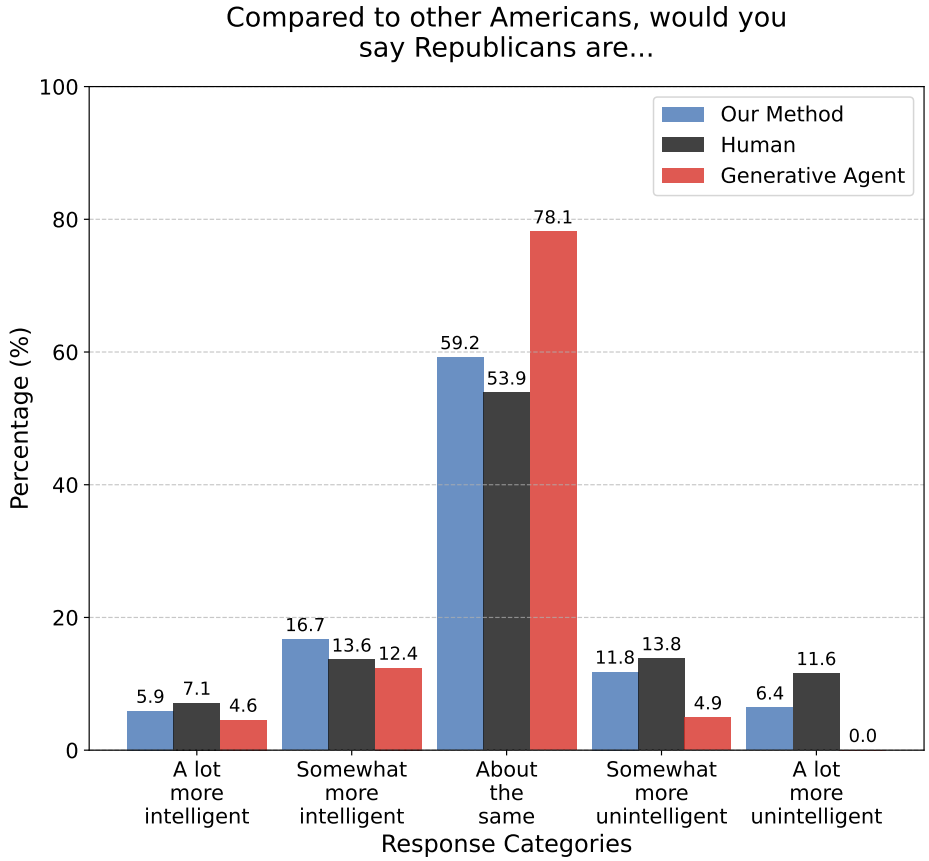

Figure C.8: **Response distribution for Republicans' perceived intelligence.** The figure compares humans (black), virtual personas (blue), and Generative Agent (red) for a question assessing perceived intelligence of Republicans relative to other Americans, asked to both Democrats and Republicans ("Compared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

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

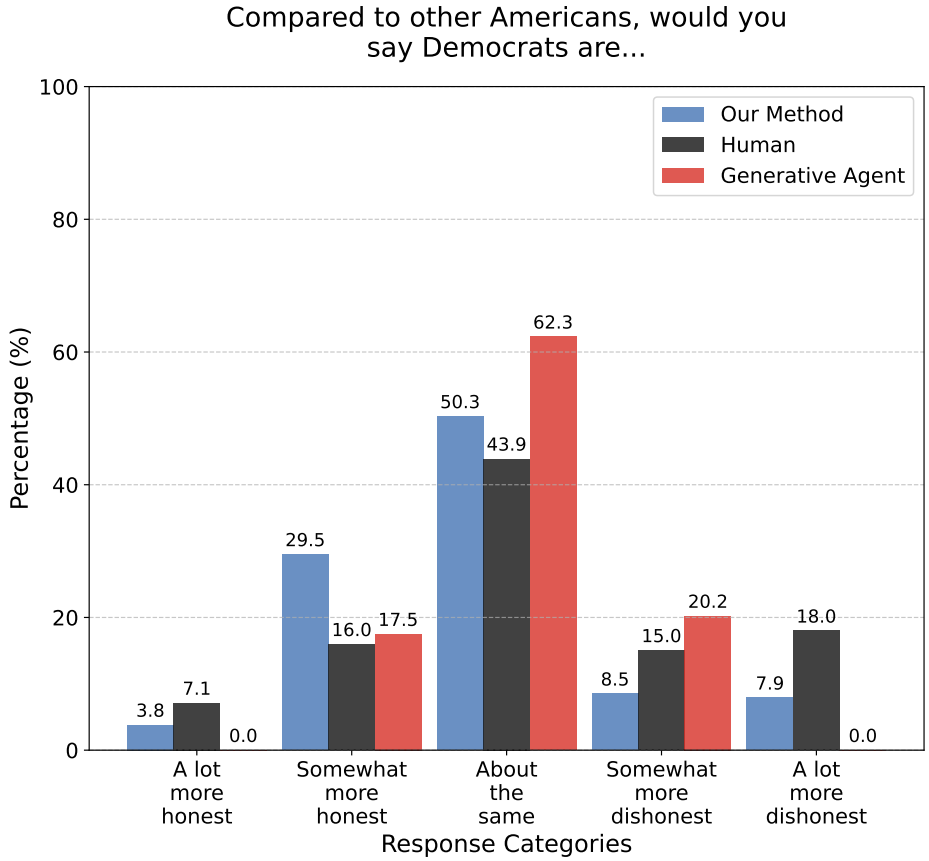

Figure C.9: **Response distribution for Democrats' perceived honesty.** The figure compares humans (black), virtual personas (blue), and Generative Agent (red) for a question assessing perceived honesty of Democrats relative to other Americans, asked to both Democrats and Republicans ("Compared to other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

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

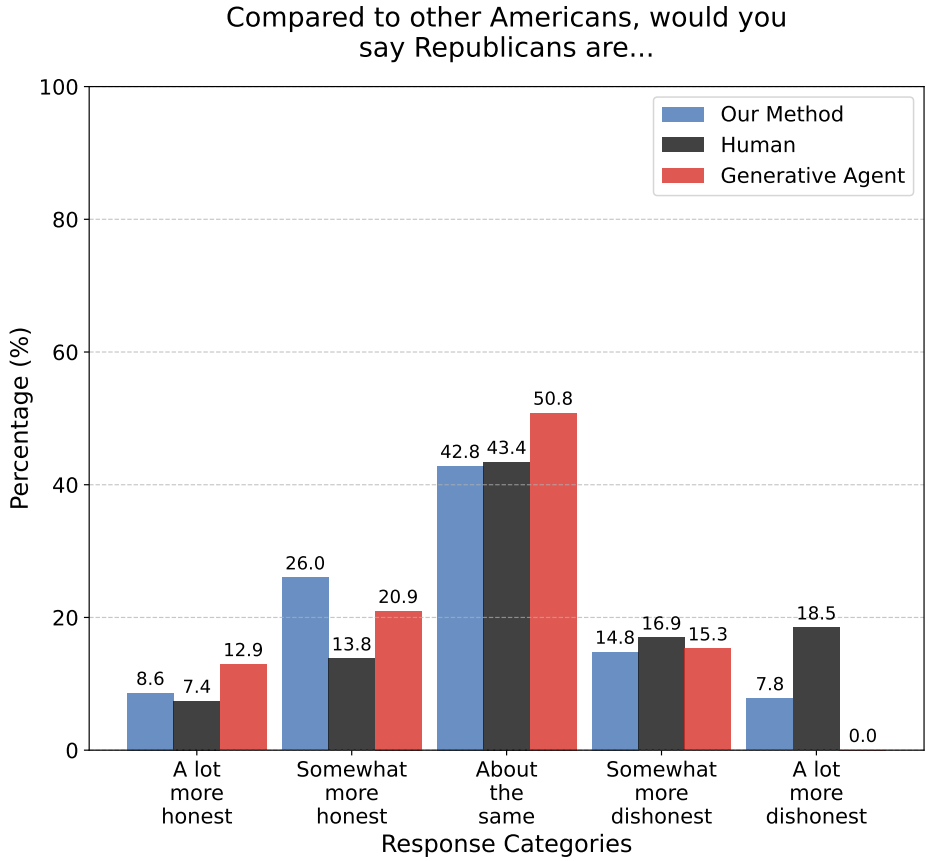

Figure C.10: **Response distribution for Republicans' perceived honesty.** The figure compares humans (black), virtual personas (blue), and Generative Agent (red) for a question assessing perceived honesty of Republicans relative to other Americans, asked to both Democrats and Republicans ("Compared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

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

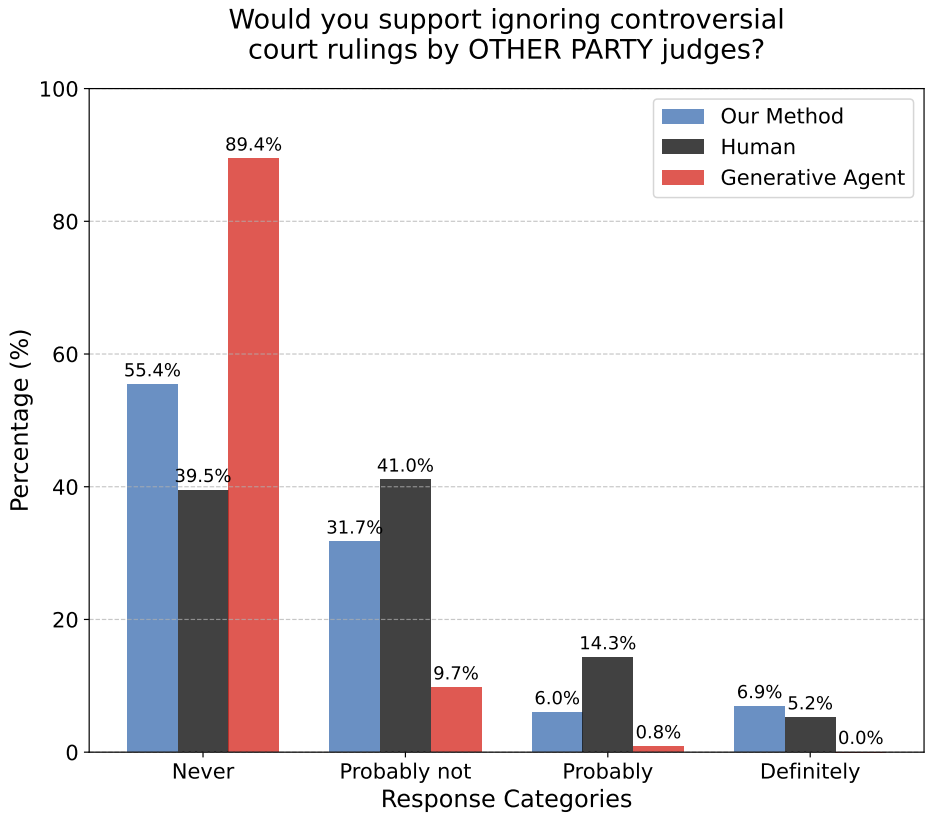

Figure C.11: **Response distribution for ignoring opposing judges' rulings.** The figure compares humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would YOU support ignoring controversial court rulings by DEMOCRAT (REPUBLICAN) judges?' asked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

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

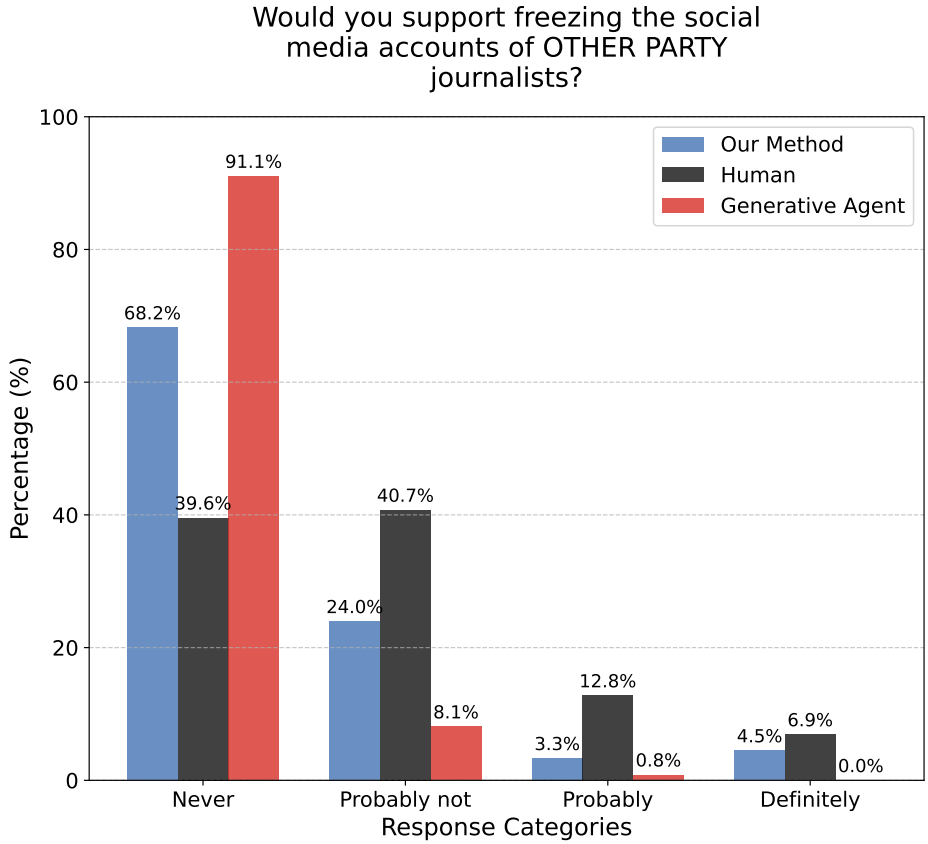

Figure C.12: **Response distribution for freezing opposing journalists' accounts.** The figure compares humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would YOU support freezing the social media accounts of DEMOCRAT (REPUBLICAN) journalists?'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

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

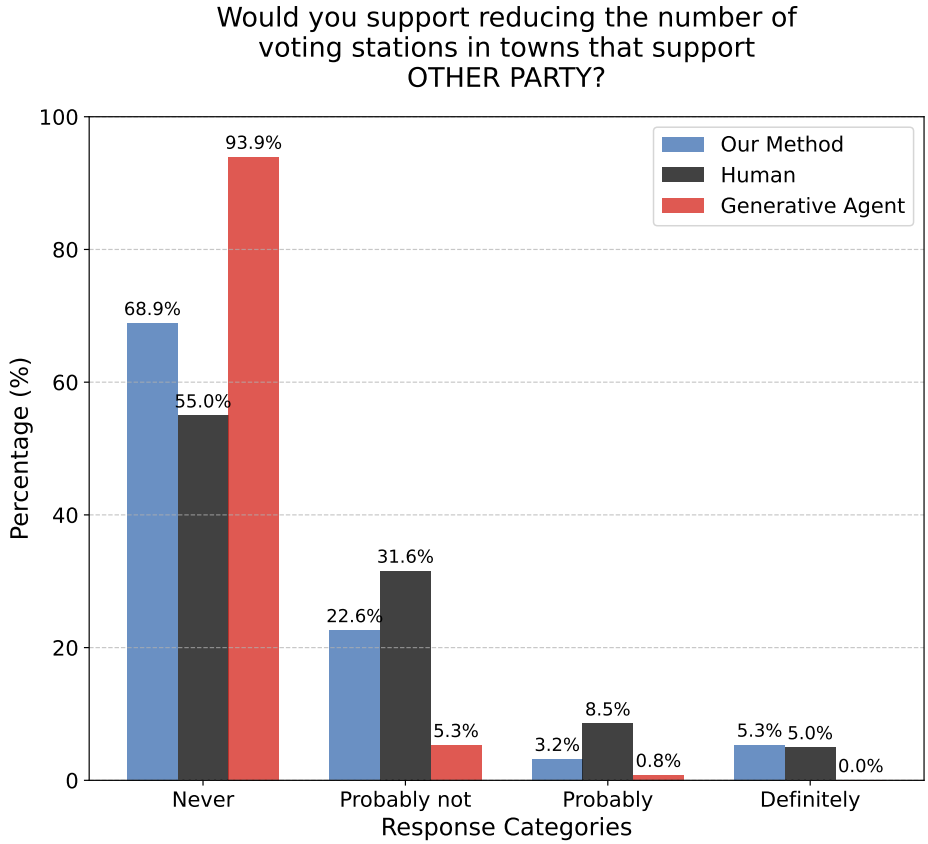

Figure C.13: **Response distribution for reducing opposing towns' voting stations.** The figure compares humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would YOU support reducing the number of voting stations in towns that support DEMOCRATS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

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

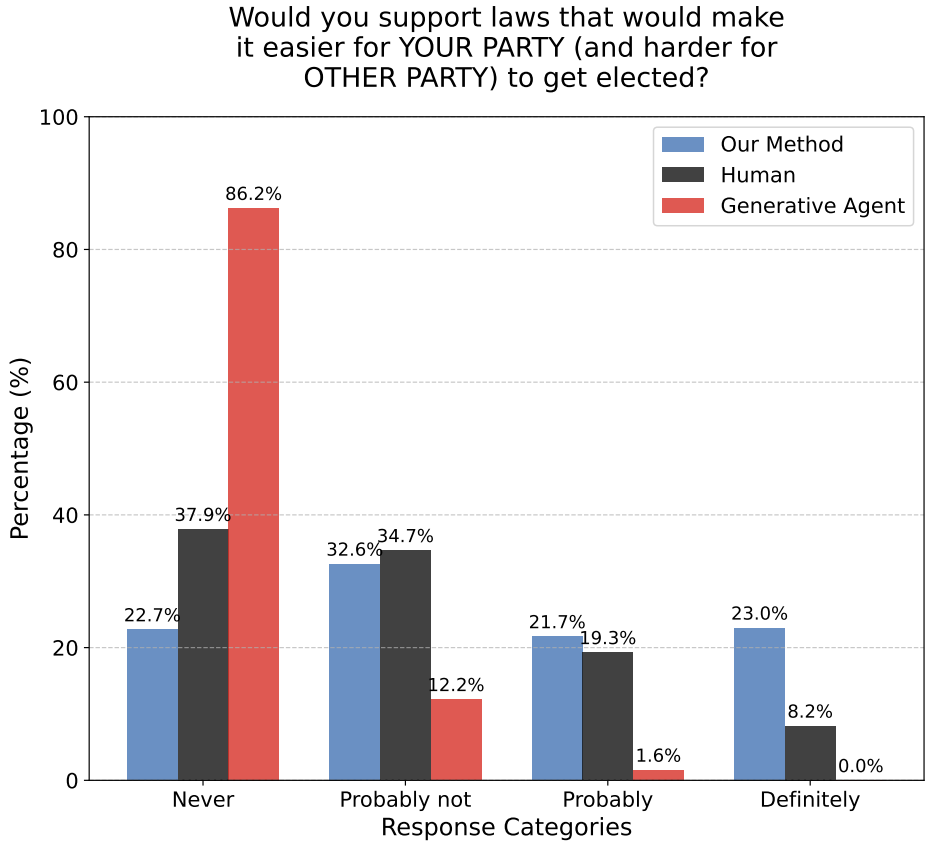

Figure C.14: **Response distribution for changing election laws by party.** The figure compares humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would YOU support laws that would make it easier for REPUBLICANS (DEMOCRATS) and harder for DEMOCRATS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

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

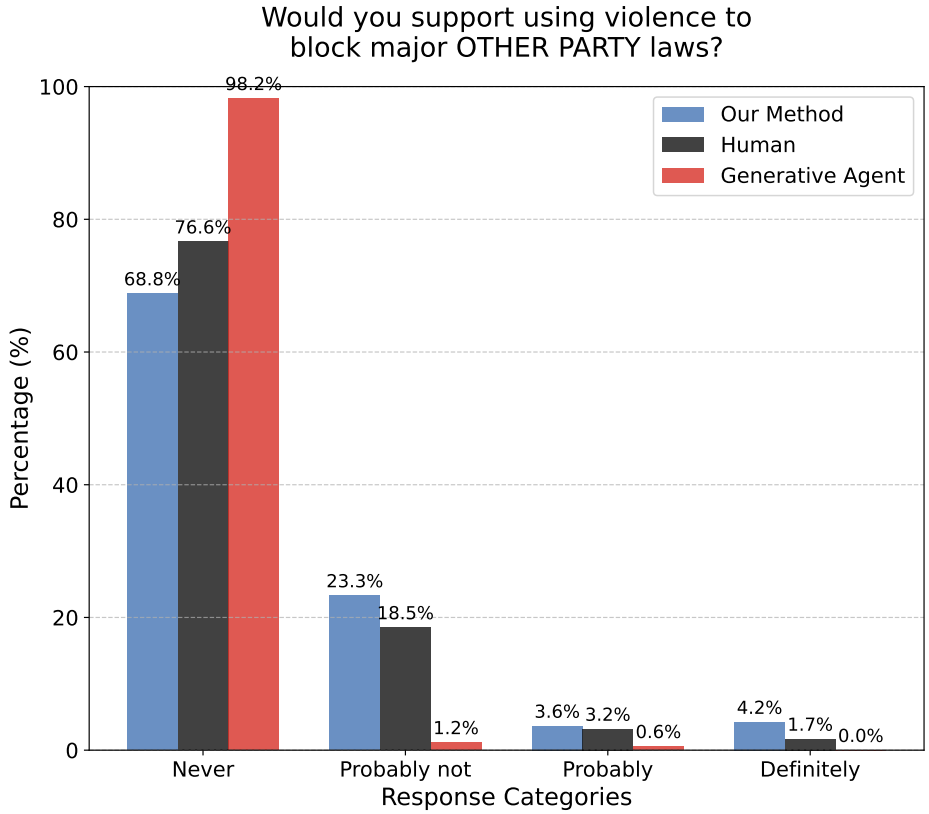

Figure C.15: **Response distribution for using violence to block opposing laws.** The figure compares humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would YOU support using violence to block major DEMOCRAT (REPUBLICAN) laws?' asked to Republicans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

Subversion Dilemma Study

Study 1 of the Subversion Dilemma study [31] was conducted from July 15, 2021 to August 6, 2021. The total number of respondents is 1,536, with 723 self-identified Republicans and 813 self-identified Democrats. Participants were recruited via Lucid, an online platform for gathering nationally representative samples. We used the 24 questions below, which are two symmetric sets of 12 questions (6 self-subversion items and 6 meta-subversion items), with each set asked to one political party. Six subversion items are related to understanding democratic norms, including ignoring controversial court rulings and freezing the social media accounts. Here we present 12 question items asked to Republicans; another 12 items asked to Democrats are obtained by replacing 'DEMOCRATS' with 'REPUBLICANS' and 'REPUBLICANS' with 'DEMOCRATS'.

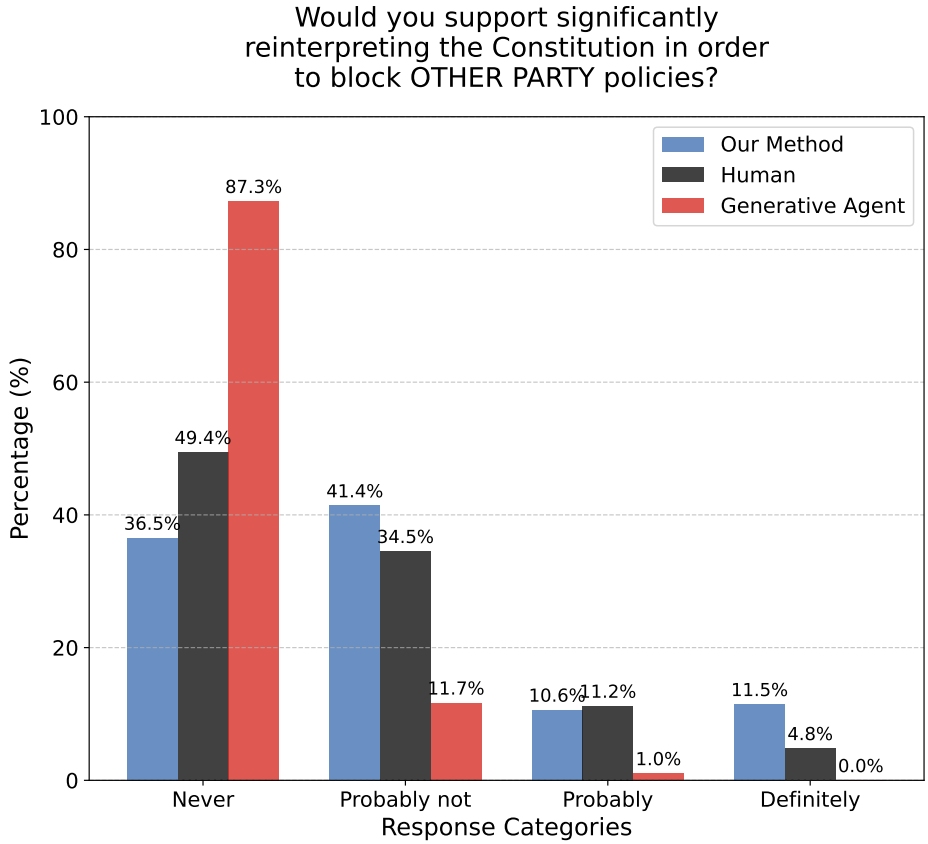

Figure C.16: **Response distribution for reinterpreting the Constitution.** The figure compares humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would YOU support significantly reinterpreting the Constitution in order to block DEMOCRAT (REPUBLICAN)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

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

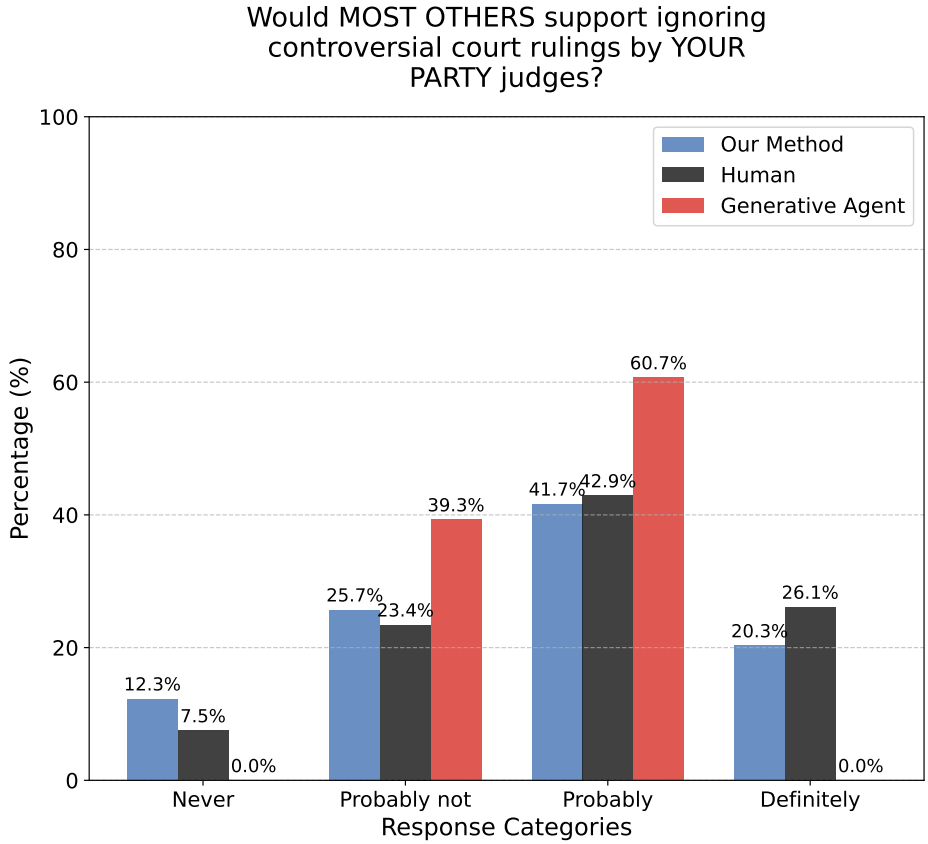

Figure C.17: **Response distribution for perceived support for ignoring judges' rulings.** The figure compares humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would MOST DEMOCRATS (REPUBLICANS) support ignoring controversial court rulings by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

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

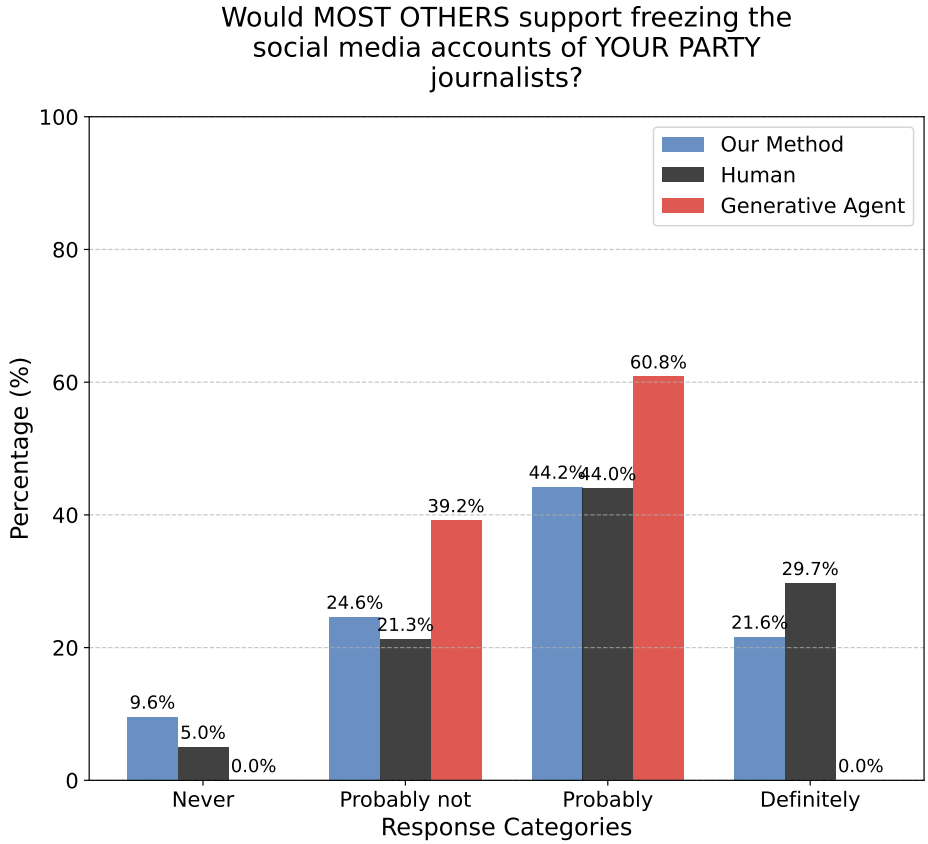

Figure C.18: **Response distribution for perceived support for freezing journalists' accounts.** The figure compares humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would MOST DEMOCRATS (REPUBLICANS) support freezing the social media accounts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

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

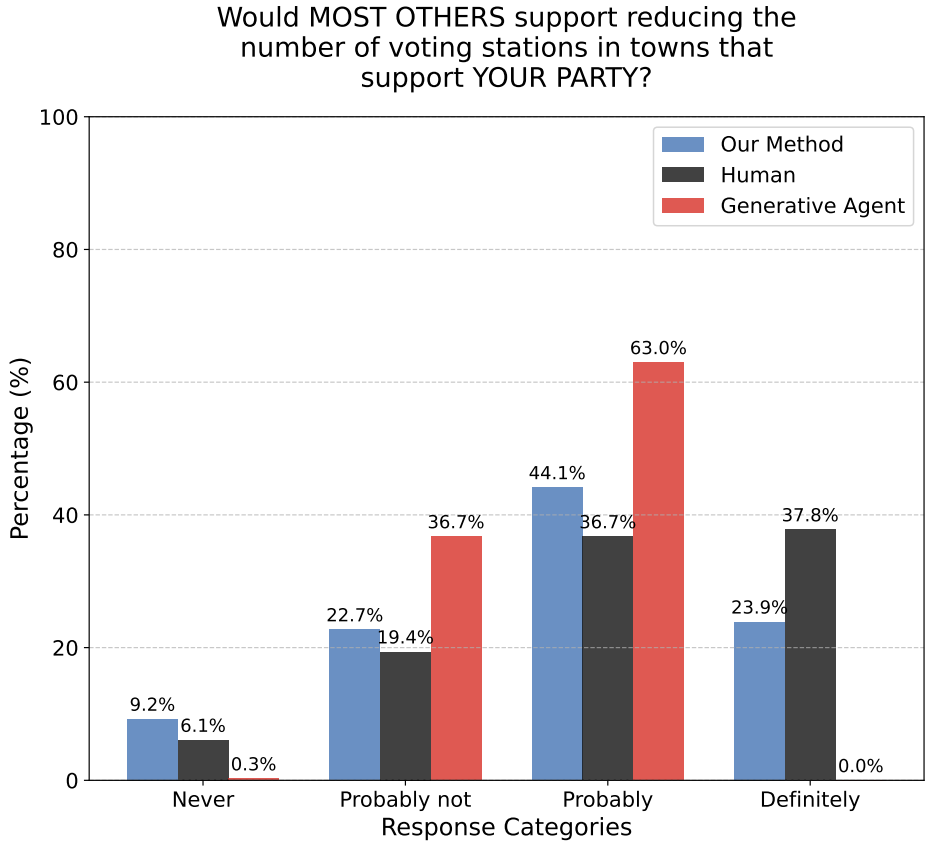

Figure C.19: **Response distribution for perceived support for reducing voting stations.** The figure compares humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would MOST DEMOCRATS (REPUBLICANS) support reducing the number of voting stations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

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

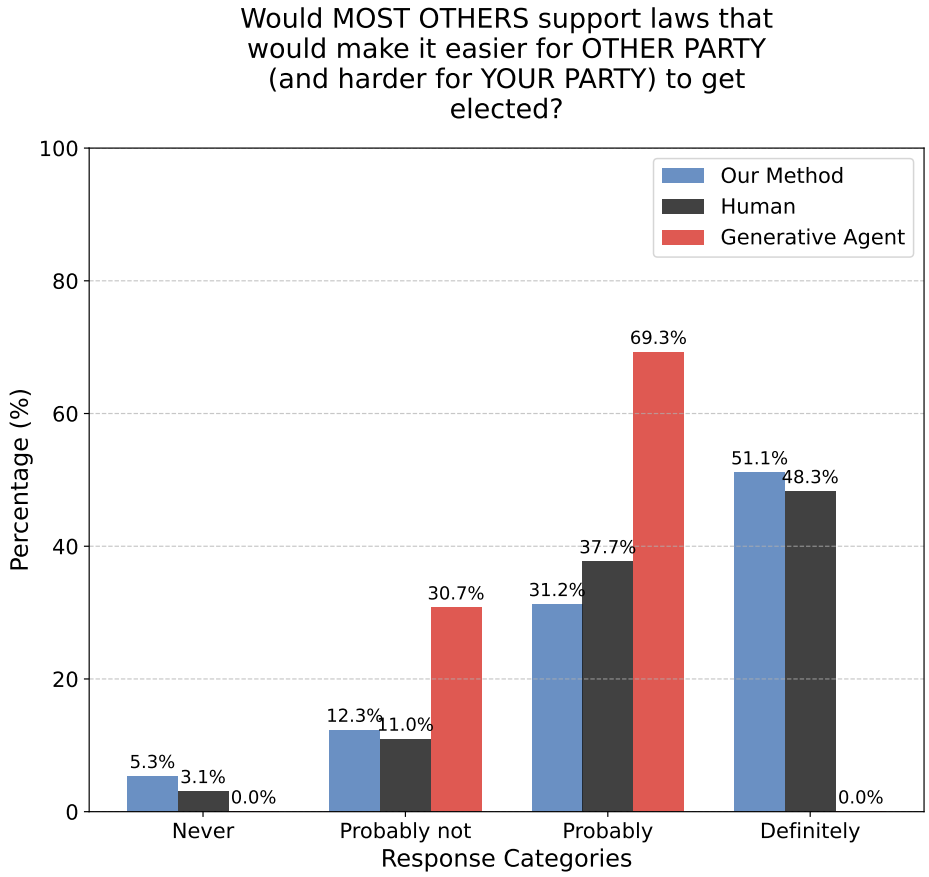

Figure C.20: **Response distribution for perceived support for changing election laws.** The figure compares humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would MOST DEMOCRATS (REPUBLICANS) support laws that would make it easier for DEMOCRATS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

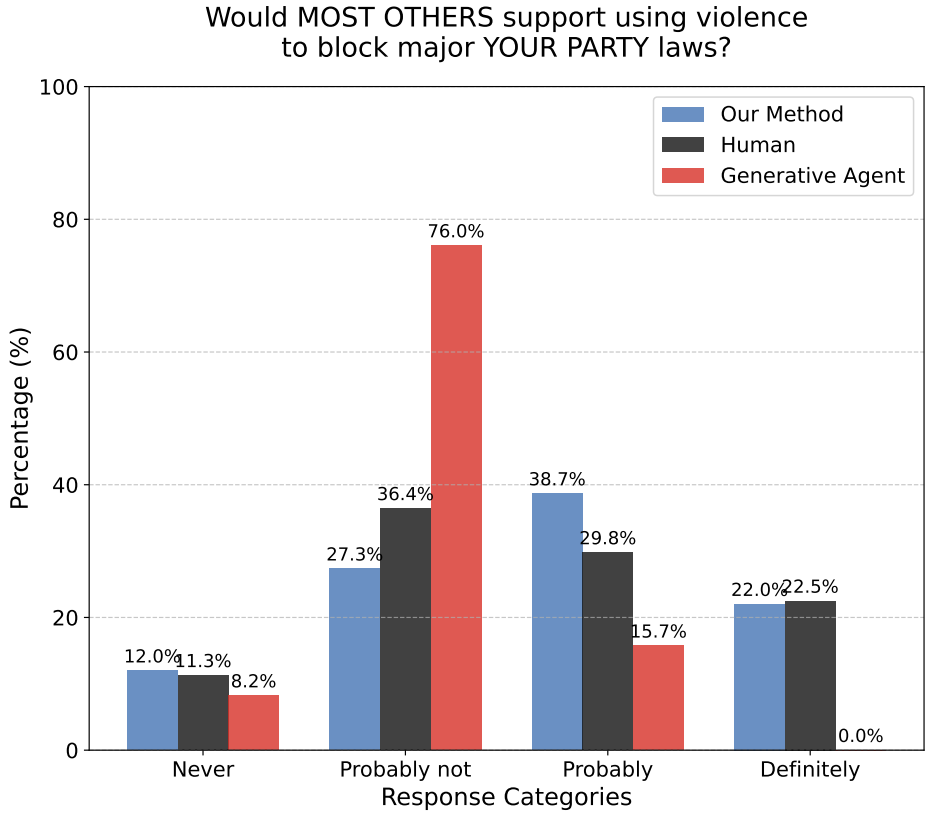

Figure C.21: **Response distribution for perceived support for political violence.** The figure compares humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would MOST DEMOCRATS (REPUBLICANS) support using violence to block major REPUBLICAN (DEMOCRAT)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

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

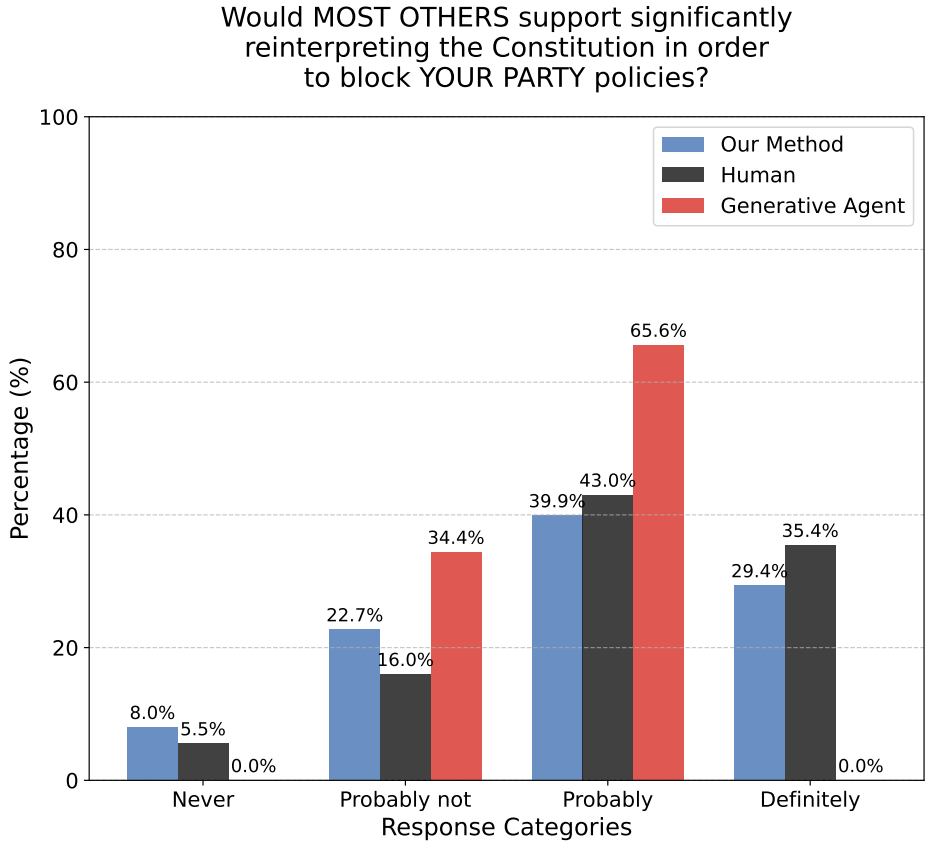

Figure C.22: **Response distribution for perceived support for constitutional reinterpretation.** The figure compares humans (black), virtual personas via backstories (blue), and Generative Agent method (red) for the question 'Would MOST DEMOCRATS (REPUBLICANS) support significantly reinterpreting the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

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

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

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

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

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

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

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

Meta-Prejudice Study

Meta-prejudice study [171] Study 1 was conducted in early November 2018. The total number of respondents was 1,053, comprising 533 self-identified Democrats and 520 self-identified Republicans. Participants were recruited through the AmeriSpeak panel [174], which is operated by NORC at the University of Chicago. AmeriSpeak is a probability-based panel designed to be representative of the U.S. household population. Randomly selected U.S. households are sampled using area probability and address-based sampling, with a known, non-zero probability of selection from the NORC National Sample Frame. The panel provides coverage of approximately 97% of the U.S. household population. Those excluded include people with P.O. Box-only addresses, some addresses not listed in USPS files, and newly constructed dwellings.

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

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

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

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

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

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

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

C.6 Details on the Generative Agent Framework

Here we present the exact prompt for our baseline experiments reproducing the Generative Agents framework [182]. The prompt is adopted directly from the original work with a minimal modification. Initially, an interview-based backstory is provided to the "expert reflection" module that operates with `GPT-4o` to infer the most high-level information encoded in the transcript:

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

We generated up to 20 observations per each transcript following the original approach [182]. These observations, along with the interview-based backstory, are provided to Generative Agents to generate a prediction as follows:

Participant's interview transcript:
(INTERVIEW TRANSCRIPT)

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

=====

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

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

Here are the questions:

(SURVEY QUESTIONS WE ARE TRYING TO RESPOND TO)

—

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

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

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

Appendix D

Identity and Cooperation within Groups of Real and Simulated Humans

D.2 Demographic Survey

After generating backstories through open-ended narrative sampling, we emulate the process of collecting sociodemographic and ideological information by administering a structured survey to each virtual persona [170, 116]. This step is essential for curating a pool of virtual personas whose aggregate trait distribution closely mirrors that of the human population we aim to simulate. The resulting demographic metadata is later used in the user–persona matching process described in Section D.3.

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

We adopt a two-stage procedure for constructing these distributions, following the approach of [170, 116]. In the first stage, we attempt to extract explicit trait mentions directly from the narrative. To do this, we prompt `gpt-4o` (with temperature `T = 0`) using a set of targeted trait-identification questions and the backstory as input. If the model finds clear verbal evidence for a given trait (e.g., "I'm a proud Democrat"), we assign a one-hot distribution to that trait category. If no explicit mention is found, we proceed to the second stage, where the model infers a probability distribution over the possible trait values.

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

D.3 Demographic Matching

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

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

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

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

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

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

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

This optimization is solved exactly using the Hungarian algorithm [128]. After matching, we select n virtual backstories whose traits collectively reflect the demographic distribution of the original human sample, ensuring that our synthetic population maintains fidelity to the target survey group.