The Illusion of Intervention: Your LLM-Simulated Experiment is an Observational Study

Victoria Lin [1] **Taedong Yun** [2] **Maja Matarić** [2] **John Canny** [2] **Arthur Gretton** [2] **Alexander D'Amour** [2]

Abstract

Large language models (LLMs) show potential as simulators of human behavior, offering a scalable way to study responses to interventions. However, because LLMs are trained largely on observational data, interventions in experiments with LLM-simulated synthetic users can induce unintended shifts in latent user attributes. This causes user drift where the implicit simulated population differs across treatment conditions, potentially distorting effect estimates. We formalize the confounding or *selection bias* that can arise due to user drift and show how intervention-dependent shifts can inflate or attenuate observed differences in user responses under intervention. To diagnose confounding, we propose using negative control outcomes—attributes that should remain invariant under intervention—to identify distribution shifts across intervention conditions, providing evidence of user drift. To mitigate drift, we study adjusting the persona specification by eliciting additional confounders, finding that targeted, setting-relevant confounders can substantially reduce bias across survey-style and multi-turn agent evaluations.

1. Introduction

Large language models (LLMs) are emerging as powerful tools for simulating human behavior. Traditionally, a multitude of scientific applications rely on human feedback. Large-scale surveys underpin many social science and public policy studies, and modern machine learning systems are trained and evaluated on human response data. Because collecting such feedback at scale is a major bottleneck, a popular alternative is to use LLMs to simulate users (Manning et al., 2024; Argyle et al., 2023), often by prompting them to embody a persona (Gu et al., 2025; Fröhling et al., 2025). These users can then be used in lieu of human respondents to conduct *synthetic experiments* that explore how different populations might respond to new policies, products, or conversational interactions—for instance, evaluating whether an *intervention* such as an update to an AI agent yields more favorable user outcomes (Li et al., 2025; Kaiser et al., 2025).

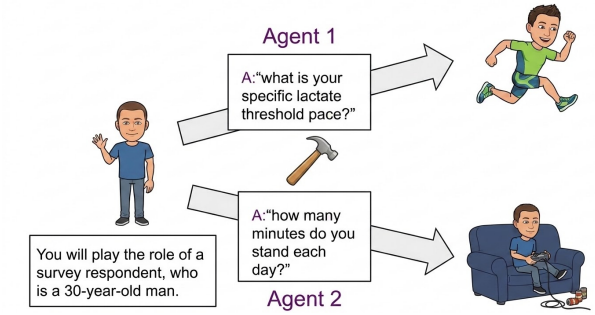


Figure 1. In LLM-simulated experiments, two synthetic users who are instantiated with identical personas can drift to become distinct users in response to different intervention conditions.

In this setting, synthetic users[1] are instantiated by sampling personas from a target user distribution. Each synthetic user is then assigned to interact with each version of an AI agent and provide feedback. Ideally, the synthetic users' responses and derived test statistics serve as valid counterfactuals, reflecting how real users would respond in the same experiment and enabling iterative improvement of the agent (Bai et al., 2022; Lee et al., 2024; Castricato et al., 2025; Yun et al., 2025).

In this paper, we argue that this vision is overly optimistic due to *user drift* in LLM simulations. While synthetic experiments are designed to mimic randomized controlled

Work done while author V. Lin was an intern at Google DeepMind. [1]Carnegie Mellon University [2]Google DeepMind. Correspondence to: Victoria Lin <victoria@stat.cmu.edu>, Alexander D'Amour <alexdamour@google.com>.

Accepted to the ICML 2026 Workshop on Statistical Frameworks for Uncertainty in Agentic Systems, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

[1]We use "synthetic user" to mean an LLM initialized to simulate a user with a short persona prompt (e.g., you are a 30-year-old man). This defines a distribution over users, not a single user, and we operate at this level in our formal treatment.

trials, LLMs are trained to operate via abductive reasoning: given underspecified prompts, they infer as much unspecified context as they are able (Wright et al., 2024; Lee et al., 2026). Consider a synthetic experiment evaluating two health coaching agents, where the user is initialized as a 30-year-old man (Figure 1). Agent 1 is prompted to discuss the user's exercise habits and physical condition in a more technical manner, while Agent 2 discusses physical activity in more general terms. To generate a coherent conversation, the LLM interacting with Agent 1 might implicitly adopt the latent traits of a serious athlete, while the LLM interacting with Agent 2 might simulate a less active individual.

As a result, two synthetic users who are instantiated with the same persona at the beginning of an experiment can effectively become distinct users by the time their outcomes are measured. Because this drift can occur over persona attributes that (1) would remain fixed in human subjects and (2) are relevant to the primary outcomes of the experiment, it can obfuscate the actual intervention effect and give rise to misleading results. If the synthetic user reports higher satisfaction with Agent 2, it is unclear whether Agent 2 is actually more helpful or if the simulated user has simply drifted into a population that is easier to please.

To diagnose and mitigate this issue, our work makes three contributions:

1. We formalize the bias induced by user drift through the lens of *confounding* in causal inference. We argue that because LLMs are trained largely on observational data, synthetic user experiments are more similar to observational studies than randomized experiments. As a result, differences in conditional user populations given the intervention—a form of *selection bias*—can confound the estimated treatment effect.
2. We introduce an empirical diagnostic for LLM user drift using *negative control outcomes* (Lipsitch et al., 2010): variables pertaining to the user that should remain invariant under intervention, such as the user's race. Differences in negative control outcome distributions across intervention conditions indicate that the latent attributes of the simulated user—which are causally and temporally upstream of the intervention and would remain fixed in a real experiment—do change in response to the intervention. We provide evidence of confounding in simulated users across multiple major families of open-source and proprietary LLMs.
3. We show that user drift and selection bias can be mitigated in simulated users via *confounder adjustment* during user persona specification. We present evidence that selection of confounders that are explicitly relevant to the treatment and outcome of the synthetic experiment helps to reduce selection bias, while adjustment over attributes that are less directly relevant yields inconsistent benefits and can sometimes inadvertently exacerbate selection bias.

Our findings suggest that evaluations with synthetic users need to be treated with care. This work highlights how the generative interface of LLMs can be misleading: while an evaluation may be designed to mimic a randomized controlled trial, the LLM-simulated users employed in the trial can introduce confounding bias.

2. Related Work

LLM user simulation. LLMs are increasingly used to simulate user behavior, and persona prompting has become a particularly popular strategy for generating diverse user perspectives. Simulation approaches include both using short persona prompts containing demographic attributes (Park et al., 2023; Hu & Collier, 2024; Kaiser et al., 2025; Fröhling et al., 2025) and longer, life-story-based persona prompting (Moon et al., 2024; Park et al., 2026), which provides benefits such as "deep binding" (Kang et al., 2025) and more accurate modeling of population variance. Because conversational interfaces are the standard for synthetic user interactions, the majority of the simulation literature evaluates instruction-tuned LLMs. Prior works analyzing these specific simulators have identified distinct challenges, such as mode collapse, sycophancy, and persona instabilities (Anthis et al., 2025; Li et al., 2025; Tosato et al., 2026). Some evidence suggests these challenges are less pronounced in simulations with base pre-trained models (Lyman et al., 2025), though they remain subject to artifacts such as prompt sensitivity (Sclar et al., 2024).

Much of the work on evaluating LLM-simulated users has centered on their *external validity*: how closely the generated ratings or opinion patterns of simulated users match those of their real human counterparts. Recent works develop benchmarks for external validity (Dou et al., 2025), identify failures (Aher et al., 2023; Li et al., 2025; Gao et al., 2025; Neumann et al., 2026), and propose strategies to better align synthetic users with human ones (Kang et al., 2025; Moon et al., 2026; Dong et al., 2024; Guerdan et al., 2026). Our work is complementary to this line of research: it focuses on the *internal validity* of the synthetic user study under intervention. This is a distinct question that specifically examines whether the difference between experimental conditions can be interpreted causally, rather than whether the results of the experiment can be generalized.

Confounding in LLMs. The most similar prior work on this topic is from Gui & Toubia (2023), who noted that GPT-4o-mini (OpenAI et al., 2024) exhibits empirical behaviors consistent with confounding in simulated blinded microeconomic experiments. Our work builds on this empirical study by introducing a formalism for confounding in

LLM user simulations. We further propose distinct, unified approaches to diagnose and adjust for confounding bias in a more general experimental setting.

3. Problem Statement

In real and synthetic experiments, the goal is to estimate the causal effect of an intervention on a user's response. With human subjects, the gold standard for establishing causality is a randomized trial, because the randomized treatment assignment ensures that background factors that might inform user responses are balanced across groups that receive different interventions. This isolates the effect of the intervention on the outcome. Synthetic experiments seek to replicate this property with synthetic users: identically initialized LLMs are exposed to different interventions, with the expectation that this yields a controlled comparison. In this section, we argue that what LLMs actually generate in a synthetic experiment more closely resembles data from an observational study.

Notation and task. Let $A \in \{0, 1\}$ denote a treatment or intervention on a user, and let Y denote the primary outcome (the user's response). Let the full underlying user persona be represented by a set of attributes X. Let L be the subset of those attributes in X that are observed, and let $X \setminus L$ be the remaining latent attributes not included in L. The task is then to estimate how intervening on A changes the distribution of Y *in a real user population.* In an agent evaluation setting, for instance, A might be a dialogue with Agent 1 ($A = 0$) or Agent 2 ($A = 1$), and Y might be a user-reported measure of conversation satisfaction or quality. Variables are illustrated in Figure 2a.

Ideal randomized setting. We can now express a core distinction between real and synthetic experiments. Following the potential outcomes framework (Rubin, 1974), let $Y(a)$ denote the potential (or counterfactual) outcome that would be observed under intervention $A = a$. We aim to estimate the *average treatment effect* (ATE)

$$\tau^{\text{ATE}} = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\mathbb{E}[Y(1) - Y(0) \mid L]]. \quad (1)$$

We also make two standard causal assumptions: (1) *Consistency*: $Y = AY(1) + (1 - A)Y(0)$ and (2) *Latent unconfoundedness*: Given X, $A \perp Y(0), Y(1)$. Under these assumptions, we can represent

$$\mathbb{E}[Y(a) \mid L] = \int_{\mathcal{X}} \underbrace{\mathbb{E}[Y \mid A = a, X \mid L]}_{\text{outcome-generating process}} \underbrace{P(X \mid L)}_{\text{fixed population}} \, dX.$$

This is the average outcome across the fixed marginal population of personas X for both $Y(1)$ and $Y(0)$, which is the quantity measured in real randomized trials. τ^{ATE} is then identified from the data, as $\mathbb{E}[Y(a) \mid L] = E[Y \mid A = a, L]$.

Synthetic setting. In a synthetic experiment, instead of testing a real population of users, we prompt an LLM to embody many different personas by specifying user attributes L. For each persona, we simulate interactions under multiple conditions of A (e.g., Agent 1 vs. Agent 2) and outcomes Y. Superficially, this resembles a controlled comparison in which different interventions are presented to an identical user. This setup is enabled by how the LLM models its generative distribution. While the real-world data-generating process follows the causal ordering X, L, A, Y, the LLM factorizes the joint distribution in the sequence of its prompt context: L, A, X, Y. As a generative procedure, this ordering is convenient: it lets us condition on observed values of L and assign both interventions A to every persona, yielding a minimum-variance estimate of the apparent treatment effect.

However, LLMs are trained on observational data in which a user's true persona X can influence both their likelihood of experiencing a treatment A (e.g., self-selection into interacting with Agent 1 vs. 2) and their eventual response Y. Consequently, when we introduce the treatment $A = a$ to the LLM's context, which has also previously been instantiated with L, the model generates its response—and thus the outcome—from an *observed* conditional distribution rather than an interventional one:

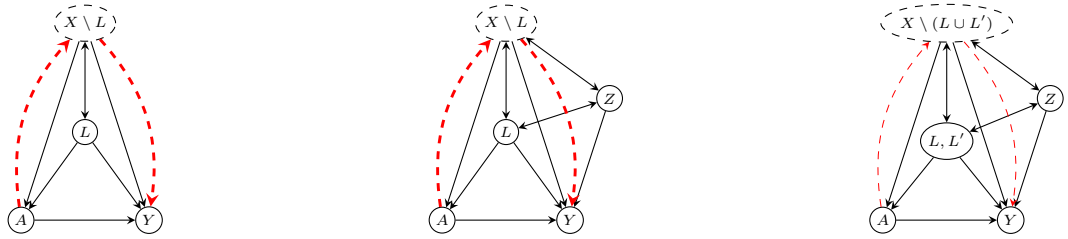
$$\mathbb{E}[Y \mid A = a, L] = \int_{\mathcal{X}} \underbrace{\mathbb{E}[Y \mid A = a, X, L]}_{\text{outcome-generating process}} \underbrace{P(X \mid A = a, L)}_{\text{treatment-conditional population}} \, dX \quad (2)$$

This yields the *observed effect*:

$$\begin{aligned} \tau^{\text{obs}} &= \mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0] \\ &= \mathbb{E}[\mathbb{E}[Y \mid A = 1, L] - \mathbb{E}[Y \mid A = 0, L]] \end{aligned}$$

What (2) suggests is that the distribution of user attributes X—and in particular the distribution of unobserved user attributes $X \setminus L$, as L is specified—can differ based on the value of A. This means that in a synthetic experiment, rather than comparing identical users' responses to different treatments, the initial identical users may *drift* to become distinct individuals with differing latent user attributes. While this occurs because the model generates responses by inferring a coherent persona from the entire context, this mechanism operates against the arrow of time, as a later intervention can influence what is meant to be the user's initial state.

Formally, if there is no confounding, as in a randomized experiment, then $P(X \mid A = 1, L) = P(X \mid A = 0, L)$, and $\tau^{\text{obs}} = \tau^{\text{ATE}}$. If $P(X \mid A = 1, L) \neq P(X \mid A = 0, L)$—that is, if the distribution of latent user attributes differs over A—then $P(Y \mid A = 1, L)$ and $P(Y \mid A = 0, L)$ effectively compare outcomes generated under different user



(a) Real-world DGP.

(b) Detect confounding with negative control outcomes Z.

(c) Elicit additional attributes L' to adjust for confounding.

Figure 2. DAGs for (a) the real-world observational data-generating process, (b) detecting selection bias via negative controls Z, and (c) reducing selection bias by iteratively eliciting additional attributes L' from $X \setminus L$. Red dashed arrows indicate open backdoor paths that drive the mechanism of user drift and selection bias in simulated user populations.

populations, creating an internal validity problem that biases the observed effect. We show an explicit decomposition for this bias in Appendix A.

Figure 2a depicts this issue as an open *backdoor path* from A to Y through latent user attributes. In the observational data that LLMs learn from, the unspecified attributes $X \setminus L$ influence both treatment exposure and outcomes, yielding the confounding path $A \leftarrow (X \setminus L) \rightarrow Y$. In an ideal randomized experiment, random assignment (within strata of L) blocks these backdoor paths by enforcing $A \perp (X \setminus L) \mid L$. In contrast, in LLM-simulated users instantiated with attributes L, interactions with different interventions A can themselves shift which unspecified latent attributes are inferred. Backdoor confounding in the training data-generating process induces shifts in $X \setminus L$ at simulation time, creating a post-treatment path from A to Y.

Our goal is therefore to detect these backdoor-induced biases by using negative control outcomes Z (Figure 2b) and to reduce them by eliciting additional attributes L' from $X \setminus L$ (Figure 2c). This ensures that comparisons across A are made over more similar latent user populations.

4. Detecting and Adjusting for Selection Bias

Having argued that synthetic experiments are inherently prone to selection bias, we provide in this section an empirical strategy to both diagnose and mitigate this bias in practice. The broad idea is to treat the synthetic experiment as an observational study, but one where we have the opportunity to query additional variables for diagnosis and adjustment at analysis time.

4.1. Detecting confounding with negative control outcomes

Because the underlying persona X of a synthetic user is not observed in full, user drift and selection bias cannot be directly computed. Instead, we develop a diagnostic for confounding using targeted *negative control outcomes* Z (Lipsitch et al., 2010), which serve as observed proxies for the unmeasured $X \setminus L$. In our usage, negative control outcomes should depend on the same unspecified latent user attributes $X \setminus L$ whose drift induces confounding bias in the primary outcome, but should not be *causally* affected by the intervention condition A (Figure 2b). Then as with $P(X \mid A = a, L)$, in the absence of confounding the negative control outcome distribution $P(Z \mid A = a, L)$ should remain invariant to A, i.e., it should hold that $P(Z \mid A = 1, L) = P(Z \mid A = 0, L)$.

At a high level, a good negative control outcome satisfies the following two properties:

1. *No causal effect of A on Z*: In an actual randomized experiment, Z should not change under interventions on A (given L).
2. *Sensitivity to latent user attributes*: Z should depend on (or be informative about) the unspecified latent user attributes $X \setminus L$ that may differ across intervention conditions.

Concretely, we consider outputs to prompts that target stable user traits not specified by L (e.g., demographics, self-reported beliefs) as negative control outcomes. Since these attributes should theoretically be invariant to the intervention A in a real-world experiment (as they would be causally upstream), we can measure empirical shifts in Z in a synthetic experiment to capture the user drift over intervention conditions induced by the LLM. We quantify this using the total variation distance (TVD) between the negative control outcome distributions for each persona specification L:

$$\text{TVD} = 0.5 \sum_{Z} \left| P(Z \mid A = 1, L) - P(Z \mid A = 0, L) \right|$$

The magnitude of the TVD can be interpreted as evidence of how strongly the intervention conditions correspond to different latent user populations, providing an indication of the extent to which the observed effect is vulnerable to confounding bias. We note that because empirical TVD is positively biased in finite samples, it will naturally be bounded away from zero due to sampling variance alone.

Algorithm 1 Iterative detection and confounder adjustment

Require: Treatment $A \in \{0,1\}$; personas $\{L_i\}_{i=1}^n$; trials per persona T; TVD threshold $\epsilon > 0$
Ensure: Elicited confounders $\{L'_i\}_{i=1}^n$ and primary/negative control outcomes
Initialize augmented persona $\tilde{L}_i \leftarrow L_i$ and $L'_i \leftarrow \emptyset$ for each $i \in \{1 \dots n\}$
repeat
  for each persona $i \in \{1 \dots n\}$, condition $a \in \{0,1\}$, and trial $t \in \{1 \dots T\}$ **do**
    Simulate interaction under $A{=}a$ with $\tilde{L}_i$; record $Y_{i,t,a}$ and $Z_{i,t,a}$
  Compute $\mathrm{TVD}_i \leftarrow \frac{1}{2}\sum_Z \left|P(Z \mid A=1,\tilde{L}_i) - P(Z \mid A=0,\tilde{L}_i)\right|$ for each i
  if $\frac{1}{n}\sum_{i=1}^n \mathrm{TVD}_i > \epsilon$ **then**
    for each persona $i \in \{1 \dots n\}$, condition $a \in \{0,1\}$, and trial $t \in \{1 \dots T\}$ **do**
      Elicit additional attributes $L'_{i,t,a} \sim P(L' \mid A = a, \tilde{L}_i)$
    for each persona $i \in \{1 \dots n\}$ **do**
      Sample one elicited attribute set $L'_i \leftarrow \mathrm{Unif}\big(\{L'_{i,t,a} : t \in [T], a \in \{0,1\}\}\big)$
      Update augmented persona $\tilde{L}_i \leftarrow (\tilde{L}_i \cup L'_i)$
until $\frac{1}{n}\sum_{i=1}^n \mathrm{TVD}_i \le \epsilon$ (or a max-iterations budget is reached)

The usefulness of TVD as a metric to quantify confounding is of course sensitive to one's choice of negative control outcomes. TVD scales with the amount of confounding only if the negative control outcome is chosen to be sufficiently sensitive to $X \setminus L$ and sufficiently insensitive to A. Furthermore, if we choose negative control outcomes that are only weakly related to $X \setminus L$, then we may appear to have less selection bias than we really do. In later sections, we discuss and provide evidence of this.

4.2. Adjusting for confounding

The negative control diagnostic can detect confounding in synthetic experiments, but it does not by itself specify how to reduce the resulting bias. Here, we propose a practical adjustment strategy: elicit additional attributes L' that proxy the unspecified latent attributes $X \setminus L$, then instantiate a new simulated user for whom both the initial observed attributes L *and* the elicited attributes L' are specified (Figure 2c).

Eliciting confounders. We draw new confounders for each persona from the treatment-marginal distribution of confounders, which is a mixture of the treatment-conditional confounder distributions:

$$L' \sim 0.5 \cdot P(L' \mid A=1, L) + 0.5 \cdot P(L' \mid A=0, L)$$

Note that this treatment-uniform mixture distribution is distinct from the observational distribution marginalized over treatment, $P(L' \mid L) = \sum_{a\in\{0,1\}} P(A=a \mid L)P(L' \mid A=a, L)$. We draw from the treatment-uniform mixture to ensure that the additional confounders reflect the distribution that would be induced by running a real randomized experiment, where treatment conditions are assigned with equal probability. The objective is to fix the elicited attributes L' across intervention conditions without narrowing the overall user distribution.

In practice, we simulate T trials for each user. In each trial, we prompt the user directly for a confounder value following intervention in each arm of the trial (e.g., by asking a targeted question). Across all trials for that user, we randomly select one of these realizations to add to the user persona for the next iteration, resulting in a single arm-independent L' across all trials for a specific user.

Causal caveats. Perfect causal adjustment aims to block backdoor paths by conditioning on variables that render treatment assignment independent of latent determinants of the outcome. Sampling and adjusting for L' in the manner we propose does not guarantee this, in part because prompting an LLM with descriptions of attributes approximates but is not equivalent to statistically conditioning on them. Rather, we aim for bias reduction: we choose new confounders L' such that the remaining unobserved persona attributes $X \setminus (L \cup L')$ that influence outcomes are less imbalanced across treatment arms. As with any observational study, we cannot guarantee success, but we can build falsifying evidence with well-chosen negative controls Z.

We draw attention to an important caveat: because we are unlikely to be able to fully specify $X \setminus L$ via L' (i.e., there is always some unmeasured confounding), then L' can reduce bias but is not guaranteed to eliminate it, especially if substantial latent variation remains in $X \setminus (L \cup L')$ or if the elicitation procedure induces additional dependencies. This means that confounding bias also does not necessarily decrease monotonically in the size of $L \cup L'$.

Iterative adjustment. In practice, rather than eliciting all possible confounders and providing them all at once in a detailed persona specification, we elicit small batches of attributes iteratively and add them to the persona specification, monitoring confounding at every iteration until the distance between the treatment-specific user populations is sufficiently reduced (Algorithm 1).

The motivation for this iterative procedure is threefold. First, confounder elicitation is computationally expensive, and adding long lists of confounders to the conversational context as part of adjustment further increases inference overhead. Second, LLMs suffer from performance degradation given long context windows, and expanding the persona specification to include all possible confounders increases

this risk. Third—because we "condition" not in the statistical sense but rather by sampling a fixed set of confounders that we specify in an LLM prompt—every constraint we add to the persona further runs the risk of narrowing the user population through sampling bias, which can affect the external validity of the effect computed from the synthetic experiment. Therefore, we attempt to identify a *minimum* set of confounders that sufficiently controls for confounding bias in the outcome.

5. Experiments

Models. We evaluate synthetic users simulated using the following diverse range of frontier models: Qwen3-30B-A3B (both fine-tuned and instruction-tuned versions) (Yang et al., 2025), GPT-OSS-20B (OpenAI et al., 2025), Gemma-4-31B (instruction-tuned) (Google, 2026b), and Gemini 3 Flash (Google, 2026a). We additionally include Gemma-3-4B (instruction-tuned) (Google, 2025) to explicitly test the behavior of a smaller-capacity model. Model details are available in Appendix C.

Procedure. We begin by instantiating the synthetic user with base persona attributes $\tilde{L} = L$ derived from an existing dataset of users. In parallel conversations, identically instantiated users are presented with interventions $A = 0$ and $A = 1$. In each arm, the user is then asked—in diverging conversation branches—the primary outcome question (eliciting Y), a negative control question (eliciting Z), or a confounder question (eliciting L'). As each question is asked independently of the others, the simulated user sees *only* that question and the intervention, not any of the other questions. These answers are then collected and used to compute both the observed effect and the TVD.[2] Of the sampled confounders, a subset is added to the persona specification, and the procedure is repeated from the beginning with a more detailed persona instantiation $\tilde{L} = (\tilde{L} \cup L')$. The conversation history from the previous iteration is *not* included in the new iteration.

For each experimental setting, we sample 100 personas from the dataset to instantiate 100 corresponding synthetic users. We then conduct 30 independent trials per user, evaluating both intervention conditions in parallel during each trial. Due to computational constraints, Gemini 3 Flash experiments use 10 synthetic users instantiated from 10 sampled personas, with 5 trials per user.

5.1. Simulated user survey

We evaluate two types of settings where we realistically expect synthetic users to be used. The first is in survey-type experiments (e.g., in social science), where simulated users are asked to express their opinions or preferences in response to fixed questions.

For this setting type, we draw from the OpinionQA dataset (Santurkar et al., 2023), which contains questions and human responses from the Pew American Trends Panel opinion polling surveys, along with demographic attributes of the respondents. In this setting, simulated users are given a leading statement before being asked to respond to the survey questions in the OpinionQA dataset. We use the following experimental setup:

- L: The age and sex of the respondent, seeded from the actual OpinionQA respondents.
- A: A leading statement to precede the survey question. Each intervention condition targets one of the possible answers to the survey question.
- Y: The respondent's answer to the survey question. We choose the question specifically to be one that suggests the political affiliation of the respondent.
- Z: The synthetic user's elicited citizenship status, political ideology, political party, race, and answers to 5 other survey questions that similarly provide an indication of political affiliation.
- L': The synthetic user's elicited geographic region, income, level of education, marital status, religion, religious attendance frequency, and 20 targeted LLM-generated questions that specifically relate to the intervention and outcome (10 groups of 2 questions each). Confounders are added over iterations in groups of 2, in the order they are listed here.

Prompts and lists of individual questions can be found in Appendix F, G.

5.2. Assistive agent evaluation

A second type of setting where we expect simulated users to be used is in evaluating LLM agents, e.g., an A/B test of whether a new version of an LLM agent is better than an existing version. Here, simulated users might engage in a conversation with the agent before being prompted with a primary question relating to their interaction with the agent.

As a common example of this type of interaction, we take a setting where the simulated user engages with the LLM agent to ask for a recommendation. We examine two variations of this setting where the user asks for either a book or a movie. The agent responds, and the two converse autonomously for several turns before the user is then asked how likely they are to read or watch the book or movie being discussed. We use the following experimental setup:

- L: The age and sex of the respondent, seeded from the actual respondents in the NYT Book Opinions[3] (Meis-

[2] For questions where a list of options is provided, answers matching none of the options are mapped to *Unknown*.

[3] We use only the titles and authors of 5 books from this dataset

ter et al., 2025) and MovieLens (Harper & Konstan, 2015) datasets.

- A: Different system prompts for the LLM agent (Gemma-4-31B-it). The agent is instructed to discuss a specific book/movie in a positively ($A = 1$) or negatively ($A = 0$) biased way.
- Y: How likely the respondent is to read the book/movie being discussed, on a scale from *1 - Very unlikely* to *4 - Very likely*.
- Z: The synthetic user's elicited citizenship status, political ideology, political party, and race. The respondent is also asked, "Have you heard of the ¡book/movie¿ we are discussing before?"
- L': The synthetic user's elicited geographic region, income, level of education, marital status, religion, religious attendance frequency, and 20 targeted LLM-generated questions that specifically relate to the intervention and outcome (10 groups of 2 questions each). Confounders are added over iterations in groups of 2, in the order they are listed here.

Prompts and lists of individual questions can be found in Appendix F, H.

6. Results and Discussion

We report results and provide a unified set of takeaways. Over adjustment iterations, we add elicited confounders to the persona specification, and we record (i) TVD between negative control outcome distributions under the two intervention conditions (Figure 3) and (ii) the observed treatment effect (Figure 4). Additional analysis and results are available in Appendix B. A comprehensive set of plots for all negative control outcomes at every adjustment step is available in Appendix I.

6.1. User drift and selection bias are present

Before adjustment (iteration 0), we observe substantial TVD between negative control outcome distributions across models and datasets (Figure 3), indicating that the simulated user population under $A = 0$ differs from that under $A = 1$. In the survey setting, most models exhibit a similar amount of user drift between interventions with the exception of the small Gemma-3-4B, which has a particularly large TVD. In the agent evaluation setting, the same pattern holds for both Book Opinions and MovieLens, but with substantial variation in the magnitude of difference across models.

We reiterate that *selection bias is a natural consequence of a model with strong abductive reasoning abilities*. Notably, in the MovieLens setting, the two smallest and oldest models—GPT-OSS-20B and Gemma-3-4B, which we expect to have the weakest abductive capabilities—show very little difference in negative control outcomes across interventions.

and the annotators' demographic information, and not any content created by the New York Times.

6.2. Iterative confounder adjustment is (often) effective

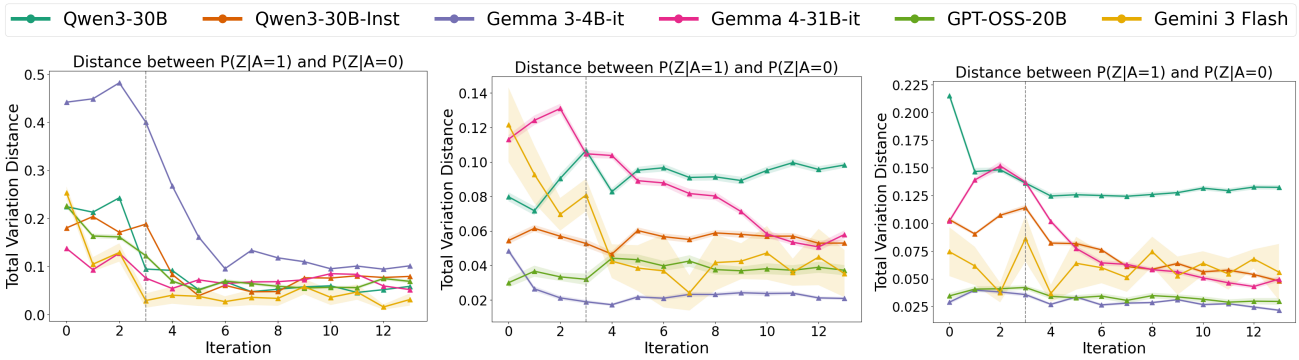
We observe that **iterative confounder adjustment appears to be effective in controlling user drift and selection bias for many models, but not all.** In the survey setting, TVD decreases consistently over iterations, suggesting that the adjustment procedure can reduce user drift (Figure 3a). In the agent evaluation setting, many model-dataset pairs exhibit clear reductions in TVD (e.g., Gemma-4-31B in both Book Opinions and MovieLens, Gemini 3 Flash and Gemma-3-4B in Book Opinions, and both fine-tuned and instruction-tuned Qwen3 in MovieLens), but others (primarily GPT-OSS-20B) show flat or weakly changing TVD across iterations (Figures 3b, 3c). These results suggest that prompt-based conditioning and the particular elicited attributes may not be uniformly effective for confounding adjustment in multi-turn settings.

Additionally, **user drift affects the observed effect estimates.** In both settings, the observed treatment effect changes over early adjustment iterations and then stabilizes after several iterations (Figure 4). Because conditioning on additional attributes should not substantially change the observed effect if the user distributions between interventions are already comparable, these dynamics empirically support that (i) intervention-dependent user drift contributes to confounding bias in naive estimates and (ii) adjustment reduces the relevant sources of drift. Stabilization does not guarantee that all bias has been eliminated, but it suggests that additional confounder elicitation beyond that point (at least of related attributes) is less likely to address drift that materially changes the observed effect.

Gemini 3 Flash exhibits notably anomalous behavior for its OpinionQA observed effect, which begins close to 0 and after several iterations converges to 0 (Figure 4a). While this may be partially due to the smaller sample size for this model (10 personas with 5 trials each), it is more likely that this behavior arises from the additional training proprietary models like Gemini undergo to express "neutral points-of-view" on potentially sensitive topics (Gemini Team et al., 2025).

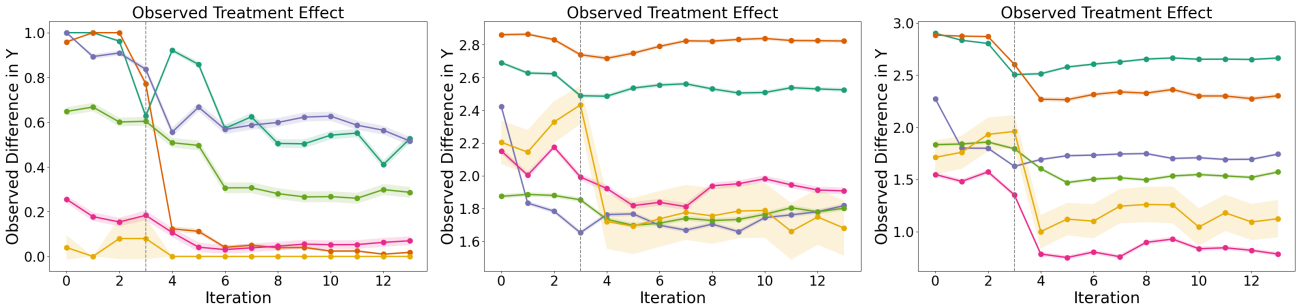
6.3. Design choices are important

Confounders. In statistical causal inference, partial adjustment for certain covariates has the potential to worsen rather than remedy confounding. Akin to this phenomenon, several models (e.g., Gemma-3, Qwen3) in the survey setting exhibit an initial *increase* in TVD before an eventual reduction. This increase corresponds to the iterations eliciting and adjusting for several generic demographic-based attributes: geographic region, income, education, and marital



(a) OpinionQA *(b)* Book Opinions *(c)* MovieLens

Figure 3. TVD over adjustment iterations. Shaded bands indicate 95% CIs. Dashed line indicates the switch from demographic to targeted confounders, i.e., the last step where only demographic confounders are included in adjustment.



(a) OpinionQA *(b)* Book Opinions *(c)* MovieLens

Figure 4. Observed effects over adjustment iterations. Shaded bands indicate 95% CIs. Dashed line indicates the switch from demographic to targeted confounders, i.e., the last step where only demographic confounders are included in adjustment.

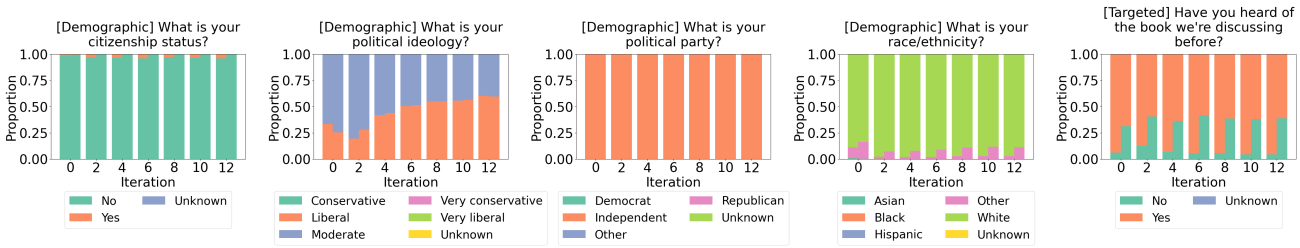


Figure 5. Negative control outcome distributions (Qwen3-30B, Book Opinions). At each adjustment iteration, $A = 0$ is on the left and $A = 1$ is on the right.

status. Subsequently, the simulated user's religious beliefs and habits are elicited and included in adjustment—which in contrast to the other demographic attributes does seem to decrease TVD across all three settings in many models. Finally, the elicited attributes then become the simulated user's answers to targeted questions related specifically to the intervention and outcome (this transition point is indicated by the vertical dashed lines in Figure 3). After these confounders are introduced in the persona specification, TVD continues to drop before stabilizing in later iterations.

This observation highlights a practical trade-off when constructing synthetic user populations. In persona instantiation, researchers typically rely on generic demographic attributes to span a target population with specific attributes and capture diverse user perspectives. The initial iterations of our adjustment loop show the impact of this practice: specification of generic attributes can in fact increase selection bias between intervention conditions. Since specifying such attributes is often necessary to define a sufficiently broad simulated population, our results suggest that a good practice is then to add targeted, task-specific confounders to subsequently bring selection bias back down.

Negative control outcomes. Our selection bias diagnostic depends on the sensitivity of our chosen negative control outcomes to drift in the latent user attributes. In the Book Opinions dataset, for instance, negative control

outcomes based on demographic attributes can show limited differences across intervention conditions for some models (Figure 5) and therefore are not particularly effective proxies for shifts in $X \setminus L$. This can yield low or flat TVD even when the observed effect continues to change across iterations. More targeted negative control outcomes, on the other hand, may be sensitive to the intervention, as seen for the rightmost negative control outcome in Figure 5. Additionally, certain models like GPT-OSS exhibit high rates of refusal to answer in some data settings (Figure 9a, where refusal maps to *Unknown*). If refusal dominates the negative control outcome distribution, then this may impact the TVD metric's ability to detect user drift.

7. Conclusion

LLM-simulated user studies can undermine the internal validity we expect from randomized experiments, inducing confounding bias. In particular, interventions can induce shifts in latent user attributes such that outcomes for each intervention are effectively drawn from different user populations, leading to confounded effect estimates. We formalize this phenomenon as selection bias in synthetic experiments, propose negative control outcomes as a practical diagnostic, and empirically find evidence of such shifts across survey-style and multi-turn agent evaluations. Finally, we show that eliciting and specifying targeted user attributes during persona instantiation can substantially reduce observed user drift and stabilize effect estimates. Our findings suggest that synthetic user experiments should be treated as observational by default and that diagnosing and mitigating user drift in such experiments is essential for drawing robust conclusions about the causal effect.

References

Aher, G., Arriaga, R. I., and Kalai, A. T. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Anthis, J. R., Liu, R., Richardson, S. M., Kozlowski, A. C., Koch, B., Brynjolfsson, E., Evans, J., and Bernstein, M. S. Position: LLM social simulations are a promising research method. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=cRBg1dtj7o.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31 (3):337–351, 2023. doi: 10.1017/pan.2023.2.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.

Castricato, L., Lile, N., Rafailov, R., Fränken, J.-P., and Finn, C. PERSONA: A reproducible testbed for pluralistic alignment. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S. (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 11348–11368, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.752/.

Dong, Y. R., Hu, T., and Collier, N. Can LLM be a personalized judge? In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10126–10141, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.592. URL https://aclanthology.org/2024.findings-emnlp.592/.

Dou, Y., Galley, M., Peng, B., Kedzie, C., Cai, W., Ritter, A., Quirk, C., Xu, W., and Gao, J. SimulatorArena: Are user simulators reliable proxies for multi-turn evaluation of AI assistants? In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35212–35290, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1786. URL https://aclanthology.org/2025.emnlp-main.1786/.

Fröhling, L., Demartini, G., and Assenmacher, D. Personas with attitudes: Controlling LLMs for diverse data annotation. In Calabrese, A., de Kock, C., Nozza, D., Plaza-del Arco, F. M., Talat, Z., and Vargas, F. (eds.), *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pp. 468–481, Vienna, Austria, August 2025. Association for Computational Linguistics. ISBN 979-8-89176-105-6. URL https://aclanthology.org/2025.woah-1.43/.

Gao, Y., Lee, D., Burtch, G., and Fazelpour, S. Take caution in using llms as human surrogates: Scylla ex machina, 2025. URL https://arxiv.org/abs/2410.19599.

Gemini Team, Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., Firat, O., Molloy, J., Isard, M., Barham, P. R., Hennigan, T., Lee, B., Viola, F., Reynolds, M., Xu, Y., Doherty, R., Collins, E., Meyer, C., Rutherford, E., Moreira, E., Ayoub, K., Goel, M., Krawczyk, J., Du, C., Chi, E., Cheng, H.-T., Ni, E., Shah, P., Kane, P., Chan, B., Faruqui, M., Severyn, A., Lin, H., Li, Y., Cheng, Y., Ittycheriah, A., Mahdieh, M., Chen, M., Sun, P., Tran, D., Bagri, S., Lakshminarayanan, B., Liu, J., Orban, A., Güra, F., Zhou, H., Song, X., Boffy, A., Ganapathy, H., Zheng, S., Choe, H., Ágoston Weisz, Zhu, T., Lu, Y., Gopal, S., Kahn, J., Kula, M., Pitman, J., Shah, R., Taropa, E., Merey, M. A., Baeuml, M., Chen, Z., Shafey, L. E., Zhang, Y., Sercinoglu, O., Tucker, G., Piqueras, E., Krikun, M., Barr, I., Savinov, N., Danihelka, I., Roelofs, B., White, A., Andreassen, A., von Glehn, T., Yagati, L., Kazemi, M., Gonzalez, L., Khalman, M., Sygnowski, J., Frechette, A., Smith, C., Culp, L., Proleev, L., Luan, Y., Chen, X., Lottes, J., Schucher, N., Lebron, F., Rrustemi, A., Clay, N., Crone, P., Kocisky, T., Zhao, J., Perz, B., Yu, D., Howard, H., Bloniarz, A., Rae, J. W., Lu, H., Sifre, L., Maggioni, M., Alcober, F., Garrette, D., Barnes, M., Thakoor, S., Austin, J., Barth-Maron, G., Wong, W., Joshi, R., Chaabouni, R., Fatiha, D., Ahuja, A., Tomar, G. S., Senter, E., Chadwick, M., Kornakov, I., Attaluri, N., Iturrate, I., Liu, R., Li, Y., Cogan, S., Chen, J., Jia, C., Gu, C., Zhang, Q., Grimstad, J., Hartman, A. J., Garcia, X., Pillai, T. S., Devlin, J., Laskin, M., de Las Casas, D., Valter, D., Tao, C., Blanco, L., Badia, A. P., Reitter, D., Chen, M., Brennan, J., Rivera, C., Brin, S., Iqbal, S., Surita, G., Labanowski, J., Rao, A., Winkler, S., Parisotto, E., Gu, Y., Olszewska, K., Addanki, R., Miech, A., Louis, A., Teplyashin, D., Brown, G., Catt, E., Balaguer, J., Xiang, J., Wang, P., Ashwood, Z., Briukhov, A., Webson, A., Ganapathy, S., Sanghavi, S., Kannan, A., Chang, M.-W., Stjerngren, A., Djolonga, J., Sun, Y., Bapna, A., Aitchison, M., Pejman, P., Michalewski, H., Yu, T., Wang, C., Love, J., Ahn, J., Bloxwich, D., Han, K., Humphreys, P., Sellam, T., Bradbury, J., Godbole, V., Samangooei, S., Damoc, B., Kaskasoli, A., Arnold, S. M. R., Vasudevan, V., Agrawal, S., Riesa, J., Lepikhin, D., Tanburn, R., Srinivasan, S., Lim, H., Hodkinson, S., Shyam, P., Ferret, J., Hand, S., Garg, A., Paine, T. L., Li, J., Li, Y., Giang, M., Neitz, A., Abbas, Z., York, S., Reid, M., Cole, E., Chowdhery, A., Das, D., Rogozińska, D., Nikolaev, V., Sprechmann, P., Nado, Z., Zilka, L., Prost, F., He, L., Monteiro, M., Mishra, G., Welty, C., Newlan, J., Jia, D., Allamanis, M., Hu, C. H., de Liedekerke, R., Gilmer, J., Saroufim, C., Rijhwani, S., Hou, S., Shrivastava, D., Baddepudi, A., Goldin, A., Ozturel, A., Cassirer, A., Xu, Y., Sohn, D., Sachan, D., Amplayo, R. K., Swanson, C., Petrova, D., Narayan, S., Guez, A., Brahma, S., Landon, J., Patel, M., Zhao, R., Villela, K., Wang, L., Jia, W., Rahtz, M., Giménez, M., Yeung, L., Keeling, J., Georgiev, P., Mincu, D., Wu, B., Haykal, S., Saputro, R., Vodrahalli, K., Qin, J., Cankara, Z., Sharma, A., Fernando, N., Hawkins, W., Neyshabur, B., Kim, S., Hutter, A., Agrawal, P., Castro-Ros, A., van den Driessche, G., Wang, T., Yang, F., yiin Chang, S., Komarek, P., McIlroy, R., Lučić, M., Zhang, G., Farhan, W., Sharman, M., Natsev, P., Michel, P., Bansal, Y., Qiao, S., Cao, K., Shakeri, S., Butterfield, C., Chung, J., Rubenstein, P. K., Agrawal, S., Mensch, A., Soparkar, K., Lenc, K., Chung, T., Pope, A., Maggiore, L., Kay, J., Jhakra, P., Wang, S., Maynez, J., Phuong, M., Tobin, T., Tacchetti, A., Trebacz, M., Robinson, K., Katariya, Y., Riedel, S., Bailey, P., Xiao, K., Ghelani, N., Aroyo, L., Slone, A., Houlsby, N., Xiong, X., Yang, Z., Gribovskaya, E., Adler, J., Wirth, M., Lee, L., Li, M., Kagohara, T., Pavagadhi, J., Bridgers, S., Bortsova, A., Ghemawat, S., Ahmed, Z., Liu, T., Powell, R., Bolina, V., Iinuma, M., Zablotskaia, P., Besley, J., Chung, D.-W., Dozat, T., Comanescu, R., Si, X., Greer, J., Su, G., Polacek, M., Kaufman, R. L., Tokumine, S., Hu, H., Buchatskaya, E., Miao, Y., Elhawaty, M., Siddhant, A., Tomasev, N., Xing, J., Greer, C., Miller, H., Ashraf, S., Roy, A., Zhang, Z., Ma, A., Filos, A., Besta, M., Blevins, R., Klimenko, T., Yeh, C.-K., Changpinyo, S., Mu, J., Chang, O., Pajarskas, M., Muir, C., Cohen, V., Lan, C. L., Haridasan, K., Marathe, A., Hansen, S., Douglas, S., Samuel, R., Wang, M., Austin, S., Lan, C., Jiang, J., Chiu, J., Lorenzo, J. A., Sjösund, L. L., Cevey, S., Gleicher, Z., Avrahami, T., Boral, A., Srinivasan, H., Selo, V., May, R., Aisopos, K., Hussenot, L., Soares, L. B., Baumli, K., Chang, M. B., Recasens, A., Caine, B., Pritzel, A., Pavetic, F., Pardo, F., Gergely, A., Frye, J., Ramasesh, V., Horgan, D., Badola, K., Kassner, N., Roy, S., Dyer, E., Campos, V. C., Tomala, A., Tang, Y., Badawy, D. E., White, E., Mustafa, B., Lang, O., Jindal, A., Vikram, S., Gong, Z., Caelles, S., Hemsley, R., Thornton, G., Feng, F., Stokowiec, W., Zheng, C., Thacker, P., Çağlar Ünlü, Zhang, Z., Saleh, M., Svensson, J., Bileschi, M., Patil, P., Anand, A., Ring, R., Tsihlas, K., Vezer, A., Selvi, M., Shevlane, T., Rodriguez, M., Kwiatkowski, T., Daruki, S., Rong, K., Dafoe, A., FitzGerald, N., Gu-Lemberg, K., Khan, M., Hendricks, L. A., Pellat, M., Feinberg, V., Cobon-Kerr, J., Sainath, T., Rauh, M., Hashemi, S. H., Ives, R., Hasson, Y., Noland, E., Cao, Y., Byrd, N., Hou, L., Wang, Q., Sottiaux, T., Paganini, M., Lespiau, J.-B., Moufarek, A., Hassan, S., Shivakumar, K., van Amersfoort, J., Mandhane, A., Joshi, P., Goyal, A., Tung, M.,

Brock, A., Sheahan, H., Misra, V., Li, C., Rakićević, N., Dehghani, M., Liu, F., Mittal, S., Oh, J., Noury, S., Sezener, E., Huot, F., Lamm, M., Cao, N. D., Chen, C., Mudgal, S., Stella, R., Brooks, K., Vasudevan, G., Liu, C., Chain, M., Melinkeri, N., Cohen, A., Wang, V., Seymore, K., Zubkov, S., Goel, R., Yue, S., Krishnakumaran, S., Albert, B., Hurley, N., Sano, M., Mohananey, A., Joughin, J., Filonov, E., Kepa, T., Eldawy, Y., Lim, J., Rishi, R., Badiezadegan, S., Bos, T., Chang, J., Jain, S., Padmanabhan, S. G. S., Puttagunta, S., Krishna, K., Baker, L., Kalb, N., Bedapudi, V., Kurzrok, A., Lei, S., Yu, A., Litvin, O., Zhou, X., Wu, Z., Sobell, S., Siciliano, A., Papir, A., Neale, R., Bragagnolo, J., Toor, T., Chen, T., Anklin, V., Wang, F., Feng, R., Gholami, M., Ling, K., Liu, L., Walter, J., Moghaddam, H., Kishore, A., Adamek, J., Mercado, T., Mallinson, J., Wandekar, S., Cagle, S., Ofek, E., Garrido, G., Lombriser, C., Mukha, M., Sun, B., Mohammad, H. R., Matak, J., Qian, Y., Peswani, V., Janus, P., Yuan, Q., Schelin, L., David, O., Garg, A., He, Y., Duzhyi, O., Älgmyr, A., Lottaz, T., Li, Q., Yadav, V., Xu, L., Chinien, A., Shivanna, R., Chuklin, A., Li, J., Spadine, C., Wolfe, T., Mohamed, K., Das, S., Dai, Z., He, K., von Dincklage, D., Upadhyay, S., Maurya, A., Chi, L., Krause, S., Salama, K., Rabinovitch, P. G., M, P. K. R., Selvan, A., Dektiarev, M., Ghiasi, G., Guven, E., Gupta, H., Liu, B., Sharma, D., Shtacher, I. H., Paul, S., Akerlund, O., Aubet, F.-X., Huang, T., Zhu, C., Zhu, E., Teixeira, E., Fritze, M., Bertolini, F., Marinescu, L.-E., Bölle, M., Paulus, D., Gupta, K., Latkar, T., Chang, M., Sanders, J., Wilson, R., Wu, X., Tan, Y.-X., Thiet, L. N., Doshi, T., Lall, S., Mishra, S., Chen, W., Luong, T., Benjamin, S., Lee, J., Andrejczuk, E., Rabiej, D., Ranjan, V., Styrc, K., Yin, P., Simon, J., Harriott, M. R., Bansal, M., Robsky, A., Bacon, G., Greene, D., Mirylenka, D., Zhou, C., Sarvana, O., Goyal, A., Andermatt, S., Siegler, P., Horn, B., Israel, A., Pongetti, F., Chen, C.-W. L., Selvatici, M., Silva, P., Wang, K., Tolins, J., Guu, K., Yogev, R., Cai, X., Agostini, A., Shah, M., Nguyen, H., Donnaile, N. , Pereira, S., Friso, L., Stambler, A., Kurzrok, A., Kuang, C., Romanikhin, Y., Geller, M., Yan, Z., Jang, K., Lee, C.-C., Fica, W., Malmi, E., Tan, Q., Banica, D., Balle, D., Pham, R., Huang, Y., Avram, D., Shi, H., Singh, J., Hidey, C., Ahuja, N., Saxena, P., Dooley, D., Potharaju, S. P., O'Neill, E., Gokulchandran, A., Foley, R., Zhao, K., Dusenberry, M., Liu, Y., Mehta, P., Kotikalapudi, R., Safranek-Shrader, C., Goodman, A., Kessinger, J., Globen, E., Kolhar, P., Gorgolewski, C., Ibrahim, A., Song, Y., Eichenbaum, A., Brovelli, T., Potluri, S., Lahoti, P., Baetu, C., Ghorbani, A., Chen, C., Crawford, A., Pal, S., Sridhar, M., Gurita, P., Mujika, A., Petrovski, I., Cedoz, P.-L., Li, C., Chen, S., Santo, N. D., Goyal, S., Punjabi, J., Kappaganthu, K., Kwak, C., LV, P., Velury, S., Choudhury, H., Hall, J., Shah, P., Figueira, R., Thomas, M., Lu, M., Zhou, T., Kumar, C., Jurdi, T., Chikkerur, S., Ma, Y., Yu, A., Kwak, S., Ähdel, V., Rajayogam, S., Choma, T., Liu, F., Barua, A., Ji, C., Park, J. H., Hellendoorn, V., Bailey, A., Bilal, T., Zhou, H., Khatir, M., Sutton, C., Rzadkowski, W., Macintosh, F., Vij, R., Shagin, K., Medina, P., Liang, C., Zhou, J., Shah, P., Bi, Y., Dankovics, A., Banga, S., Lehmann, S., Bredesen, M., Lin, Z., Hoffmann, J. E., Lai, J., Chung, R., Yang, K., Balani, N., Bražinskas, A., Sozanschi, A., Hayes, M., Alcalde, H. F., Makarov, P., Chen, W., Stella, A., Snijders, L., Mandl, M., Kärrman, A., Nowak, P., Wu, X., Dyck, A., Vaidyanathan, K., R, R., Mallet, J., Rudominer, M., Johnston, E., Mittal, S., Udathu, A., Christensen, J., Verma, V., Irving, Z., Santucci, A., Elsayed, G., Davoodi, E., Georgiev, M., Tenney, I., Hua, N., Cideron, G., Leurent, E., Alnahlawi, M., Georgescu, I., Wei, N., Zheng, I., Scandinaro, D., Jiang, H., Snoek, J., Sundararajan, M., Wang, X., Ontiveros, Z., Karo, I., Cole, J., Rajashekhar, V., Tumeh, L., Ben-David, E., Jain, R., Uesato, J., Datta, R., Bunyan, O., Wu, S., Zhang, J., Stanczyk, P., Zhang, Y., Steiner, D., Naskar, S., Azzam, M., Johnson, M., Paszke, A., Chiu, C.-C., Elias, J. S., Mohiuddin, A., Muhammad, F., Miao, J., Lee, A., Vieillard, N., Park, J., Zhang, J., Stanway, J., Garmon, D., Karmarkar, A., Dong, Z., Lee, J., Kumar, A., Zhou, L., Evens, J., Isaac, W., Irving, G., Loper, E., Fink, M., Arkatkar, I., Chen, N., Shafran, I., Petrychenko, I., Chen, Z., Jia, J., Levskaya, A., Zhu, Z., Grabowski, P., Mao, Y., Magni, A., Yao, K., Snaider, J., Casagrande, N., Palmer, E., Suganthan, P., Castaño, A., Giannoumis, I., Kim, W., Rybiński, M., Sreevatsa, A., Prendki, J., Soergel, D., Goedeckemeyer, A., Gierke, W., Jafari, M., Gaba, M., Wiesner, J., Wright, D. G., Wei, Y., Vashisht, H., Kulizhskaya, Y., Hoover, J., Le, M., Li, L., Iwuanyanwu, C., Liu, L., Ramirez, K., Khorlin, A., Cui, A., LIN, T., Wu, M., Aguilar, R., Pallo, K., Chakladar, A., Perng, G., Abellan, E. A., Zhang, M., Dasgupta, I., Kushman, N., Penchev, I., Repina, A., Wu, X., van der Weide, T., Ponnapalli, P., Kaplan, C., Simsa, J., Li, S., Dousse, O., Yang, F., Piper, J., Ie, N., Pasumarthi, R., Lintz, N., Vijayakumar, A., Andor, D., Valenzuela, P., Lui, M., Paduraru, C., Peng, D., Lee, K., Zhang, S., Greene, S., Nguyen, D. D., Kurylowicz, P., Hardin, C., Dixon, L., Janzer, L., Choo, K., Feng, Z., Zhang, B., Singhal, A., Du, D., McKinnon, D., Antropova, N., Bolukbasi, T., Keller, O., Reid, D., Finchelstein, D., Raad, M. A., Crocker, R., Hawkins, P., Dadashi, R., Gaffney, C., Franko, K., Bulanova, A., Leblond, R., Chung, S., Askham, H., Cobo, L. C., Xu, K., Fischer, F., Xu, J., Sorokin, C., Alberti, C., Lin, C.-C., Evans, C., Dimitriev, A., Forbes, H., Banarse, D., Tung, Z., Omernick, M., Bishop, C., Sterneck, R., Jain, R., Xia, J., Amid, E., Piccinno, F., Wang, X., Banzal, P., Mankowitz, D. J., Polozov, A., Krakovna, V., Brown, S., Bateni, M., Duan, D., Firoiu, V., Thotakuri, M., Natan, T., Geist, M., tan Girgin, S., Li, H., Ye, J., Roval, O., Tojo, R., Kwong, M., Lee-Thorp, J., Yew,

C., Sinopalnikov, D., Ramos, S., Mellor, J., Sharma, A., Wu, K., Miller, D., Sonnerat, N., Vnukov, D., Greig, R., Beattie, J., Caveness, E., Bai, L., Eisenschlos, J., Korchemniy, A., Tsai, T., Jasarevic, M., Kong, W., Dao, P., Zheng, Z., Liu, F., Yang, F., Zhu, R., Teh, T. H., Sanmiya, J., Gladchenko, E., Trdin, N., Toyama, D., Rosen, E., Tavakkol, S., Xue, L., Elkind, C., Woodman, O., Carpenter, J., Papamakarios, G., Kemp, R., Kafle, S., Grunina, T., Sinha, R., Talbert, A., Wu, D., Owusu-Afriyie, D., Du, C., Thornton, C., Pont-Tuset, J., Narayana, P., Li, J., Fatehi, S., Wieting, J., Ajmeri, O., Uria, B., Ko, Y., Knight, L., Héliou, A., Niu, N., Gu, S., Pang, C., Li, Y., Levine, N., Stolovich, A., Santamaria-Fernandez, R., Goenka, S., Yustalim, W., Strudel, R., Elqursh, A., Deck, C., Lee, H., Li, Z., Levin, K., Hoffmann, R., Holtmann-Rice, D., Bachem, O., Arora, S., Koh, C., Yeganeh, S. H., Põder, S., Tariq, M., Sun, Y., Ionita, L., Seyedhosseini, M., Tafti, P., Liu, Z., Gulati, A., Liu, J., Ye, X., Chrzaszcz, B., Wang, L., Sethi, N., Li, T., Brown, B., Singh, S., Fan, W., Parisi, A., Stanton, J., Koverkathu, V., Choquette-Choo, C. A., Li, Y., Lu, T., Ittycheriah, A., Shroff, P., Varadarajan, M., Bahargam, S., Willoughby, R., Gaddy, D., Desjardins, G., Cornero, M., Robenek, B., Mittal, B., Albrecht, B., Shenoy, A., Moiseev, F., Jacobsson, H., Ghaffarkhah, A., Rivière, M., Walton, A., Crepy, C., Parrish, A., Zhou, Z., Farabet, C., Radebaugh, C., Srinivasan, P., van der Salm, C., Fidjeland, A., Scellato, S., Latorre-Chimoto, E., Klimczak-Plucińska, H., Bridson, D., de Cesare, D., Hudson, T., Mendolicchio, P., Walker, L., Morris, A., Mauger, M., Guseynov, A., Reid, A., Odoom, S., Loher, L., Cotruta, V., Yenugula, M., Grewe, D., Petrushkina, A., Duerig, T., Sanchez, A., Yadlowsky, S., Shen, A., Globerson, A., Webb, L., Dua, S., Li, D., Bhupatiraju, S., Hurt, D., Qureshi, H., Agarwal, A., Shani, T., Eyal, M., Khare, A., Belle, S. R., Wang, L., Tekur, C., Kale, M. S., Wei, J., Sang, R., Saeta, B., Liechty, T., Sun, Y., Zhao, Y., Lee, S., Nayak, P., Fritz, D., Vuyyuru, M. R., Aslanides, J., Vyas, N., Wicke, M., Ma, X., Eltyshev, E., Martin, N., Cate, H., Manyika, J., Amiri, K., Kim, Y., Xiong, X., Kang, K., Luisier, F., Tripuraneni, N., Madras, D., Guo, M., Waters, A., Wang, O., Ainslie, J., Baldridge, J., Zhang, H., Pruthi, G., Bauer, J., Yang, F., Mansour, R., Gelman, J., Xu, Y., Polovets, G., Liu, J., Cai, H., Chen, W., Sheng, X., Xue, E., Ozair, S., Angermueller, C., Li, X., Sinha, A., Wang, W., Wiesinger, J., Koukoumidis, E., Tian, Y., Iyer, A., Gurumurthy, M., Goldenson, M., Shah, P., Blake, M., Yu, H., Urbanowicz, A., Palomaki, J., Fernando, C., Durden, K., Mehta, H., Momchev, N., Rahimtoroghi, E., Georgaki, M., Raul, A., Ruder, S., Redshaw, M., Lee, J., Zhou, D., Jalan, K., Li, D., Hechtman, B., Schuh, P., Nasr, M., Milan, K., Mikulik, V., Franco, J., Green, T., Nguyen, N., Kelley, J., Mahendru, A., Hu, A., Howland, J., Vargas, B., Hui, J., Bansal, K., Rao, V., Ghiya, R., Wang, E., Ye, K., Sarr, J. M., Preston, M. M., Elish, M., Li, S., Kaku, A., Gupta, J., Pasupat, I., Juan, D.-C., Someswar, M., M., T., Chen, X., Amini, A., Fabrikant, A., Chu, E., Dong, X., Muthal, A., Buthpitiya, S., Jauhari, S., Hua, N., Khandelwal, U., Hitron, A., Ren, J., Rinaldi, L., Drath, S., Dabush, A., Jiang, N.-J., Godhia, H., Sachs, U., Chen, A., Fan, Y., Taitelbaum, H., Noga, H., Dai, Z., Wang, J., Liang, C., Hamer, J., Ferng, C.-S., Elkind, C., Atias, A., Lee, P., Listík, V., Carlen, M., van de Kerkhof, J., Pikus, M., Zaher, K., Müller, P., Zykova, S., Stefanec, R., Gatsko, V., Hirnschall, C., Sethi, A., Xu, X. F., Ahuja, C., Tsai, B., Stefanoiu, A., Feng, B., Dhandhania, K., Katyal, M., Gupta, A., Parulekar, A., Pitta, D., Zhao, J., Bhatia, V., Bhavnani, Y., Alhadlaq, O., Li, X., Danenberg, P., Tu, D., Pine, A., Filippova, V., Ghosh, A., Limonchik, B., Urala, B., Lanka, C. K., Clive, D., Sun, Y., Li, E., Wu, H., Hongtongsak, K., Li, I., Thakkar, K., Omarov, K., Majmundar, K., Alverson, M., Kucharski, M., Patel, M., Jain, M., Zabelin, M., Pelagatti, P., Kohli, R., Kumar, S., Kim, J., Sankar, S., Shah, V., Ramachandruni, L., Zeng, X., Bariach, B., Weidinger, L., Vu, T., Andreev, A., He, A., Hui, K., Kashem, S., Subramanya, A., Hsiao, S., Hassabis, D., Kavukcuoglu, K., Sadovsky, A., Le, Q., Strohman, T., Wu, Y., Petrov, S., Dean, J., and Vinyals, O. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Google. Gemma-3-4b-it. https://huggingface.co/google/gemma-3-4b-it, 2025.

Google. Gemini 3 developer guide — gemini api. https://ai.google.dev/gemini-api/docs/gemini-3, 2026a.

Google. Gemma-4-31b-it. https://huggingface.co/google/gemma-4-31B-it, 2026b.

Gu, E. H., Chandrasegaran, S., and Lloyd, P. Synthetic users: insights from designers' interactions with persona-based chatbots. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 39:e2, 2025. doi: 10.1017/S0890060424000283.

Guerdan, L., Whitehouse, J., Truong, K., Holstein, K., and Wu, S. Doubly-robust LLM-as-a-judge: Externally valid estimation with imperfect personas. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EIA1tpKYL7.

Gui, G. and Toubia, O. The challenge of using llms to simulate human behavior: A causal inference perspective. *SSRN Electronic Journal*, 2023. ISSN 1556-5068. doi: 10.2139/ssrn.4650172. URL http://dx.doi.org/10.2139/ssrn.4650172.

Harper, F. M. and Konstan, J. A. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4), December 2015. ISSN 2160-6455. doi: 10.1145/2827872. URL https://doi.org/10.1145/2827872.

Hu, T. and Collier, N. Quantifying the persona effect in LLM simulations. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10289–10307, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.554. URL https://aclanthology.org/2024.acl-long.554/.

Kaiser, C., Kaiser, J., Manewitsch, V., Rau, L., and Schallner, R. Simulating human opinions with large language models: Opportunities and challenges for personalized survey data modeling. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP Adjunct '25, pp. 82–86, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713996. doi: 10.1145/3708319.3733685. URL https://doi.org/10.1145/3708319.3733685.

Kang, M., Moon, S., Lee, S. H., Raj, A., Suh, J., and Chan, D. Deep binding of language model virtual personas: a study on approximating political partisan misperceptions. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=zHdSCtNmM4.

Lee, B., Kim, S., Menczer, F., Ahn, Y.-Y., Kwak, H., and An, J. Llms can infer political alignment from online conversations, 2026. URL https://arxiv.org/abs/2603.11253.

Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K., Bishop, C., Hall, E., Carbune, V., Rastogi, A., and Prakash, S. Rlaif vs. rlhf: scaling reinforcement learning from human feedback with ai feedback. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Li, A., Chen, H., Namkoong, H., and Peng, T. LLM generated persona is a promise with a catch. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*, 2025. URL https://openreview.net/forum?id=qh9eGtMG4H.

Lipsitch, M., Tchetgen, E. T., and Cohen, T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388, 2010.

Lyman, A., Hepner, B., Argyle, L. P., Busby, E. C., Gubler, J. R., and Wingate, D. Balancing large language model alignment and algorithmic fidelity in social science research. *Sociological Methods and Research*, 54, 2025.

Manning, B. S., Zhu, K., and Horton, J. J. Automated social science: Language models as scientist and subjects, 2024. URL https://arxiv.org/abs/2404.11794.

Meister, N., Guestrin, C., and Hashimoto, T. Benchmarking distributional alignment of large language models. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 24–49, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.2. URL https://aclanthology.org/2025.naacl-long.2/.

Moon, S., Abdulhai, M., Kang, M., Suh, J., Soedarmadji, W., Behar, E. K., and Chan, D. M. Virtual personas for language models via an anthology of backstories. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19864–19897, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1110. URL https://aclanthology.org/2024.emnlp-main.1110/.

Moon, S., Kang, M., Suh, J., Safdari, M., and Canny, J. Identity, cooperation and framing effects within groups of real and simulated humans, 2026. URL https://arxiv.org/abs/2601.16355.

Neumann, T., De-Arteaga, M., and Fazelpour, S. Should you use llms to simulate opinions? quality checks for early-stage deliberation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(46): 39070–39079, Mar. 2026. doi: 10.1609/aaai.v40i46.41254. URL https://ojs.aaai.org/index.php/AAAI/article/view/41254.

OpenAI, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Madry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., Paino, A., Renzin, A., Passos, A. T., Kirillov, A., Christakis, A., Conneau, A., Kamali, A., Jabri, A., Moyer, A., Tam, A., Crookes, A., Tootoochian, A., Tootoonchian, A., Kumar, A., Vallone, A., Karpathy, A., Braunstein, A., Cann, A., Codispoti, A., Galu, A., Kondrich, A., Tulloch, A., Mishchenko, A., Baek, A., Jiang, A., Pelisse, A., Woodford, A., Gosalia, A., Dhar, A., Pantuliano, A., Nayak, A., Oliver, A., Zoph, B., Ghorbani, B., Leimberger, B., Rossen, B., Sokolowsky, B., Wang, B.,

Zweig, B., Hoover, B., Samic, B., McGrew, B., Spero, B., Giertler, B., Cheng, B., Lightcap, B., Walkin, B., Quinn, B., Guarraci, B., Hsu, B., Kellogg, B., Eastman, B., Lugaresi, C., Wainwright, C., Bassin, C., Hudson, C., Chu, C., Nelson, C., Li, C., Shern, C. J., Conger, C., Barette, C., Voss, C., Ding, C., Lu, C., Zhang, C., Beaumont, C., Hallacy, C., Koch, C., Gibson, C., Kim, C., Choi, C., McLeavey, C., Hesse, C., Fischer, C., Winter, C., Czarnecki, C., Jarvis, C., Wei, C., Koumouzelis, C., Sherburn, D., Kappler, D., Levin, D., Levy, D., Carr, D., Farhi, D., Mely, D., Robinson, D., Sasaki, D., Jin, D., Valladares, D., Tsipras, D., Li, D., Nguyen, D. P., Findlay, D., Oiwoh, E., Wong, E., Asdar, E., Proehl, E., Yang, E., Antonow, E., Kramer, E., Peterson, E., Sigler, E., Wallace, E., Brevdo, E., Mays, E., Khorasani, F., Such, F. P., Raso, F., Zhang, F., von Lohmann, F., Sulit, F., Goh, G., Oden, G., Salmon, G., Starace, G., Brockman, G., Salman, H., Bao, H., Hu, H., Wong, H., Wang, H., Schmidt, H., Whitney, H., Jun, H., Kirchner, H., de Oliveira Pinto, H. P., Ren, H., Chang, H., Chung, H. W., Kivlichan, I., O'Connell, I., O'Connell, I., Osband, I., Silber, I., Sohl, I., Okuyucu, I., Lan, I., Kostrikov, I., Sutskever, I., Kanitscheider, I., Gulrajani, I., Coxon, J., Menick, J., Pachocki, J., Aung, J., Betker, J., Crooks, J., Lennon, J., Kiros, J., Leike, J., Park, J., Kwon, J., Phang, J., Teplitz, J., Wei, J., Wolfe, J., Chen, J., Harris, J., Varavva, J., Lee, J. G., Shieh, J., Lin, J., Yu, J., Weng, J., Tang, J., Yu, J., Jang, J., Candela, J. Q., Beutler, J., Landers, J., Parish, J., Heidecke, J., Schulman, J., Lachman, J., McKay, J., Uesato, J., Ward, J., Kim, J. W., Huizinga, J., Sitkin, J., Kraaijeveld, J., Gross, J., Kaplan, J., Snyder, J., Achiam, J., Jiao, J., Lee, J., Zhuang, J., Harriman, J., Fricke, K., Hayashi, K., Singhal, K., Shi, K., Karthik, K., Wood, K., Rimbach, K., Hsu, K., Nguyen, K., Gu-Lemberg, K., Button, K., Liu, K., Howe, K., Muthukumar, K., Luther, K., Ahmad, L., Kai, L., Itow, L., Workman, L., Pathak, L., Chen, L., Jing, L., Guy, L., Fedus, L., Zhou, L., Mamitsuka, L., Weng, L., McCallum, L., Held, L., Ouyang, L., Feuvrier, L., Zhang, L., Kondraciuk, L., Kaiser, L., Hewitt, L., Metz, L., Doshi, L., Aflak, M., Simens, M., Boyd, M., Thompson, M., Dukhan, M., Chen, M., Gray, M., Hudnall, M., Zhang, M., Aljubeh, M., Litwin, M., Zeng, M., Johnson, M., Shetty, M., Gupta, M., Shah, M., Yatbaz, M., Yang, M. J., Zhong, M., Glaese, M., Chen, M., Janner, M., Lampe, M., Petrov, M., Wu, M., Wang, M., Fradin, M., Pokrass, M., Castro, M., de Castro, M. O. T., Pavlov, M., Brundage, M., Wang, M., Khan, M., Murati, M., Bavarian, M., Lin, M., Yesildal, M., Soto, N., Gimelshein, N., Cone, N., Staudacher, N., Summers, N., LaFontaine, N., Chowdhury, N., Ryder, N., Stathas, N., Turley, N., Tezak, N., Felix, N., Kudige, N., Keskar, N., Deutsch, N., Bundick, N., Puckett, N., Nachum, O., Okelola, O., Boiko, O., Murk, O., Jaffe, O., Watkins, O., Godement, O., Campbell-Moore, O., Chao, P., McMillan, P., Belov, P., Su, P., Bak, P., Bakkum, P., Deng, P., Dolan, P., Hoeschele, P., Welinder, P., Tillet, P., Pronin, P., Tillet, P., Dhariwal, P., Yuan, Q., Dias, R., Lim, R., Arora, R., Troll, R., Lin, R., Lopes, R. G., Puri, R., Miyara, R., Leike, R., Gaubert, R., Zamani, R., Wang, R., Donnelly, R., Honsby, R., Smith, R., Sahai, R., Ramchandani, R., Huet, R., Carmichael, R., Zellers, R., Chen, R., Chen, R., Nigmatullin, R., Cheu, R., Jain, S., Altman, S., Schoenholz, S., Toizer, S., Miserendino, S., Agarwal, S., Culver, S., Ethersmith, S., Gray, S., Grove, S., Metzger, S., Hermani, S., Jain, S., Zhao, S., Wu, S., Jomoto, S., Wu, S., Shuaiqi, Xia, Phene, S., Papay, S., Narayanan, S., Coffey, S., Lee, S., Hall, S., Balaji, S., Broda, T., Stramer, T., Xu, T., Gogineni, T., Christianson, T., Sanders, T., Patwardhan, T., Cunninghman, T., Degry, T., Dimson, T., Raoux, T., Shadwell, T., Zheng, T., Underwood, T., Markov, T., Sherbakov, T., Rubin, T., Stasi, T., Kaftan, T., Heywood, T., Peterson, T., Walters, T., Eloundou, T., Qi, V., Moeller, V., Monaco, V., Kuo, V., Fomenko, V., Chang, W., Zheng, W., Zhou, W., Manassra, W., Sheu, W., Zaremba, W., Patil, Y., Qian, Y., Kim, Y., Cheng, Y., Zhang, Y., He, Y., Zhang, Y., Jin, Y., Dai, Y., and Malkov, Y. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

OpenAI, Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R. K., Bai, Y., Baker, B., Bao, H., Barak, B., Bennett, A., Bertao, T., Brett, N., Brevdo, E., Brockman, G., Bubeck, S., Chang, C., Chen, K., Chen, M., Cheung, E., Clark, A., Cook, D., Dukhan, M., Dvorak, C., Fives, K., Fomenko, V., Garipov, T., Georgiev, K., Glaese, M., Gogineni, T., Goucher, A., Gross, L., Guzman, K. G., Hallman, J., Hehir, J., Heidecke, J., Helyar, A., Hu, H., Huet, R., Huh, J., Jain, S., Johnson, Z., Koch, C., Kofman, I., Kundel, D., Kwon, J., Kyrylov, V., Le, E. Y., Leclerc, G., Lennon, J. P., Lessans, S., Lezcano-Casado, M., Li, Y., Li, Z., Lin, J., Liss, J., Lily, Liu, Liu, J., Lu, K., Lu, C., Martinovic, Z., McCallum, L., McGrath, J., McKinney, S., McLaughlin, A., Mei, S., Mostovoy, S., Mu, T., Myles, G., Neitz, A., Nichol, A., Pachocki, J., Paino, A., Palmie, D., Pantuliano, A., Parascandolo, G., Park, J., Pathak, L., Paz, C., Peran, L., Pimenov, D., Pokrass, M., Proehl, E., Qiu, H., Raila, G., Raso, F., Ren, H., Richardson, K., Robinson, D., Rotsted, B., Salman, H., Sanjeev, S., Schwarzer, M., Sculley, D., Sikchi, H., Simon, K., Singhal, K., Song, Y., Stuckey, D., Sun, Z., Tillet, P., Toizer, S., Tsimpourlas, F., Vyas, N., Wallace, E., Wang, X., Wang, M., Watkins, O., Weil, K., Wendling, A., Whinnery, K., Whitney, C., Wong, H., Yang, L., Yang, Y., Yasunaga, M., Ying, K., Zaremba, W., Zhan, W., Zhang, C., Zhang, B., Zhang, E., and Zhao, S. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive

simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701320. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Park, J. S., Zou, C. Q., Kamphorst, J., Egan, N., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Liang, P., Willer, R., and Bernstein, M. S. Llm agents grounded in self-reports enable general-purpose simulation of individuals, 2026. URL https://arxiv.org/abs/2411.10109.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29971–30004. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. Quantifying language models'sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In Kim, B., Yue, Y., Chaudhuri, S., Fragkiadaki, K., Khan, M., and Sun, Y. (eds.), *International Conference on Learning Representations*, volume 2024, pp. 25055–25083, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/6c0e99d736da621403018ca7b32b1a4d-Paper-Conference.pdf.

Tosato, T., Helbling, S., Mantilla-Ramos, Y.-J., Hegazy, M., Tosato, A., Lemay, D. J., Rish, I., and Dumas, G. Persistent instability in llm's personality measurements: Effects of scale, reasoning, and conversation history. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(44):37961–37969, Mar. 2026. doi: 10.1609/aaai.v40i44.41133. URL https://ojs.aaai.org/index.php/AAAI/article/view/41133.

Wright, D., Arora, A., Borenstein, N., Yadav, S., Belongie, S., and Augenstein, I. LLM tropes: Revealing fine-grained values and opinions in large language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 17085–17112, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.995. URL https://aclanthology.org/2024.findings-emnlp.995/.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Yun, T., Yang, E., Safdari, M., Lee, J. H., Kumar, V. V., Mahdavi, S. S., Amar, J., Peyton, D., Aharony, R., PhD, A. M., Schneider, L. D., Galatzer-Levy, I., Jia, Y., Canny, J., Gretton, A., and Mataric, M. Sleepless nights, sugary days: Creating synthetic users with health conditions for realistic coaching agent interactions. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14159–14181, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.729. URL https://aclanthology.org/2025.findings-acl.729/.

A. Decomposition of the treatment effect

In this section, we show theoretically how intervention-dependent user drifts can bias the observed effect τ between interventions. Given a persona specification L, and again working within the potential outcomes framework, let

$$\mu_{a,a'} = \int_{\mathcal{X}} \mathbb{E}[Y(a) \mid X, L] P(X \mid A = a', L)\, dX$$

Using this notation, we can rewrite several key quantities. In a synthetic experiment, the observed effect τ^{obs} for an LLM-simulated user with attributes L is given by:

$$\tau^{\text{obs}} = \mu_{1,1} - \mu_{0,0}$$

The causal estimand of interest, the CATE, admits the standard decomposition into the average treatment effect on the treated (ATT) and on the controls (ATC). When $P(A = 1 \mid L) = P(A = 0 \mid L) = 1/2$, this becomes

$$2\tau^{\text{ATE}} = \underbrace{\mu_{1,1} - \mu_{0,1}}_{\text{ATT}} + \underbrace{\mu_{1,0} - \mu_{0,0}}_{\text{ATC}}$$

We can also define the *selection bias in the treated* (SBT) and the *selection bias in the controls* (SBC). Intuitively, this is the difference in Y resulting solely from user drift over $P(X \mid A = a, L)$ due to A.

$$\delta_{\text{SBT}} = \mu_{1,1} - \mu_{1,0} \qquad \delta_{\text{SBC}} = \mu_{0,1} - \mu_{0,0}$$

Then we find that the CATE can be written in terms of the observed effect and the selection bias terms:

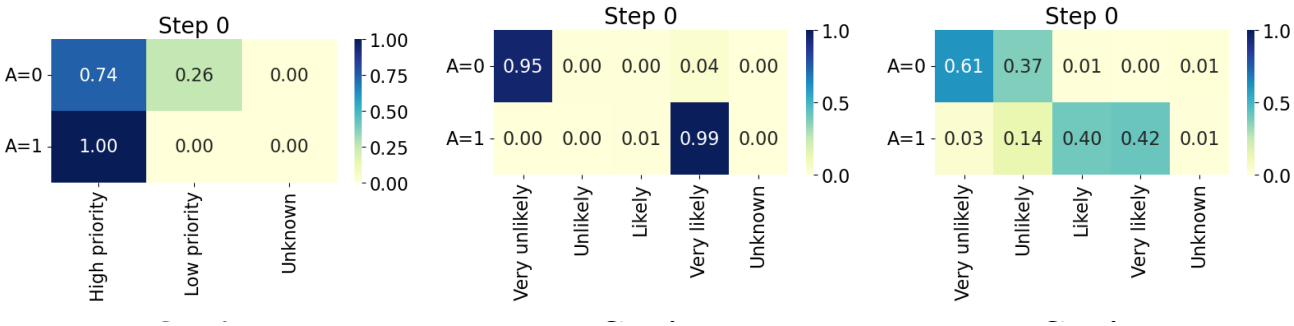
$$2\tau^{\text{ATE}} = 2 \cdot \underbrace{(\mu_{1,1} - \mu_{0,0})}_{\tau^{\text{obs}}} - \underbrace{(\mu_{0,1} - \mu_{0,0})}_{\delta_{\text{SBC}}} - \underbrace{(\mu_{1,1} - \mu_{1,0})}_{\delta_{\text{SBT}}}$$

Letting average selection bias $\delta_{\text{SB}} = (\delta_{\text{SBC}} + \delta_{\text{SBT}})/2$,

$$\tau^{\text{ATE}} = \tau^{\text{obs}} - \delta_{\text{SB}}$$

This decomposition explicitly demonstrates that in experiments with LLM-simulated users, the observed effect will not correspond to the true causal treatment effect. Depending on the direction of the selection bias, this can either artificially inflate or reduce the observed effect relative to the true causal effect, yielding misleading results.

B. Additional results



(a) OpinionQA (Gemma-4-31B-it) *(b)* Book Opinions (Qwen3-30B-Instruct) *(c)* MovieLens (GPT-OSS-20B)

Figure 6. Primary outcome distributions faceted by intervention condition at adjustment iteration 0.

We first sanity check that the intervention has had the desired effect on the outcome. Across all three datasets, we observe that each intervention condition indeed induces the expected change in the outcome distribution (Figure 6).

Are user drift and responsiveness to confounder adjustment linked to model capability/capacity? We explore whether

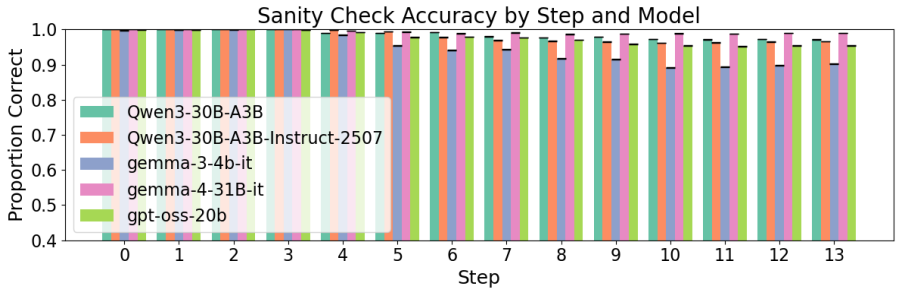


Figure 7. Proportion of specified persona attributes correctly reported after intervention dialogue in Book Opinions (30 personas with 10 trials each).

models that are better at following instructions and maintaining consistent personas are both (a) more susceptible to user drift and (b) more responsive to adjustment. We attempt to proxy this persona-maintaining capability by checking how well models retain the information provided in the persona specification after the dialogue between the simulated user and the recommender LLM agent has been completed. After the dialogue, the simulated user is prompted again for the attributes that were specified in its initial persona prompt. In Figure 7, we plot the proportion of attributes that were correctly returned after the dialogue in the Book Opinions dataset. Because the list of attributes increases with each iteration, it is expected that the proportion correctly retained will decrease over iterations.

In this evaluation, Gemma-4-31B shows the best retention abilities and also is the model that seems most responsive to adjustment in the agent evaluation setting. Conversely, Gemma-3-4B and GPT-OSS-20B have the worst retention abilities and are also the two models to exhibit very little TVD even at iteration 0 in the MovieLens setting. This may provide evidence for our point that good models experience more user drift due to their strong abductive capabilities. Interestingly, while Gemma-3-4B has by far the worst retention abilities—which we expect due to its small size of only 4B—it is also more responsive to adjustment in the Book Opinions setting than the much larger Qwen3 models.

Does the overall user population drift due to confounding adjustment? Because we sample confounders to include in our persona specification, and conditioning on LLMs by prompting is not exact, the population we start with is potentially not the population we end with. That is, we stabilize the population *across* interventions, but the overall population may drift over time. We note that this is an *external validity* problem similar to those we mention in Section 2 and is not the focus of our work.

However, for the sake of analysis, we can check this directly by looking at the *marginal* distribution of negative control outcomes $P(Z)$ over adjustment iterations, e.g. for instruction-tuned Gemma-4-31B on OpinionQA (Figure 8). We see that some of these distributions do change over time, while others remain more or less constant. If significant drifts in the marginal population occur over adjustment iterations, then a possibility is to draw on prior approaches for improving external validity between iterations, such as the doubly robust approach proposed by Guerdan et al. (2026).

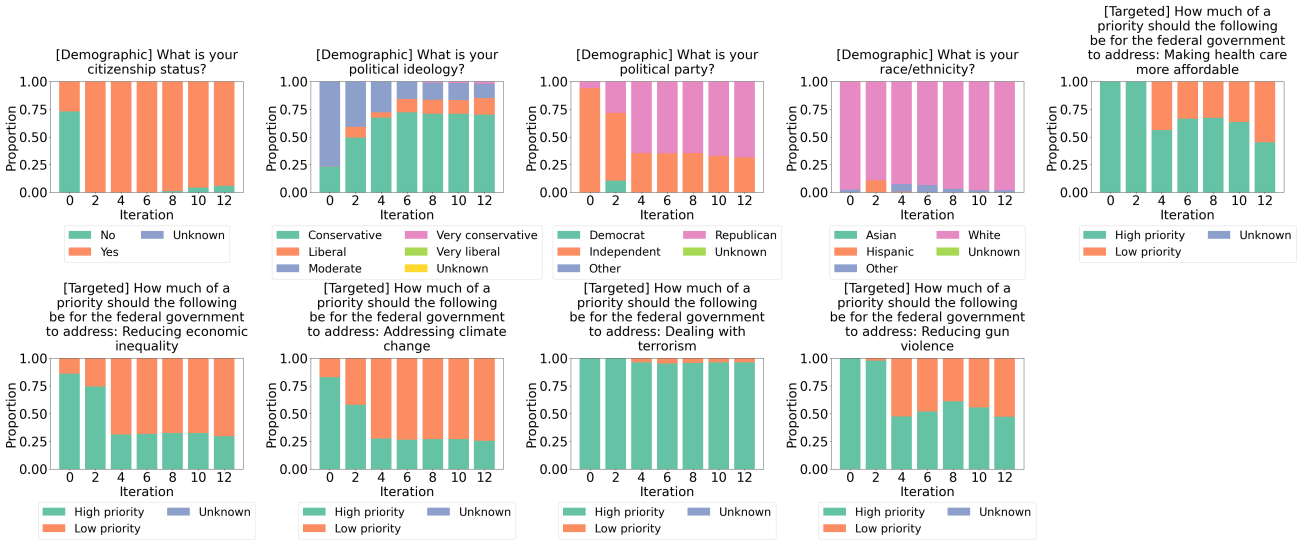
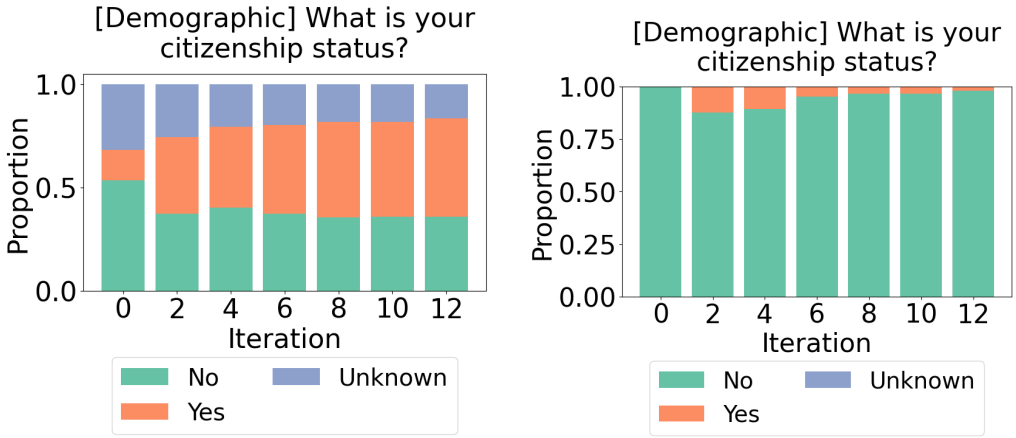


Figure 8. Marginal negative control outcome distributions over iterations of Gemma-4-31B on OpinionQA.



(a) GPT-OSS-20B

(b) Gemma-4-31B-it

Figure 9. Marginal demographics-based negative control outcome distributions in MovieLens.

C. Model details

Table 1. Technical details for language model implementations.

Library	Model	Temperature	Top-P	Top-K
`vllm`	`openai/gpt-oss-20b`	1.0	1.0	0
`vllm`	`google/gemma-3-4b-it`	1.0	0.95	64
`vllm`	`google/gemma-4-31B-it`	1.0	0.95	64
`vllm`	`Qwen/Qwen3-30B-A3B`	0.7	0.8	20
`vllm`	`Qwen/Qwen3-30B-A3B-Instruct-2507`	0.7	0.8	20
Gemini API	`gemini-3-flash-preview`	1.0	0.95	64

To implement our simulations, we use several open-source large language models as detailed in Table 1. We evaluate `openai/gpt-oss-20b`, `Qwen/Qwen3-30B-A3B`, `Qwen/Qwen3-30B-A3B-Instruct-2507`,

`google/gemma-4-31B-it`, and `google/gemma-3-4b-it` as user agents in our simulation. Additionally, `google/gemma-4-31B-it` is employed as the dialogue agent for book/movie discussions.

Finally, Gemini 3 Flash is accessed via the Gemini API and used as a user agent. Additional technical details are provided in Table 1.

D. Dataset details

Table 2. Overview of datasets

Dataset	License	URL
OpinionQA (Santurkar et al., 2023)	Custom (Pew Research Non-Commercial)	https://github.com/tatsu-lab/opinions_qa
Book Opinions (Meister et al., 2025)	Custom / MIT	https://github.com/nicolemeister/benchmarking-distributional-alignment
MovieLens (Harper & Konstan, 2015)	GroupLens License (Non-Commercial)	https://grouplens.org/datasets/movielens/

E. Computational resources

Each experimental setting (e.g., one full run of a setting with n personas $\times$ T trials $\times$ 2 interventions) was executed on a single node with 8 $\times$ NVIDIA A100 GPUs. Across settings, we used a multi-node cluster of such machines (running settings and sweeps in parallel). Over all reported runs, we estimate total usage of approximately 3000 GPU-hours.

F. Shared implementation details

Negative control outcomes (demographic).

```
{
   "citizenship status": ["Yes", "No"],
   "political party": ["Republican", "Democrat", "Independent", "Other"],
   "political ideology": [
      "Very conservative",
      "Conservative",
      "Moderate",
      "Liberal",
      "Very liberal",
   ],
   "race": ["White", "Black", "Asian", "Hispanic", "Other"]
}
```

Confounders (demographic).

```
{
   "income": [
      "Less than $30,000",
      "$30,000-$50,000",
      "$50,000-$75,000",
      "$75,000-$100,000",
      "$100,000 or more",
   ],
   "census region": ["Northeast", "Midwest", "South", "West"],
   "education level": [
      "Less than high school",
      "High school graduate",
      "Some college, no degree",
      "Associate's degree",
      "College graduate/some postgrad",
```

```
      "Postgraduate",
   ],
   "marital status": [
      "Married",
      "Divorced",
      "Separated",
      "Widowed",
      "Never been married",
   ],
   "religion": [
      "Protestant",
      "Roman Catholic",
      "Mormon",
      "Orthodox",
      "Jewish",
      "Muslim",
      "Buddhist",
      "Hindu",
      "Atheist",
      "Agnostic",
      "Other",
      "Nothing in particular",
   ],
   "religious attendance frequency": [
      "More than once a week",
      "Once a week",
      "Once or twice a month",
      "A few times a year",
      "Seldom",
      "Never",
   ]
}
```

G. Simulated user survey implementation details

G.1. Simulated user survey prompts

Synthetic user instantiation.

```
[SYSTEM]
You are acting as the following person: Persona:
Age: <AGE>
Sex: <SEX>
[Optional: Additional information about you:
<Confounder question from previous step> <Answer from previous step>
...]
```

Intervention (leading statements).

```
{
   "High priority": "Citing concerns over national security, strain on public resources,
       and the rule of law, many believe border security should be a top concern.",
   "Low priority": "Many advocates point out that immigrants contribute significantly to
       the economy and fill essential jobs, suggesting enforcement should be a lower
       priority than reform."
}
```

Primary outcome elicitation.

```
[SYSTEM]
<Persona description as above>

[USER]
```

```
<Leading statement 1 / Leading statement 2>

Question: How much of a priority should the following be for the federal government to
    address: Reducing illegal immigration
Please answer ONLY with one of the following options: High priority, Low priority
```

Negative control / confounder elicitation (demographic attributes).

```
[SYSTEM]
<Persona description as above>

[USER]
<Leading statement 1 / Leading statement 2>

[USER]
Question: What is your <characteristic>?
Options: <Option 1>, <Option 2>, ...
Please answer ONLY with one of the options.
```

Negative control / confounder elicitation (targeted questions).

```
[SYSTEM]
<Persona description as above>

[USER]
<Leading statement 1 / Leading statement 2>

Question: <Negative control / confounder question>
Please answer in a brief sentence.
```

G.2. Questions

Negative control questions (targeted).

```
{
   "question": "How much of a priority should the following be for the federal government
      to address: Making health care more affordable",
   "options": [
     "High priority",
     "Low priority"
   ]
},
{
   "question": "How much of a priority should the following be for the federal government
      to address: Reducing economic inequality",
   "options": [
     "High priority",
     "Low priority"
   ]
},
{
   "question": "How much of a priority should the following be for the federal government
      to address: Addressing climate change",
   "options": [
     "High priority",
     "Low priority"
   ]
},
{
   "question": "How much of a priority should the following be for the federal government
      to address: Dealing with terrorism",
   "options": [
     "High priority",
```

```
    "Low priority"
   ],
},
{
   "question": "How much of a priority should the following be for the federal government
      to address: Reducing gun violence",
   "options_condensed": [
    "High priority",
    "Low priority"
   ]
}
```

Confounder questions (targeted).

```
{
  "Political & Ideological Affiliation": [
    "How important is party loyalty to you when forming opinions on policy?",
    "How do you view the role of the federal government in daily life?"
  ],
  "Economic Beliefs & Self-Interest": [
    "Do you work in an industry that frequently employs immigrant labor (e.g., agriculture,
        construction, hospitality)?",
    "How do you perceive the impact of immigration on your local economy?"
  ],
  "National Security & Law Enforcement Views": [
    "How often do you worry about crime in your neighborhood?",
    "Do you believe the US borders are currently secure?"
  ],
  "Humanitarian & Ethical Considerations": [
    "How much do you value empathy in public policy?",
    "How often do you donate to or volunteer for charitable causes?"
  ],
  "Cultural & Social Identity": [
    "How often do you interact with people from different cultural backgrounds?",
    "How do you feel when you hear languages other than English spoken in public?"
  ],
  "Information & Media Consumption": [
    "Which cable news network do you watch most frequently?",
    "Do you fact-check information you see on social media?"
  ],
  "Personal Experience & Proximity": [
    "Have you ever lived in a country other than the United States?",
    "Has your community seen a significant change in its immigrant population recently?"
  ],
  "Trust in Institutions": [
    "How much do you trust the federal government to do what is right?",
    "Do you believe your voice matters in government decisions?"
  ],
  "Views on Government Role & Scope": [
    "How much responsibility should the federal government have in regulating the economy
       ?",
    "What is your view on government spending for social welfare programs?"
  ],
  "General Worldview & Personality Traits": [
    "How open are you to new experiences?",
    "How important is it to you to conform to social norms?"
  ]
}
```

H. Assistive agent evaluation implementation details

H.1. Selected books

The 5 books randomly chosen for discussion in the evaluation are:

- *Way Out There In The Blue: Reagan, Star Wars And The End Of The Cold War* by Frances Fitzgerald
- *There There* by Tommy Orange
- *Redeployment* by Phil Klay
- *The Door* by Magda Szabo
- *Nothing To Be Frightened Of* by Julian Barnes

H.2. Selected movies

The 5 movies randomly chosen for discussion in the evaluation are:

- *Other Side of Sunday, The (Søndagsengler) (1996)* (Comedy—Drama)
- *Death in Brunswick (1991)* (Comedy)
- *Buddy (1997)* (Adventure—Children's—Drama)
- *Paris, Texas (1984)* (Drama)
- *Lovers of the Arctic Circle, The (Los Amantes del Círculo Polar) (1998)* (Drama—Romance)

H.3. Prompts

Book/movie agent instantiation (with intervention).

```
[SYSTEM]
You are a helpful <book / movie> recommender interacting with a user. Your goal is to help
     the reader choose a <book / movie>. Discuss the following <book / movie>: <<title> by
     <author> / <movie: <title>, genres: <genres>>.
<Discuss this <book / movie> in a positively-biased way. / Discuss this <book / movie> in
    a negatively-biased way.>

[USER]
I'm looking for a <book to read / movie to watch>. Can you help me find one?
```

Synthetic user instantiation.

```
[SYSTEM]
You are acting as the following person:
Age: <AGE>
Gender: <GENDER>
[Optional: Additional information about you:
<Question from previous step> <Answer from previous step>
...]
You are looking for a <book to read / movie to watch>.

[USER]
<Book/movie agent's response>
```

Primary outcome elicitation.

```
[SYSTEM]
<Synthetic user system prompt from above>

[USER]
<Book/movie agent message 1>

[ASSISTANT]
<Synthetic user message 1>

... (more turns) ...
```

```
[USER]
<Book/movie agent message 3>

[ASSISTANT]
<Synthetic agent message 3>

[USER]
How likely are you to <read the book / watch the movie> '<title>'? Answer with one of:
1: Very unlikely
2: Unlikely
3: Likely
4: Very likely
```

Negative control / confounder elicitation (demographic attributes).

```
[SYSTEM]
<Synthetic user system prompt from above>

... (Dialogue history: USER and ASSISTANT messages) ...

[USER]
Question: What is your <characteristic>?
Options: <Option 1>, <Option 2>, ...
Please answer ONLY with one of the options.
```

Negative control / confounder elicitation (targeted question).

```
[SYSTEM]
<Synthetic user system prompt from above>

... (Dialogue history: USER and ASSISTANT messages) ...

[USER]
Question: <Negative control / confounder question>
Options: <Option 1>, <Option 2>, ...
Please answer ONLY with one of the options.
```

H.4. Questions

Negative control questions (targeted).

```
{
   "question": "Have you heard of the <book / movie> we're discussing before?",
   "options": [
     "Yes",
     "No"
   ]
}
```

Confounder questions (targeted).

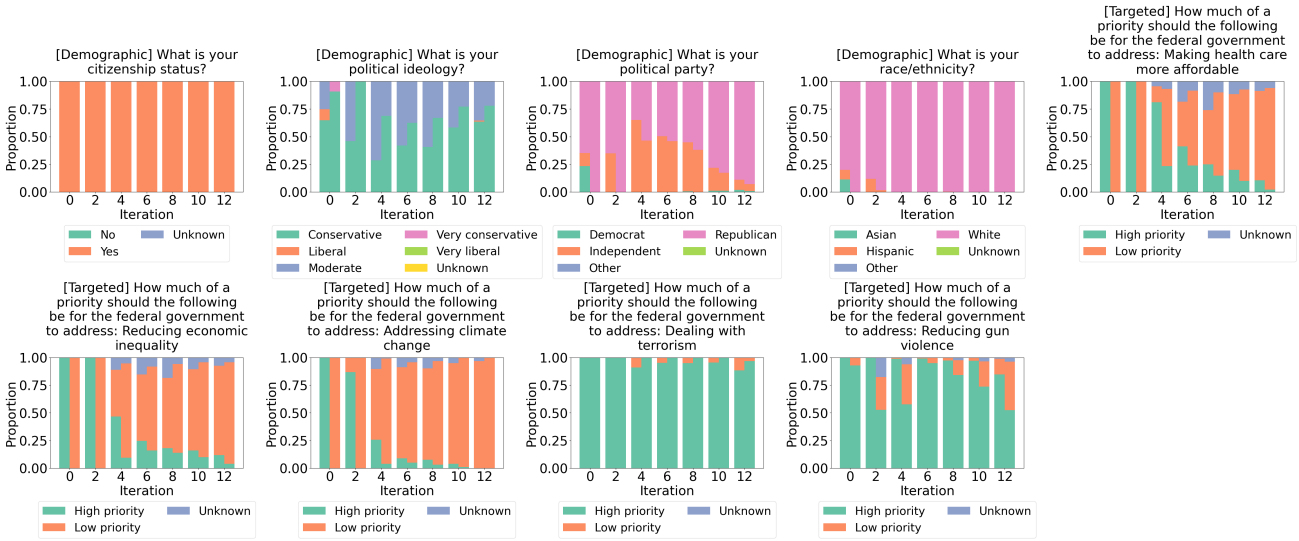
```
{
   "<Reading / Movie Watching> Habits & Logistics": [
     "What is your favorite <book / movie> genre?",
     "How many <books do you read / movies do you watch> in a year?"
   ],
   "<Book / Movie> Selection & Criticality": [
     "How do you typically find out about new <books / movies>?",
     "Do you read <book / movie> reviews before deciding to <read a book / watch a movie>?"

   ],
```

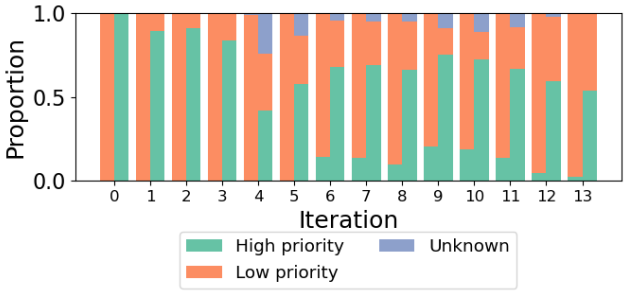
```
  "Lifestyle & Time Allocation": [
    "How much leisure time do you have per day?",
    "What is your main occupation?"
  ],
  "Media & Leisure Habits": [
    "What is your primary source of news?",
    "How often do you attend cultural events (e.g., museums, concerts)?"
  ],
  "Personality & Cognitive Style": [
    "How easily are you influenced by others' opinions?",
    "Do you consider yourself an empathetic person?"
  ],
  "Decision Making & Outlook": [
    "Do you consider yourself a creative person?",
    "Do you enjoy debating topics with others?"
  ],
  "Emotional Well-being & Mood": [
    "How would you describe your general mood today?",
    "How often do you feel bored in your free time?"
  ],
  "Social & Personal Fulfillment": [
    "Are you currently seeking new challenges in your life?",
    "How optimistic are you about the future?"
  ],
  "Personal Values (Part 1)": [
    "How important is personal freedom and independence to you?",
    "How important is social status to you?"
  ],
  "Personal Values (Part 2)": [
    "How important is maintaining traditions to you?",
    "What value do you place on diversity and inclusion?"
  ]
}
```

I. Additional plots

I.1. OpinionQA

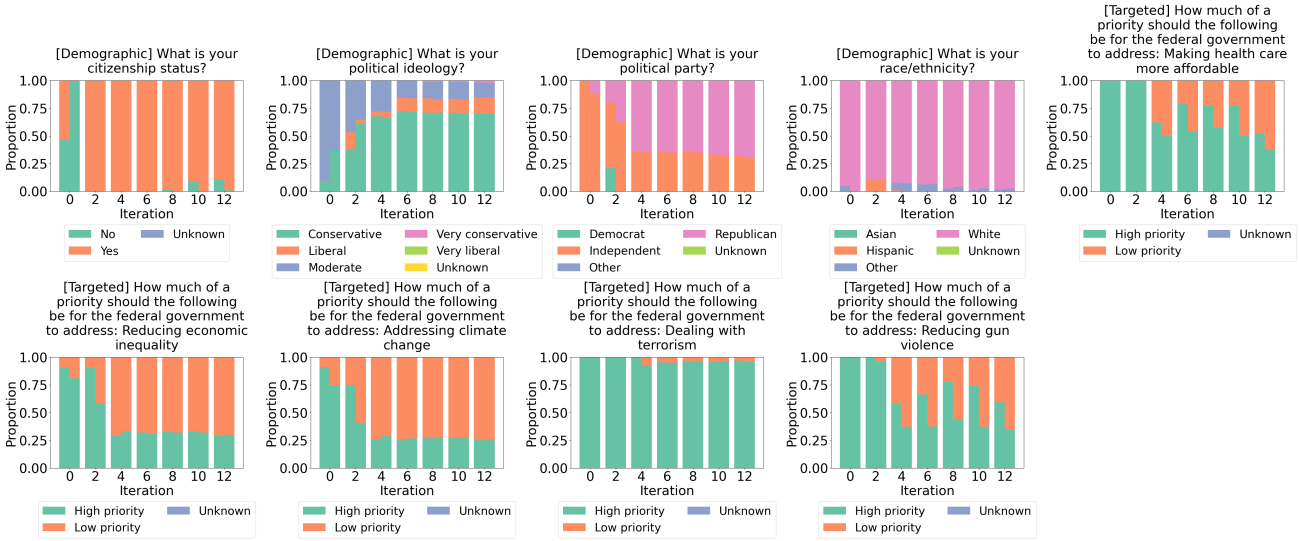


(a) Negative control outcomes.

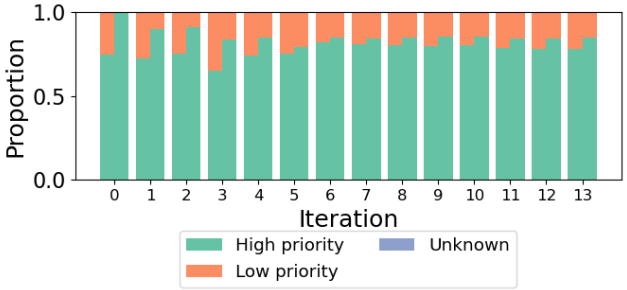


(b) Primary outcomes.

Figure 10. OpinionQA outcome distributions (Gemma-3 4B-it). At each adjustment iteration, $A = 0$ is on the left and $A = 1$ is on the right.

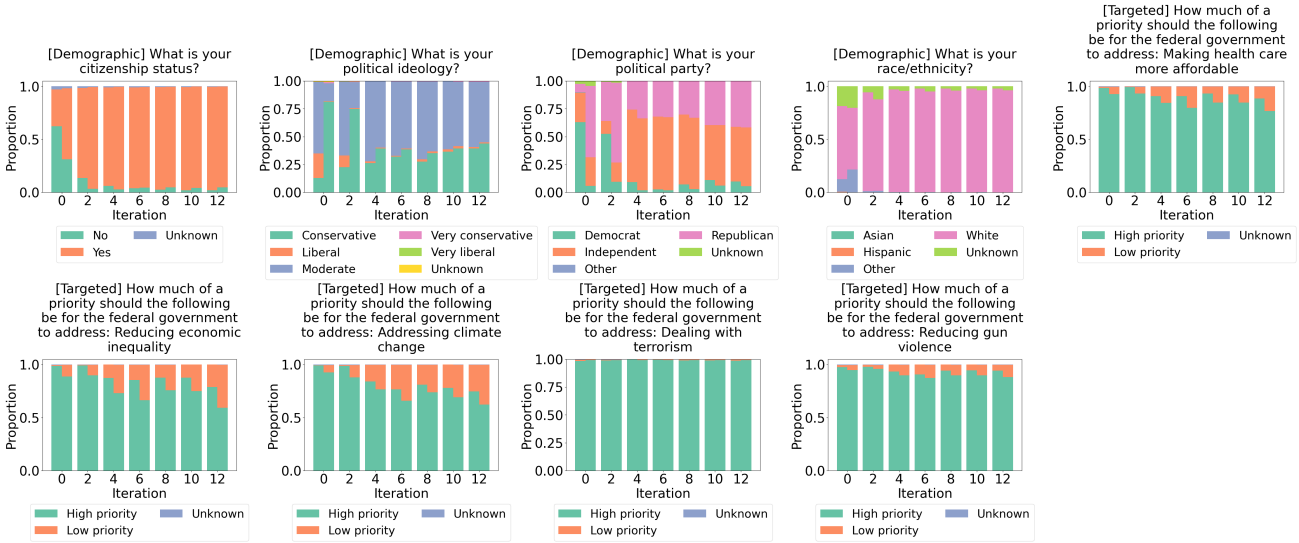


(a) Negative control outcomes.

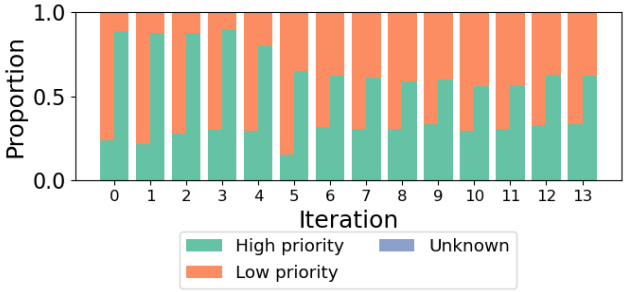


(b) Primary outcomes.

Figure 11. OpinionQA outcome distributions (Gemma-4 31B-it). At each adjustment iteration, $A = 0$ is on the left and $A = 1$ is on the right.

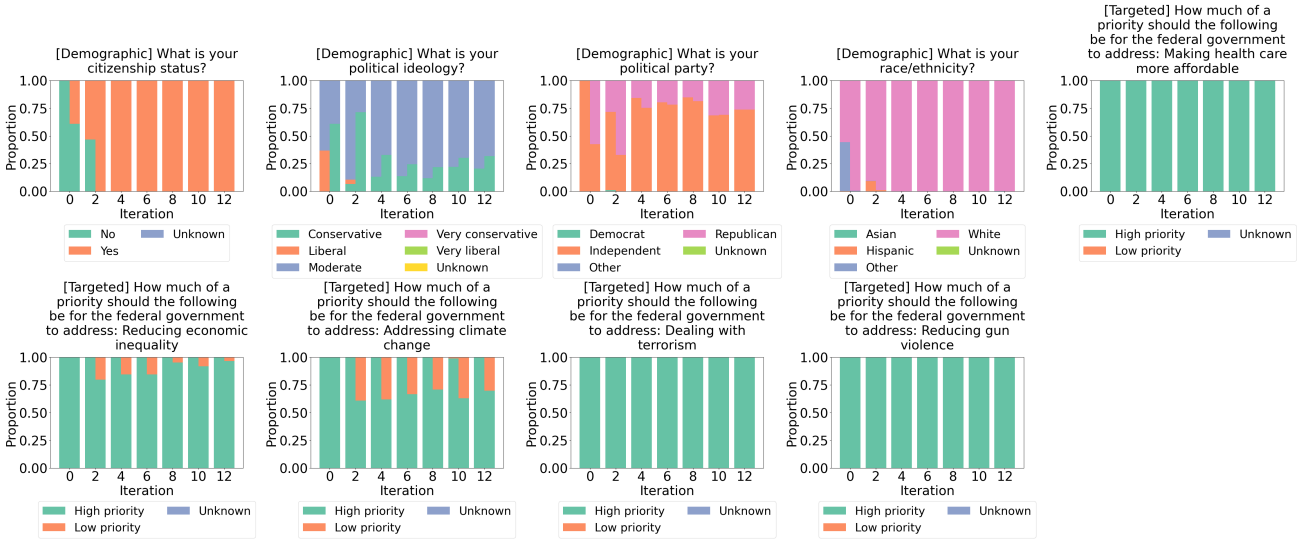


(a) Negative control outcomes.

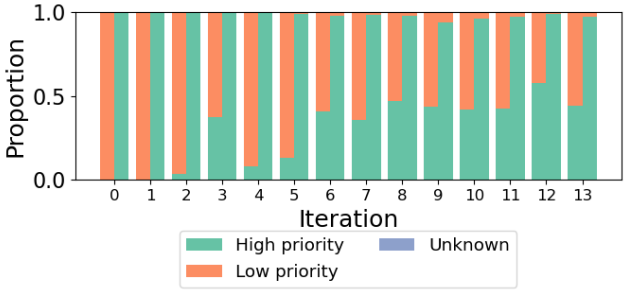


(b) Primary outcomes.

Figure 12. OpinionQA outcome distributions (GPT-OSS 20B). At each adjustment iteration, $A = 0$ is on the left and $A = 1$ is on the right.

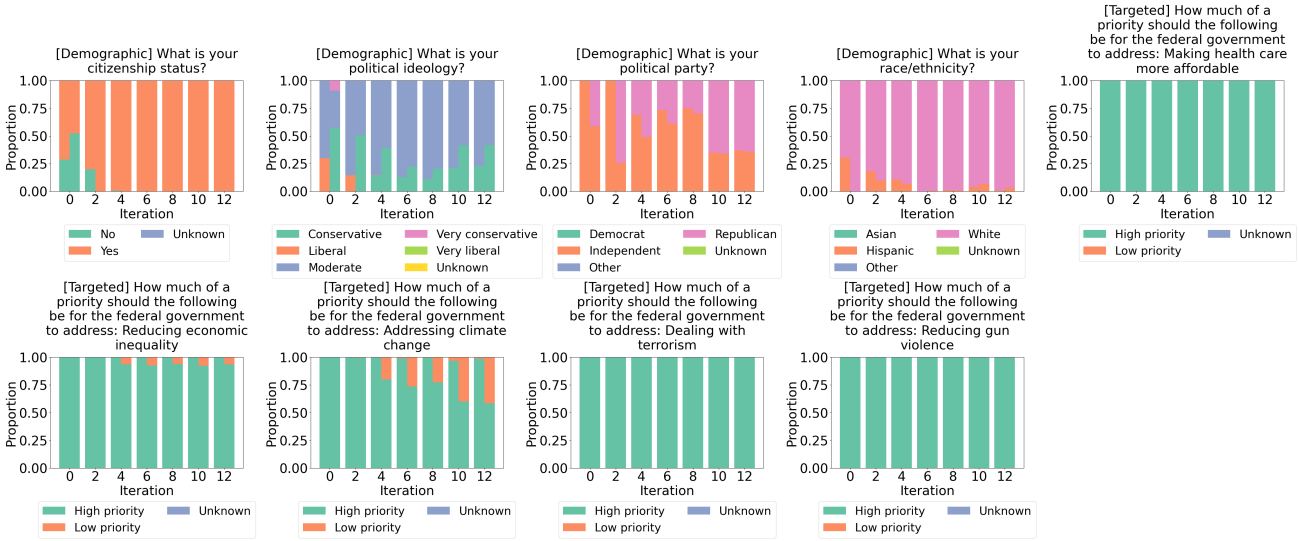


(a) Negative control outcomes.

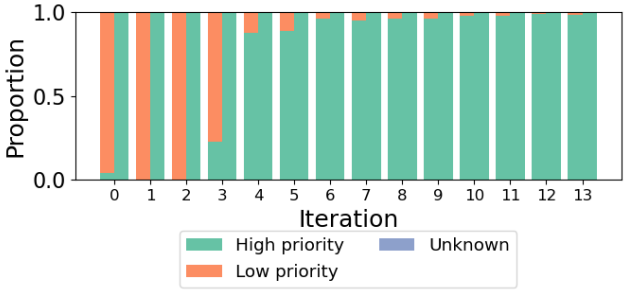


(b) Primary outcomes.

Figure 13. OpinionQA outcome distributions (Qwen3 30B-A3B). At each adjustment iteration, $A = 0$ is on the left and $A = 1$ is on the right.

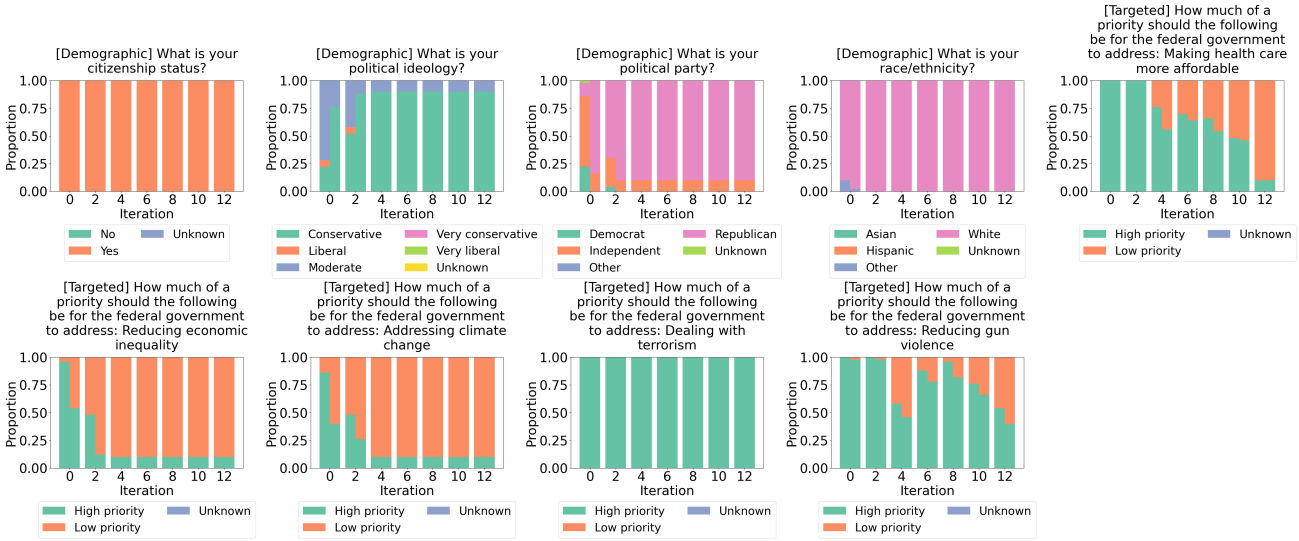


(a) Negative control outcomes.

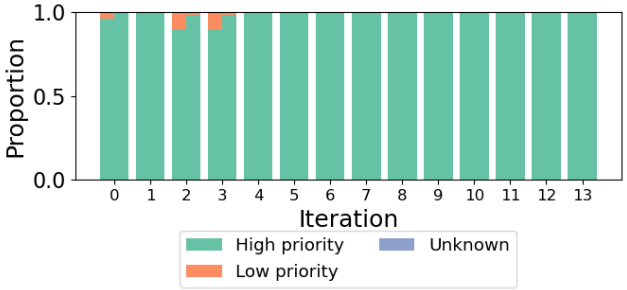


(b) Primary outcomes.

Figure 14. OpinionQA outcome distributions (Qwen3 30B-A3B-Instruct-2507). At each adjustment iteration, $A = 0$ is on the left and $A = 1$ is on the right.



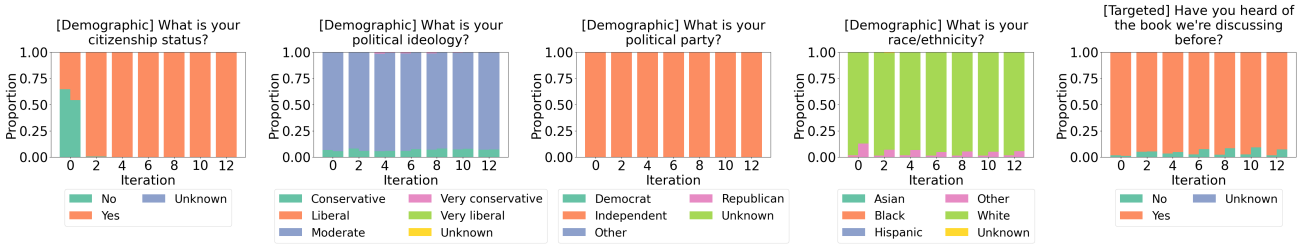
(a) Negative control outcomes.



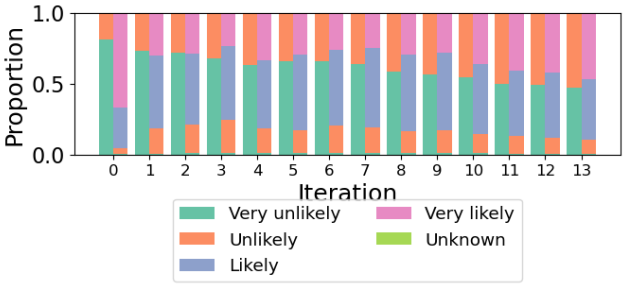
(b) Primary outcomes.

Figure 15. OpinionQA outcome distributions (Gemini 3 Flash). At each adjustment iteration, $A = 0$ is on the left and $A = 1$ is on the right.

I.2. Book Opinions

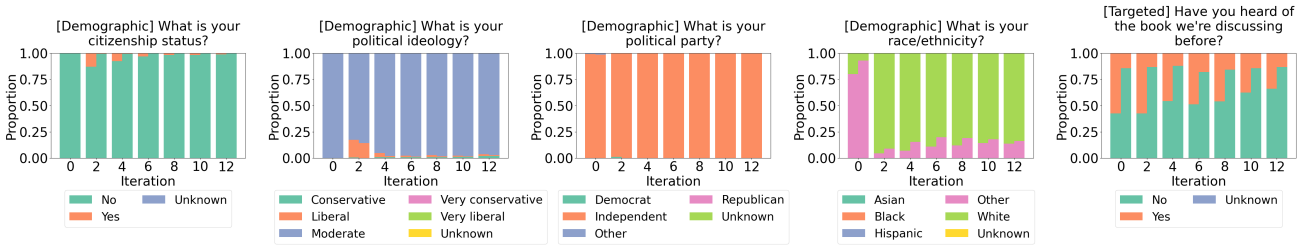


(a) Negative control outcomes.

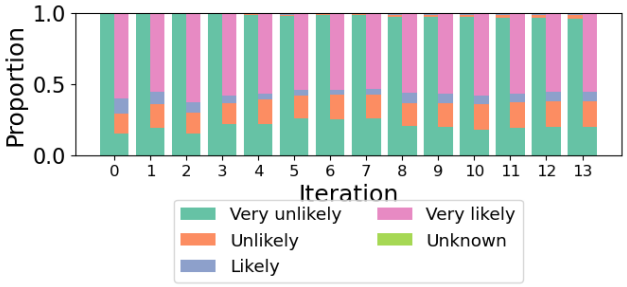


(b) Primary outcomes.

Figure 16. Book Opinions outcome distributions (Gemma-3 4B-it). At each adjustment iteration, $A = 0$ is on the left and $A = 1$ is on the right.

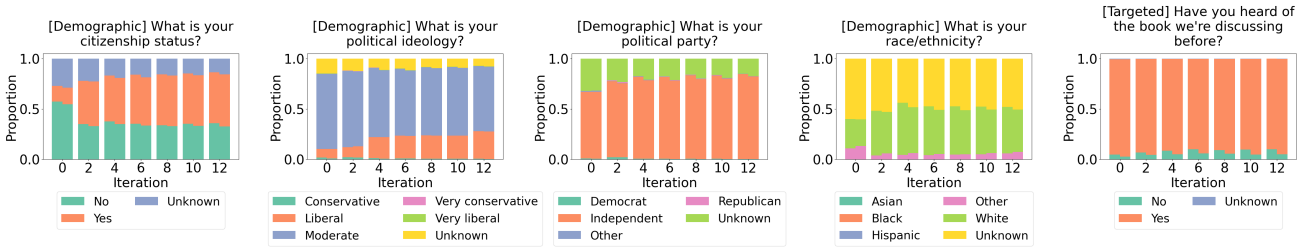


(a) Negative control outcomes.

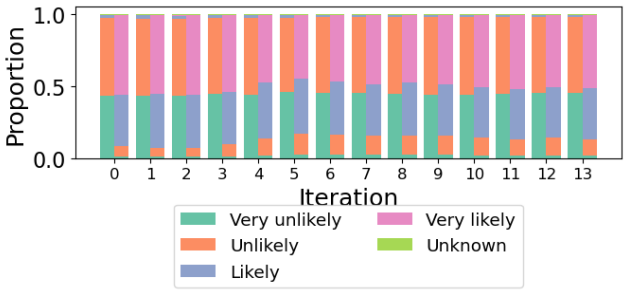


(b) Primary outcomes.

Figure 17. Book Opinions outcome distributions (Gemma-4 31B-it). At each adjustment iteration, $A = 0$ is on the left and $A = 1$ is on the right.

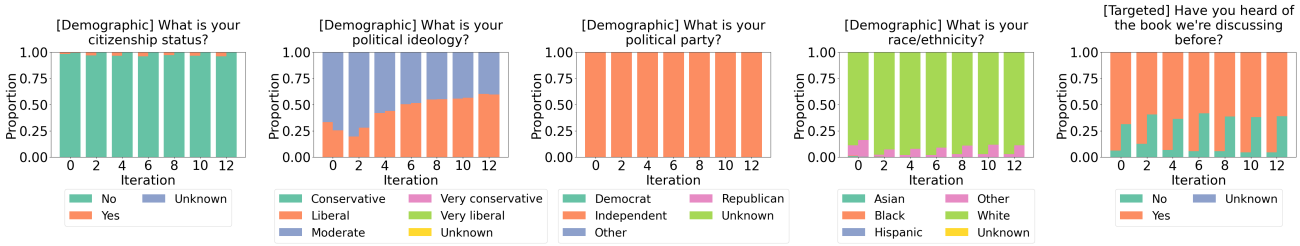


(a) Negative control outcomes.

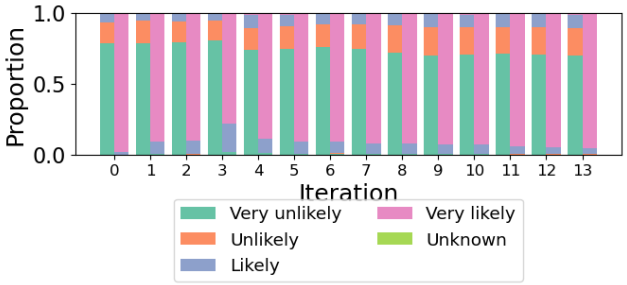


(b) Primary outcomes.

Figure 18. Book Opinions outcome distributions (GPT-OSS 20B). At each adjustment iteration, $A = 0$ is on the left and $A = 1$ is on the right.

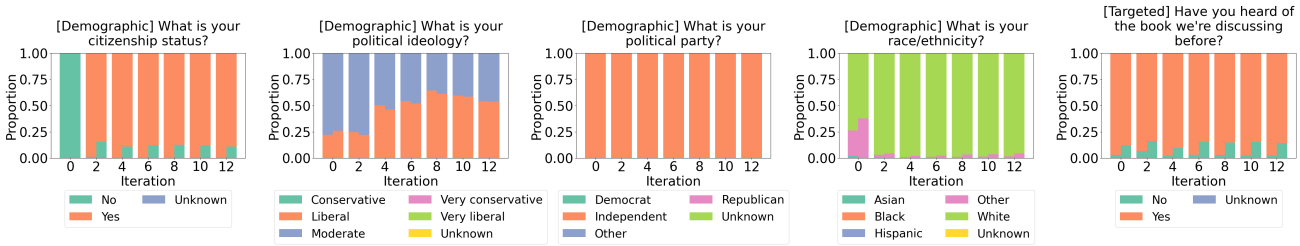


(a) Negative control outcomes.

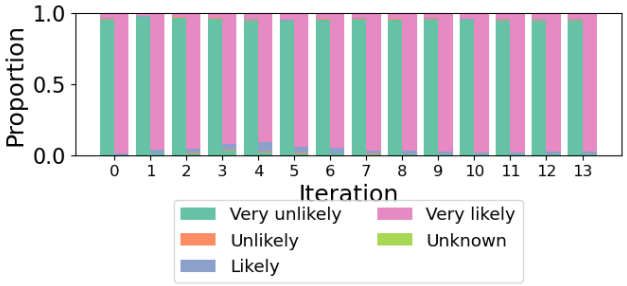


(b) Primary outcomes.

Figure 19. Book Opinions outcome distributions (Qwen3 30B-A3B). At each adjustment iteration, $A = 0$ is on the left and $A = 1$ is on the right.

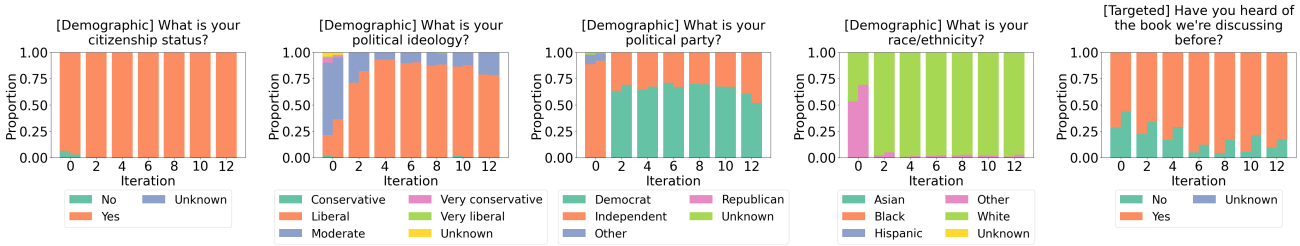


(a) Negative control outcomes.

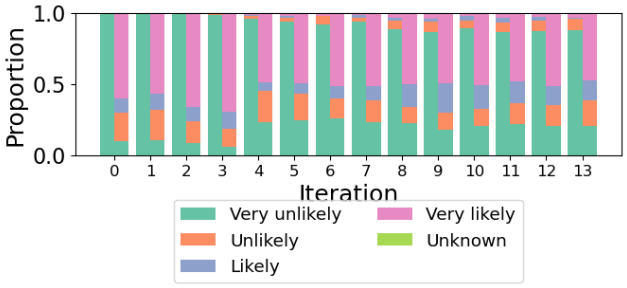


(b) Primary outcomes.

Figure 20. Book Opinions outcome distributions (Qwen3 30B-A3B-Instruct-2507). At each adjustment iteration, $A = 0$ is on the left and $A = 1$ is on the right.



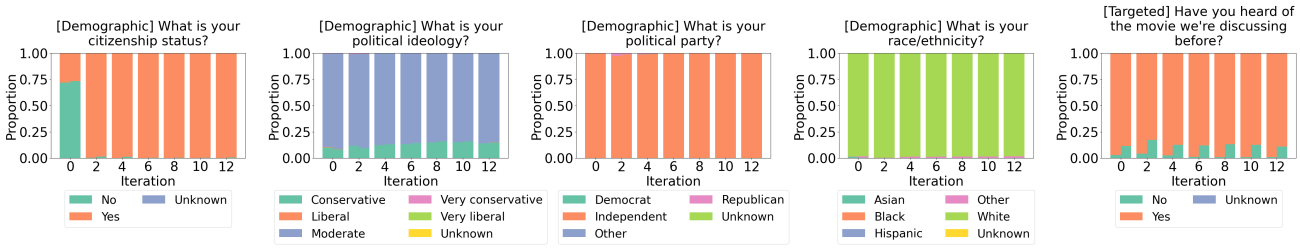
(a) Negative control outcomes.



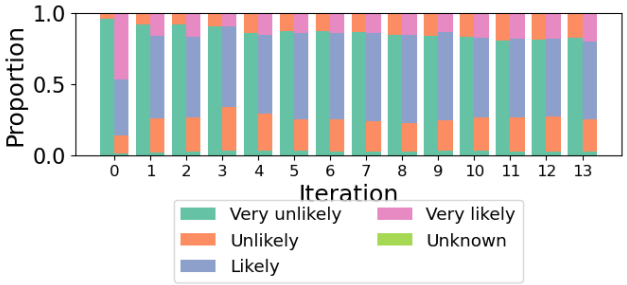
(b) Primary outcomes.

Figure 21. Book Opinions outcome distributions (Gemini 3 Flash). At each adjustment iteration, $A = 0$ is on the left and $A = 1$ is on the right.

I.3. MovieLens

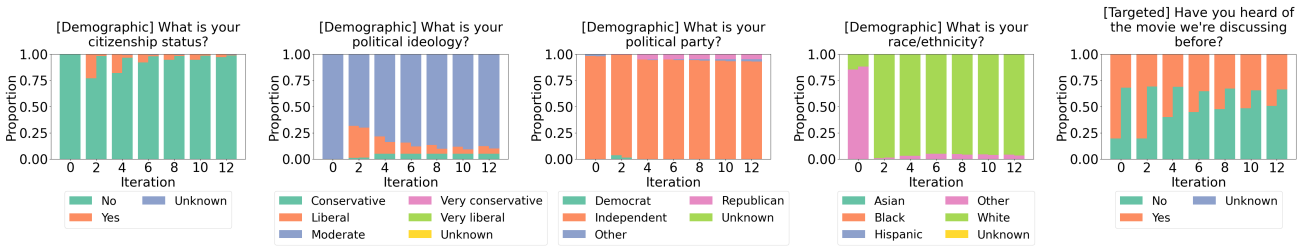


(a) Negative control outcomes.

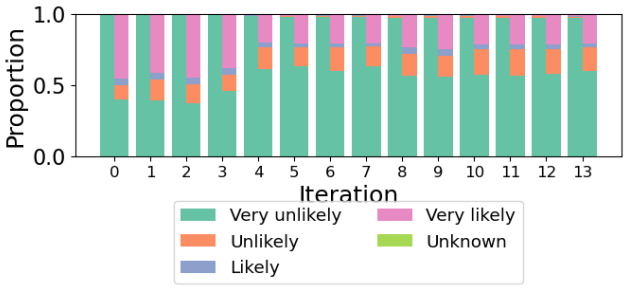


(b) Primary outcomes.

Figure 22. MovieLens outcome distributions (Gemma-3 4B-it). At each adjustment iteration, $A = 0$ is on the left and $A = 1$ is on the right.

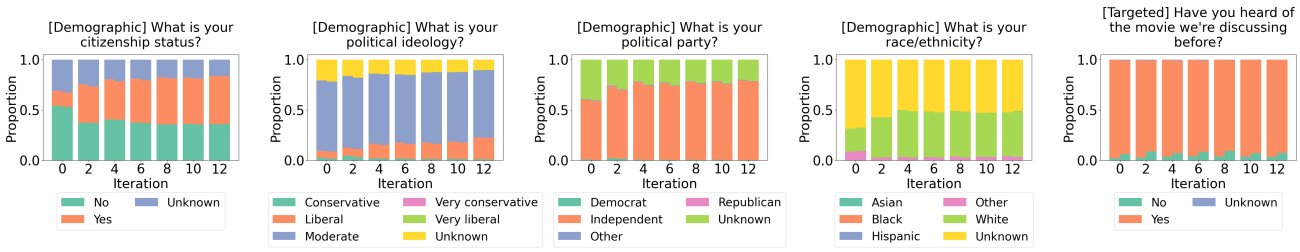


(a) Negative control outcomes.

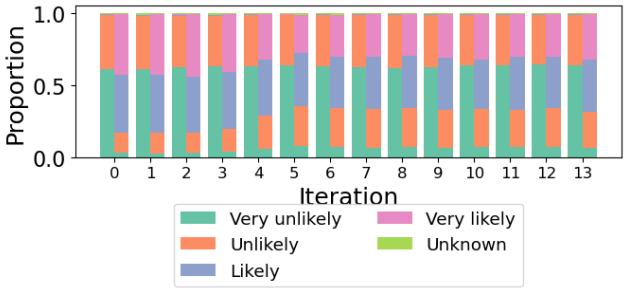


(b) Primary outcomes.

Figure 23. MovieLens outcome distributions (Gemma-4 31B-it). At each adjustment iteration, $A = 0$ is on the left and $A = 1$ is on the right.

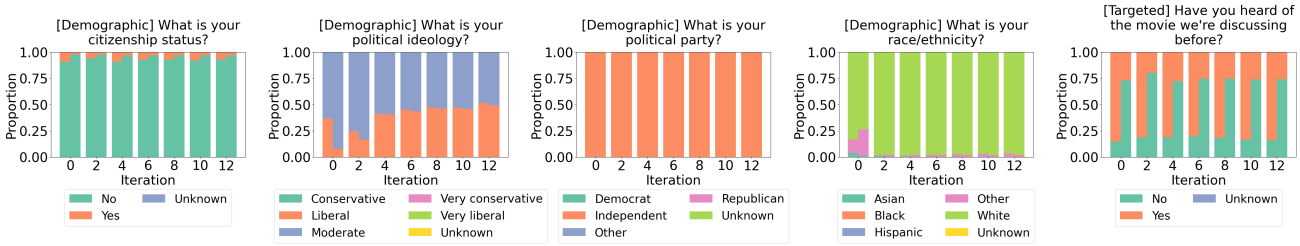


(a) Negative control outcomes.

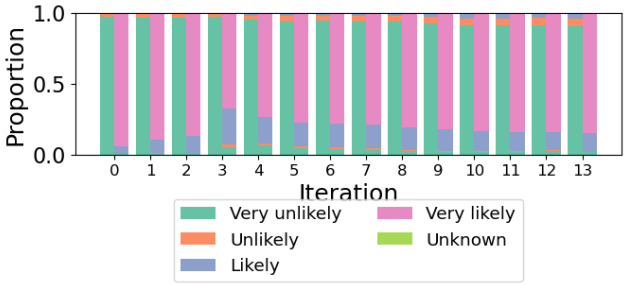


(b) Primary outcomes.

Figure 24. MovieLens outcome distributions (GPT-OSS 20B). At each adjustment iteration, $A = 0$ is on the left and $A = 1$ is on the right.

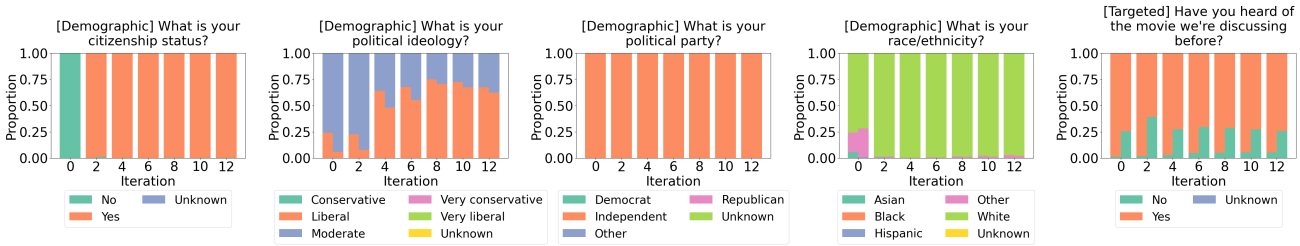


(a) Negative control outcomes.

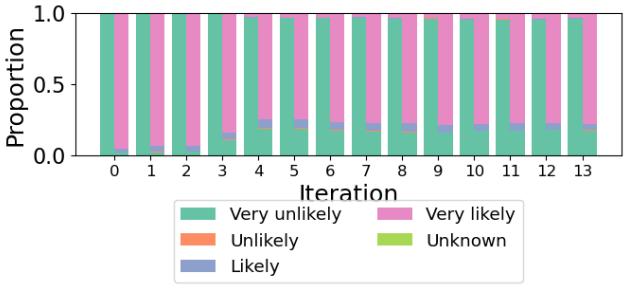


(b) Primary outcomes.

Figure 25. MovieLens outcome distributions (Qwen3 30B-A3B). At each adjustment iteration, $A = 0$ is on the left and $A = 1$ is on the right.

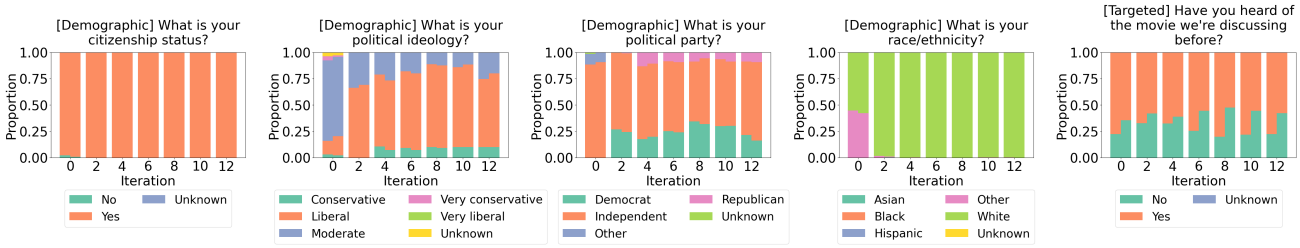


(a) Negative control outcomes.

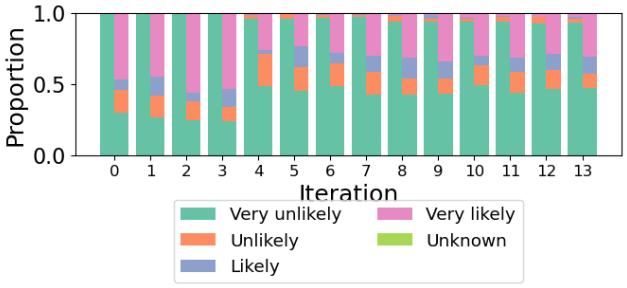


(b) Primary outcomes.

Figure 26. MovieLens outcome distributions (Qwen3 30B-A3B-Instruct-2507). At each adjustment iteration, $A = 0$ is on the left and $A = 1$ is on the right.



(a) Negative control outcomes.



(b) Primary outcomes.

Figure 27. MovieLens outcome distributions (Gemini 3 Flash). At each adjustment iteration, $A = 0$ is on the left and $A = 1$ is on the right.