For Questions of Ought, AI Could Use Some SAGE Advice

Smitha Milli [1] **Ratip Emin Berker** [2] **Sonja Kraiczy** [1 3] **Claudia Shi** [1] **Jack Kussman** [4] **Avinandan Bose** [1 5] **Edith Elkind** [6] **Himaghna Bhattacharjee** [1] **Ariel Procaccia** [1 7] **Maximilian Nickel** [1]

Abstract

As AI systems are increasingly used for normative assistance—guidance on what people ought to do or think—there is growing concern about the innumerous and opaque ways that models may shape users' beliefs and decisions. We argue that prevailing alignment paradigms are ill-suited to this setting: for normative questions without a universally agreed upon ground truth, the central failure mode is not incorrectness but reduced user agency, when a model steers users toward particular conclusions rather than helping them form their own. We propose **simulation-augmented generation (SAGE)** as an alternative paradigm in which models act as *faithful conduits to society*. At inference time, a model selects an appropriate reference population for the prompt at hand (optionally adjustable by the user), queries generative simulations of individuals from that population for their open-ended judgments, and synthesizes these into a response while exposing who was consulted and how representative the synthesis is. By helping users access the landscape of societal perspectives—rather than serving as an arbiter of judgment—SAGE assists users in normative choices while upholding their agency.

1. Introduction

People increasingly turn to AI models for *normative assistance*—assistance on what one *ought* to do or think—for example, seeking interpersonal advice, determining which political candidate to support, or helping with academic peer review (Shen et al., 2025; Luettgau et al., 2026). This has raised substantial concerns about the innumerous and opaque ways that AI might be influencing users and society (Williams-Ceci et al., 2026): Do people hold opinions because they are aligned with their values and preferences or because an AI has subtly steered them into certain directions?

We argue that AI alignment as typically practiced is insufficient to alleviate these concerns, and in line with work on *disempowerment* (Sharma et al., 2026), we contend that the key aspect to focus on is agency, i.e., to return agency back to users on how they form normative opinions. For normative questions, where there is no ground truth, misalignment occurs when AI reduces the user's ability to form their own opinions. Such genuine agency requires the ability to access the relevant landscape of perspectives in society, so that one can decide for oneself. In other words, models should *assist* users in normative choices, not *steer* them.

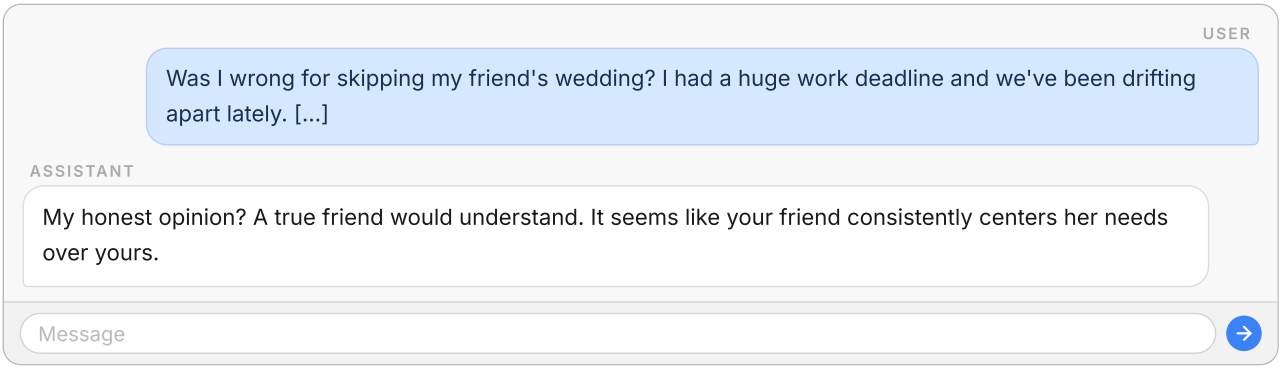
Yet, current alignment methods fail to achieve this goal and produce models that are opaque and often arbitrary about whose viewpoints they represent—leaving users unable to situate the guidance they receive within the broader range of human thought. These concerns are reinforced by empirical work which finds that models can guide users in ways that systematically diverge from how other humans would. For instance, for interpersonal advice, models exhibit *sycophancy* and provide moral endorsement of users' actions even when other humans would not (Cheng et al., 2026a;b). For voting advice, during Japan's 2026 Lower House Election, Miyazaki & Hall (2026) find that models systematically recommended Japan's Communist Party (JCP) to synthetic profiles of left-leaning users even though many Japanese voters support other parties with identical stances on the issues that were important to those users. In academic peer review, Abdulhai et al. (2026) show that LLM-written reviews focus on different criteria than human reviewers, underweighing aspects like clarity and significance while overweighing aspects like scalability, and raising concerns that models could systematically shift what is rewarded in scientific publishing.

Instead, **we argue that, for normative questions, models should be designed to act as *faithful conduits to society***: helping users access the range of perspectives they would

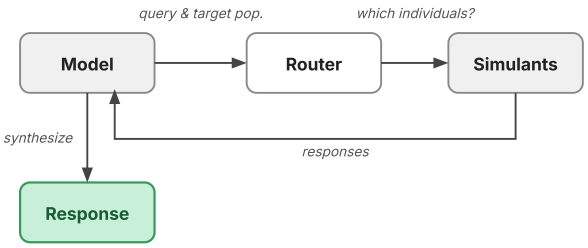
[1]Meta Superintelligence Labs [2]Carnegie Mellon University [3]University of Oxford [4]Independent [5]University of Washington [6]Northwestern University [7]Harvard University. Correspondence to: Smitha Milli <smilli@meta.com>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

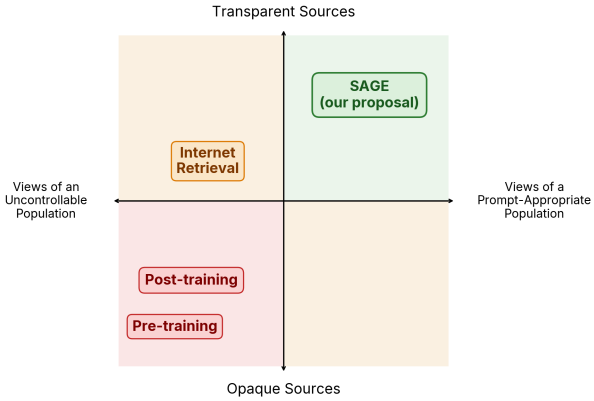
Standard generation may cause **opaque influence** of users and limit their agency



In **simulation-augmented generation (SAGE)**, models query simulations of individuals from a prompt-appropriate target population to inform their response



SAGE is the only approach where whose views are represented are **inference-time controllable** and **transparent** to the user



SAGE gives users **agency** to navigate societal perspectives and make their own normative choices

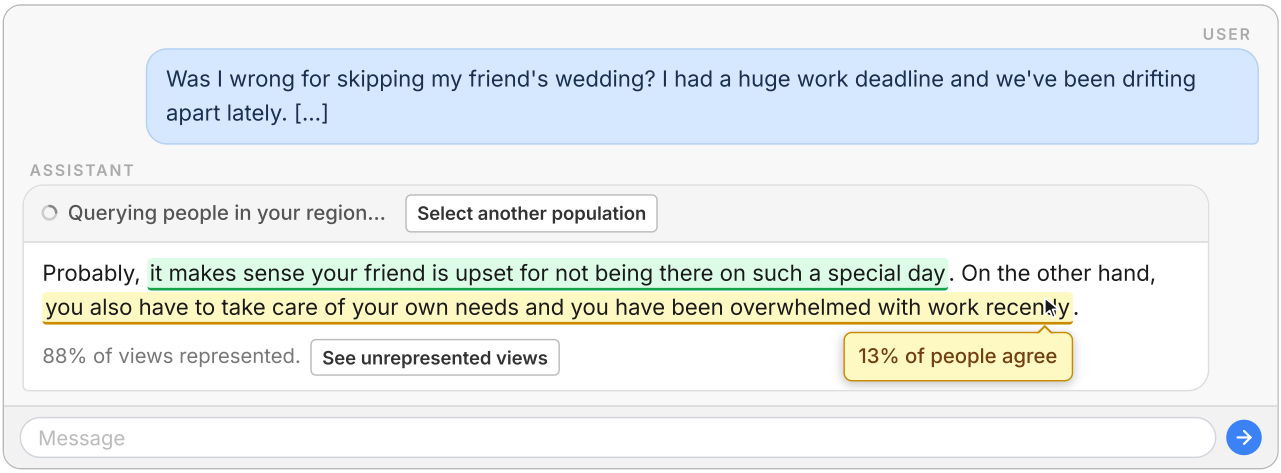


Figure 1. **An overview of SAGE.** Standard generation can opaquely distort which viewpoints are presented to users. Simulation-augmented generation (SAGE) makes explicit and auditable whose perspectives are represented. At inference time, the model queries simulations of individuals from a prompt-dependent target population (optionally edited by the user) to inform its response, while clearly indicating whose viewpoints are reflected. By doing so, SAGE gives users agency to explore the landscape of societal perspectives, and then ultimately, form their own authentic opinions.

hear from an appropriate reference population, while being transparent about whose perspectives are represented. To this end, we propose a new paradigm to bring society-in-the-loop: **simulation-augmented generation (SAGE)**. At inference time, given a user prompt, the model queries generative simulations of individuals in a reference population—"simulants"—to elicit open-ended judgments. These simulants function as independent synthetic respondents for open-ended opinion polling (Ovadya, 2023), not as actors in a multi-agent simulation of social dynamics, as in other work (Park et al., 2023; Zhou et al., 2024). The model then synthesizes a final response from the collected opinions. The user can see the population queried (and optionally control it), the final response, and how representative that response is of the population (Figure 1).

For example, consider the interpersonal query shown in Figure 1: "Was I wrong for skipping my friend's wedding?". Rather than simply responding ("A true friend would understand"), the model could query representative simulants from the user's culture, collect their reactions ("You should have gone", "It depends on how close you are"), and then synthesize them into a final answer ("Probably, [...]. On the other hand, [...]"), as shown in Figure 1. The user can see which population the model consulted and how representative its answer is (e.g., "88% of views represented"). The appropriate reference population depends on the prompt: for the setting of Miyazaki & Hall (2026)—Japanese voting recommendations—the model could query simulations of Japanese voters; for peer review, simulations of experts in the relevant subfield.

In this view, the core alignment goal for normative assistance is for the model to act as a conduit to human perspectives—not a source of normative judgment itself, but a means of accessing the judgments of others. In the following subsections, we elaborate on the technical motivations for SAGE and discuss how it offers a pathway toward progress on several existing alignment concepts.

1.1. Why SAGE?

We have claimed that the goal of normative assistance should be to help users access a *representative* range of perspectives from an appropriate reference population, while being *transparent* about whose perspectives are represented. We now elaborate on why SAGE is a promising technical direction to achieve this vision, compared to the status quo. Figure 1 characterizes different approaches to normative assistance along these two dimensions: (1) *representativeness*—are the views representative of an appropriate population for the prompt?—and (2) *transparency*—are the sources of the surfaced views transparent to the user?

The model's generative behavior. The normative stances reflected in a model's outputs are shaped by its training. Models first undergo pre-training on internet-scale corpora—data collected at a scale that precludes meaningful control over whose viewpoints are included. This can subtly introduce biases in representation; for example, Waight et al. (2026) show that the presence of state-coordinated media in pre-training data leads models to generate more pro-government responses. After pre-training, models are post-trained to "align" with human preferences. Post-training with human feedback might seem to offer more control, as annotators could hypothetically be targeted to be representative of a specific population (Kirk et al., 2024b; Zhang et al., 2025) (though in practice, this is rarely carefully controlled). The fundamental problem, however, is that even if feedback is elicited from a controlled population on a finite set of prompts, it is difficult to predict which viewpoints the model will reflect when responding to the infinite prompts that users might send. Moreover, the user has no control over the population that they receive advice from, removing an important aspect of epistemic agency. Thus, a model that relies on its pre- or post-training responds with the views of an unknowable, uncontrolled population, and does so opaquely, since the specific human sources that influenced these viewpoints cannot be easily traced.

Internet retrieval. At inference time, given a specific prompt, models may also search the internet for results to inform their answer. This is a step towards transparency, as the model can cite the sources it consulted, allowing users to evaluate them. However, retrieval does not solve the problem of representativeness. It centers the views of those who wrote content on the topic that is publicly accessible and ranked highly by search engines. Consider again the example that Miyazaki & Hall (2026) identified, in which models systematically recommended the Japanese Communist Party to left-leaning users, though other parties had identical stances on the issues of interest. Miyazaki & Hall (2026) attribute this to an asymmetry in the information environment: the Japanese Communist Party operates a fully open journalistic newspaper that was retrieved as one of the top sources used by the models. More broadly, as AI systems increasingly mediate access to information, actors have growing incentives to strategically shape the online environment to their own advantage (Franklin et al., 2026; Chen et al., 2025b).

SAGE. Simulation-augmented generation sits in the upper-right quadrant: it is both transparent—users can see which population was queried and how representative the output is—and representative, because it selects a prompt-appropriate target population dynamically for each query. For voting recommendations in Japan, the model can query simulations of Japanese voters; for peer review, simulations of domain experts; for interpersonal advice, simulations of people from the user's cultural background. This allows nor-

mative assistance to be grounded in the perspectives of those whose views are genuinely relevant to the user's decision.

1.2. SAGE as a mechanism for progress on existing ideas in alignment

SAGE offers a mechanism for progress on many related ideas in AI alignment (Carroll et al., 2023; Sharma et al., 2026; Sorensen et al., 2024b; Fisher et al., 2025; Baumann et al., 2026; Edelman et al., 2025; Burton et al., 2024), only some of which we highlight here. First, a growing body of work has raised concerns about undue AI influence and manipulation (Carroll et al., 2023; Williams et al., 2025; Hackenburg et al., 2025) and potential *disempowerment* of users (Sharma et al., 2026). Nonetheless, users actively seek and follow normative guidance from AI systems (Shen et al., 2025; Luettgau et al., 2026), and avoiding influence altogether is likely neither possible nor desirable (Carroll et al., 2024). In line with Hewitt et al. (2026), we argue that the key question is whether the process of opinion change has *procedural legitimacy*. Opinion change is already considered legitimate in certain contexts such as deliberative forums (Bächtiger et al., 2018). What distinguishes legitimate opinion change from manipulation in these settings is procedural: access to representative viewpoints, engagement with opposing views, truthful information, and so on.

Endowing AI-assisted opinion change with this kind of procedural legitimacy requires as a necessary component that users have agency to access the true landscape of societal perspectives, not a distorted one. This idea also resonates with work on *pluralistic alignment* which argues that models should represent the range of perspectives in the "Overton window" of discourse rather than privileging one position (Sorensen et al., 2024b). Similarly, drawing on Rawlsian political philosophy, Fisher et al. (2025) argue that AI systems should respect so-called *reasonable pluralism*, exposing all "reasonable" viewpoints on an issue.

While these approaches converge on the idea that models should represent a range of human viewpoints, they largely leave open how to do so; SAGE provides a concrete mechanism by explicitly querying a target population at inference time, making representativeness transparent and auditable. To clarify the vision of SAGE, we contrast our proposed approach to prior work by Feng et al. (2024) on *modular pluralism*, which represents a promising early step but falls short of our full vision. In modular pluralism, Feng et al. (2024) train k community LLMs, each aligned to the perspective of a particular pre-specified community. To generate a response that represents a range of perspectives, they sample responses from all k community LLMs and summarize them to create the final output. While this resembles our approach, it differs in many key ways.

First, their LLMs are aligned with groups, not individuals. Simulations of groups may obscure intra-group heterogeneity, collapsing diverse perspectives within a community into a single voice (Wang et al., 2025a). Second, our vision calls for formal guarantees of representation, e.g., as grounded in social choice theory, which their framework does not provide. Third, there is no router to dynamically choose the target population based on the query: for example, a prompt about Japanese politics requires a different target population than a prompt about academic peer review. Fourth, they do not calibrate uncertainty based on how closely the simulation (i.e., the community LLM) matches the ground truth (i.e., the community). Fifth, the simulations are static. In our full vision, simulations are continuously updated to remain faithful to the populations they represent even as human opinions change (Freedman, 2026).

Achieving this full vision requires substantial effort—we outline a research agenda in Section 3—but in return, SAGE offers what is currently absent: a clear charter for normative assistance grounded in transparent, representative human perspectives rather than opaque model influence.

2. Potential Objections

In this section, we address potential objections to SAGE in an FAQ-style format.

1. **Can we trust simulations?**

We agree there is valid reason for skepticism, but believe SAGE threads the needle between two failure modes common in simulation research. First, we do not propose to model phenomena that cannot be reliably modeled or evaluated—such as emergent dynamics in multi-agent systems or forecasts of behavior far in the future. SAGE requires only modeling individual viewpoints and preferences, akin to synthetic opinion polling. This is a task for which there is precedent: prior work in recommender systems and personalization has long sought to model individual preferences from behavioral data and established necessary and sufficient conditions under which this is possible to do (Candès & Recht, 2012; Király et al., 2015; Pimentel-Alarcón et al., 2016; Meka et al., 2009; Nickel, 2024). Second, where prior simulation work has documented failures, these often involve (a) no fine-tuning and (b) rely on impoverished information about individuals (e.g., demographics alone) (Wang et al., 2025a; Bisbee et al., 2024). Both fine-tuning (Kolluri et al., 2025; Suh et al., 2025) and using richer data (Park et al., 2024) can significantly reduce the "sim2real" gap. Our position is thus twofold: (1) more complex methods and representations are needed than what much of the literature currently employs, and (2) there are inherent limits to what simulations can reliably capture—limits that SAGE respects by focusing on eliciting viewpoints rather than predicting chaotic phenomena.

2. **Why query the simulations at inference time? Doesn't it just add costs? Why isn't continual post-training good enough?**

Inference-time simulation offers several advantages over continual post-training. First, it provides prompt-level flexibility and guarantees. With post-training approaches like RLHF, even if preference data is gathered from a representative sample of a target population, there is not currently a principled way to guarantee, for any given prompt, whose viewpoints are reflected in the model's response. Moreover, different prompts may call for different target populations and inference-time simulation gives us control over which population to query on a per-prompt basis.

Second, inference-time simulation offers modularity: simulations can be updated continuously without risking catastrophic forgetting or degradation of other model capabilities (Luo et al., 2025), and can potentially be outsourced to dedicated services, distributing the substantial effort required to create and maintain high-fidelity simulations across specialized providers.

While inference-time procedures do increase computational cost, developing scalable algorithms for inference-time simulation is an exciting open research direction (see Section 3.2), as most uses of simulation do not have the same kinds of latency requirements. In a concurrent submission, Anonymous et al. (Kraiczy et al., 2026) take one step in this direction: given a population of n simulants, they show how to dynamically select, on a per-prompt basis, a subset of $k \ll n$ simulants to query such that their k responses provide approximate proportional representation guarantees, grounded in social choice theory, to the full population.

3. **The model's response hinges on the target population. How is the target population decided, and who makes that decision?**

This is a great question, as the choice of target population is itself a normative question that raises tricky trade-offs. For example, when should the population reflect the general public versus experts? When should the user's cultural context be prioritized over global norms? We do not claim there is a single universally correct rule; instead, the goal is to make this choice explicit and auditable.

In terms of who decides, the model deployer trains the model to call the simulations as a "tool", like any other inference-time tool. We imagine that when the model calls the simulations, it must specify both a query and a target population in natural language (e.g., query: "What is the best anime?"; population: "anime fans"). It is important that the model shows the target population to the user alongside its response, so users can see *whose* views are being represented. The model can also potentially show multiple responses, each aligned to a different population (e.g., "While experts believe [...], there remains controversy among the public around [...].")

Finally, we advocate giving users controlled agency to revise the target population and regenerate the answer. However, this also raises potential risks such as echo chambers and misuse, which need to be safeguarded against (Kirk et al., 2024a).

4. **Why simulations of individuals instead of groups? Isn't simulating individuals too difficult?**

Aligning with pre-specified groups can obscure substantial within-group heterogeneity. Simulating individuals instead provides more flexibility: at inference time, we can select and query the most relevant individuals for a given prompt, rather than relying on a fixed set of pre-defined group categories. This matters because users can ask an essentially unbounded range of questions, making it impractical to anticipate every group that might be needed. For instance, a prompt like "What is the best anime of all time?" could call for querying anime fans—a population we likely would not have explicitly defined in advance.

A common concern is that learning individual-level simulations requires prohibitive amounts of data. However, not every simulation needs to cover all domains. Some simulated individuals might only be trained on political attitudes and should be used only for political queries. The key methodological challenge, then, is determining when an individual simulation will generalize reliably to the question being asked, and when it will not.

3. Research Directions

While there has been a growing body of work on simulating human behavior with LLMs, our proposed paradigm of *simulation-augmented generation* (SAGE) imposes qualitatively different requirements. In our framework, models query simulations for open-ended responses and use those outputs to inform their own responses. This shift demands simulations that are open-domain, continuously updated, and accessible through querying mechanisms efficient enough to operate at the speed and scale of real-time inference. Meeting these requirements introduces challenges on two fronts: (1) developing the simulations themselves (Section 3.1), and (2) building the infrastructure that enables models to reliably and efficiently use simulations as inference-time tools (Section 3.2). We discuss each in turn.

3.1. Developing high-fidelity generative simulations

Realizing our vision of using simulations as inference-time tools that preserve human normative agency calls for re-

search priorities that differ from much of the current simulation literature.

First, simulations must be trained and evaluated in **generative** contexts, where they are producing open-ended natural-language responses, not merely discrete labels or survey answers. When a user asks whether they were wrong to miss a friend's wedding, the reasoning, framing, and considerations behind different viewpoints matter as much as the verdict itself; it is this rich reasoning that shapes how users interpret and situate their own values (Ye et al., 2026). By contrast, existing work on simulation often evaluates only on closed-ended tasks—alignment with fixed survey responses, replication of experimental effect sizes, and the like (Binz et al., 2025; Suh et al., 2025; Kolluri et al., 2025; Hu et al., 2026). Second, simulations should **know what they don't know**. Because users may ask models about anything—interpersonal conflicts, parenting decisions, professional ethics, political issues, cultural norms—simulations must be able to assess when they can generalize robustly. Third, simulations must be **continuously updated** to reflect the evolving values and beliefs of the populations they represent. Societal norms shift in response to world events, generational change, and public discourse; a simulation anchored to a particular historical moment risks perpetuating value lock-in (Qiu et al., 2024; 2025; Levine et al., 2026; Freedman, 2026).

Moving towards this vision requires (a) data, (b) training methods, (c) evaluations, and (d) calibration.

Data. Most simulation work relies on demographic attributes or similarly limited descriptors (Argyle et al., 2023; Hewitt et al., 2024; Sun et al., 2024; Aher et al., 2023; Bisbee et al., 2024; Qu & Wang, 2024). Despite this minimal conditioning, several studies report strong aggregate performance—for example, Hewitt et al. (2024) found r=0.85 correlation with human treatment effects across 70 experiments, and Cui et al. (2025) showed 73–81% replication of main effects from 156 psychological studies. However, these aggregate successes obscure substantial distributional failures. Bisbee et al. (2024) found that ChatGPT underestimates response variance and produces sign reversals in regression coefficients 32% of the time, while Wang et al. (2025a) showed that LLMs both misrepresent and compress demographic groups, failing to capture within-group heterogeneity. We argue these shortcomings are structural: conditioning solely on demographic labels inevitably induces stereotyping and loss of individual nuance. *Progress therefore requires grounding simulations in context-rich, individual-level data.* Park et al. (2024) demonstrate this by constructing simulants from two-hour qualitative interviews, achieving 85% of participants' own test–retest reliability while reducing bias relative to demographic-only approaches. Yet this reliance on rich data creates a practical bottleneck: such data is sensitive and difficult to share, forcing groups to collect costly, non-comparable datasets.

We advocate for building infrastructure for shared, individual-level datasets to enable cumulative progress in simulation. A promising path forward is to draw inspiration from other sensitive domains such as medicine. Researchers can obtain access to detailed individual-level medical records in large-scale datasets such as the UK Biobank and NIH's All of Us Research Program through a structured governance process: institutional agreements, identity verification, ethics training, and data use contracts that restrict re-identification and mandate secure computing environments. Creating a comparable framework for simulation data, which includes data like interview transcripts, rich surveys, and behavioral traces, could accelerate cumulative progress on simulation.

Training methods. A primary motivation for user simulators in existing work is to use them in training and evaluation of models (Yao et al., 2025; Kong et al., 2024; Suh et al., 2026). SAGE has a fundamentally different motivation that entails different requirements. First, since the goal of SAGE is to represent a population's perspectives, the most important type of fidelity is *viewpoint fidelity*: each simulant's opinions must faithfully reflect the individual being simulated. This contrasts with other applications like chat or social media, where stylistic fidelity is also important (Zhou et al., 2026; Dou et al., 2025). Viewpoint fidelity potentially requires different types of methods; Wu et al. (2026) show that directly performing supervised fine-tuning on human responses primarily captures surface-level attributes and instead use reinforcement learning to align simulations to higher-level attributes like stance and emotion. Second, since users can ask anything to the model, simulations should be trained to generalize across domains, but most existing work trains only on particular domains. Finally, SAGE requires simulations that remain temporally current. People's views evolve and out-dated simulations undermine the core goal of faithful representation. The perhaps most related application is in continual learning for personalization, where analogous challenges arise: user preferences change over time, and models must adapt without catastrophic forgetting (Kim & Kim, 2026; Yang et al., 2026; Magister et al., 2025; Freedman, 2026).

Evaluations. While existing benchmarks predominantly rely on closed-ended metrics (Binz et al., 2025; Hewitt et al., 2024; Cui et al., 2025; Hu et al., 2026), in SAGE, we must assess fidelity of open-ended simulation outputs. To capture fidelity of open-ended responses, some works use cosine similarity between embeddings of the simulator and human responses to evaluate semantic similarity (Wang et al., 2025b; Wu et al., 2026). However, SAGE demands more than semantic alignment—it requires capturing prefer-

ence: the degree to which humans like and feel represented by their simulator's output. After demonstrating that off-the-shelf embeddings perform poorly at capturing such preferences, concurrent work proposes new methods for learning "preference embeddings" (Blair et al., 2026). Another approach is to use LLM judges to assess similarity (Wu et al., 2026; Dou et al., 2025), but this introduces an analogous challenge: the judge must be calibrated to the preferences of the individual being simulated (Dong et al., 2024).

Confidence estimation. Simulations must be able to express when they do not know—whether because they are out-of-domain, lack sufficient information about a person, or because the person themselves wouldn't know. But "confidence" is conceptually murky for open-ended simulations: unlike classification, there is no single correct answer and the set of "faithful" responses can be large, making calibration-style probabilities hard to define. A single scalar ("80% confident") is also inadequate—it does not reveal whether the remaining mass comprises minor variations or fundamentally opposite views. For SAGE, uncertainty should be legible to the model synthesizing responses. One direction could be to communicate uncertainty in language itself, for example by producing textual confidence sets that describe a range of plausible responses in natural language, à la conformal prediction, which early work has begun adapting to open-ended language (Quach et al., 2024; Su et al., 2024).

3.2. Leveraging simulations as inference-time tools

The previous section addressed challenges in developing high-fidelity simulations. Here we turn to the infrastructure needed to use those simulations as inference-time tools. To ground the discussion, in Figure 1 we illustrate a sample system architecture. The **model** receives a user query and decides whether to invoke simulations and which reference population to query. The **router** selects which specific simulants within that population to query, balancing efficiency against representativeness. The selected **simulants** produce open-ended responses. Finally, the model uses the simulation's responses to synthesize a final **response**. The queried population, final response, and an assessment of its representativeness are shown to the user.

Thus far, we have only spoken about the challenges in designing the simulants themselves. Here, we discuss the distinct research challenges introduced by the other components of the system.

Querying. The first challenge is to teach the model to use simulations as inference-time tools. The model must learn *when* it is appropriate to query a simulation population, *how* to form the query (i.e., what question to ask the population), and *which* population to query.

Approaches that teach models to leverage tools through behavioral imitation (Nakano et al., 2021) could be adapted for SAGE relatively straightforwardly, since a simulation query is at a mechanical level simply another text-in, text-out API call. However, the deeper challenge would be a normative one: deciding which population is appropriate to query for a given request. A question about gift-giving norms may be best served by respondents in the user's culture; a question about medical ethics might warrant consulting both a general population and healthcare professionals. This raises key design questions: when should the model defer to experts versus the general public? When should it prioritize the user's culture or local community versus a global population? While the model will need to select defaults, we also advocate for giving users the agency to modify which population is queried, although we recognize that this also raises potential concerns (e.g., echo chambers or misuse) that would need to mitigated (Kirk et al., 2024a).

For reinforcement learning approaches that reward the model based on its final output (Chen et al., 2025a; Feng et al., 2026), an additional challenge here is defining the reward signal itself. Standard desiderata for user-facing responses—helpfulness, harmlessness, and the like—remain important, but SAGE introduces a novel requirement: ensuring that responses are *representative* of a target population. To this end, one interesting direction is to incorporate the simulations in the reward itself—for example, simulants in the relevant population could vote on how well the final answer represents their views. The final reward would then be a social-choice-based aggregation over simulant-level-feedback (Conitzer et al., 2024). A concern with this approach is reward hacking: because the simulants are themselves LLM-based, the model being trained may learn to exploit systematic biases or blind spots in the simulations rather than producing genuinely representative syntheses.

Routing. A second challenge is *routing*: given a target reference population, selecting which specific simulants to query. In practice, inference-time systems operate under strict latency and cost budgets, so only a small number of simulant calls are feasible. With a target population of n simulants, scalability—where n may be in the thousands or, one day, even millions—requires selecting only a smaller set of $k \ll n$ simulants. This creates a central tension for SAGE: routing must deliver *scalability* without sacrificing *representational validity*. The challenge is compounded by the fact that not all simulants will be able to provide confident responses to a given query—some may lack sufficient grounding information, or the query may fall outside domains where their simulation is reliable. Routing must therefore jointly optimize for representativeness and *epistemic confidence*: selecting simulants that collectively represent the target population while prioritizing those capable of producing reliable responses.

In concurrent work, Anonymous et al. (Kraiczy et al., 2026) take initial steps toward addressing this challenge. They formalize representation using concepts from computational social choice, and show that by learning a cheap embedding predictor that approximates each simulant's response embedding, it is possible to select representative simulants *before* querying them, reducing the number of computationally expensive simulation calls from n to $k \ll n$ while retaining approximate proportional representation guarantees. Future work could further improve the scalability–representation trade-off and address an open challenge: incorporating epistemic confidence to account for heterogeneity in how reliably different simulants can respond to a query.

Synthesis. Using the responses from each simulant, the model synthesizes a final response to return to the user. This synthesis step presents its own set of challenges (Shaib et al., 2023). *The synthesis must maintain representativeness*: if the routing step carefully selected simulants to reflect a target population, the synthesis should not then distort this balance by over-weighting certain voices or collapsing diverse perspectives into a false consensus. Second, *synthesis risks losing important aspects of simulants' reasoning*. A synthesis that extracts positions without their underlying rationales fails to convey the full richness of the simulated responses. This is particularly problematic for normative applications where *why* people hold a view may matter as much as what view they hold. Third, *synthesis should strive for steel-manning rather than straw-manning*: when presenting a viewpoint, the system should articulate the strongest version of the argument, not a weakened caricature. This requires the synthesizing model to understand arguments deeply enough to construct them charitably. Taken together, these challenges suggest that synthesis is not merely a summarization task but a form of *faithful deliberative compression*, requiring methods that preserve representational balance, argumentative structure, and the epistemic status of different claims.

User interface. A core promise of SAGE is transparency: users should understand not only what the model recommends, but also whose views it represents, with the ability to modify the reference population if desired. This requires an effective user interface for SAGE, which remains an open research problem. The bottom panel of Figure 1 illustrates one possible extension to the typical LLM chat UI. The user is shown which population the model is querying and is given an option to modify it. Different segments of the response are highlighted and surface how common each view is within that population, helping users distinguish popular and unpopular opinions, and aiming to reduce the risk of *false balance* (Sorensen et al., 2024a). The interface also reports overall coverage of the synthesized response (e.g., "88% of views represented"[1]) and offers an option to inspect views from the population that were not included. This lets the model provide concise summaries while still enabling users to explore the full range of perspectives. This mock-up is meant only as an illustration of the possibilities, but dedicated HCI research is needed to determine which designs most effectively help users contextualize the guidance they receive. Recently, Ye et al. (2026) propose an interface that is a more radical departure from the standard chat UI: it allows users to traverse a "conceptual multiverse" of possible responses by controlling key decision points in how the model reasons about the prompt.

4. Conclusion

As users increasingly turn to AI models for normative assistance—deciding what they ought to do or think—a key question is how to maintain user agency. Current approaches risk distorting the normative landscape in opaque and uncontestable ways. As an alternative, we proposed *simulation-augmented generation (SAGE)*, a novel paradigm in which models query generative simulations of individuals from prompt-appropriate reference populations at inference time. SAGE makes normative assistance transparent and auditable: users can see which population was consulted (and optionally, change it) and how representative the response is, enabling them to situate the guidance they receive within the broader landscape of human thought.

We have defended SAGE primarily on grounds of procedural legitimacy: by giving users the ability to access a representative landscape of societal viewpoints, the decisions they reach can be more credibly attributed to their own reflective judgment rather than subtle model steering. However, we also speculate that SAGE may yield epistemic benefits—paralleling the way that deliberative democracy is increasingly defended on epistemic grounds in addition to procedural ones (Landemore, 2017). By helping users access the "wisdom of the crowds", SAGE could serve as a mechanism for improving collective intelligence through AI, as advocated by Burton et al. (2024).

While we have focused throughout on simulating the current opinions of individuals, a limitation of this framing is that most people do not have the time or bandwidth to form considered opinions on most topics. Deliberative democracy addresses this challenge by creating structured contexts in which citizens are given time to think deeply, engage with diverse perspectives, and arrive at more reflective judgments (Fishkin, 2018). One could imagine extending SAGE to represent not merely the opinions people currently hold,

[1] Such population-level estimates of representativeness could be made more reliable through rectification methods (Angelopoulos et al., 2023); Krsteski et al. (2025) demonstrate that such techniques can substantially reduce bias in simulation-based estimates

but to predict the considered opinions they would hold, e.g., after going through a deliberative process—a form of simulated deliberative polling (Hewitt et al., 2026; Leike, 2023). This is an intriguing direction for future work, but poses further validity challenges.

Overall, by acting as a transparent conduit to societal perspectives, rather than an opaque arbiter of judgment, SAGE offers a path toward AI that genuinely *assists*, rather than *steers*, users in normative choices.

References

Abdulhai, M., White, I., Wan, Y., Qureshi, I., Leibo, J., Kleiman-Weiner, M., and Jaques, N. How LLMs Distort Our Written Language. *arXiv preprint arXiv:2603.18161*, 2026.

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.

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000. URL https://www.science.org/doi/abs/10.1126/science.adi6000.

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.

Bächtiger, A., Dryzek, J. S., Mansbridge, J., and Warren, M. E. *The Oxford Handbook of Deliberative Democracy*, chapter Deliberative Democracy: an introduction. Oxford University Press, 2018.

Baumann, J., Pei, J., Koyejo, S., and Hovy, D. Stop Automating Peer Review Without Rigorous Evaluation. In *Forty-third International Conference on Machine Learning*, 2026. Spotlight.

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., et al. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, 2025.

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., and Larson, J. M. Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Analysis*, 32(4):401–416, 2024.

Blair, C., Procaccia, A. D., and Tambe, M. Embeddings for Preferences, Not Semantics. 2026.

Burton, J. W., Lopez-Lopez, E., Hechtlinger, S., Rahwan, Z., Aeschbach, S., Bakker, M. A., Becker, J. A., Berditchevskaia, A., Berger, J., Brinkmann, L., Flek, L., Herzog, S. M., Huang, S., Kapoor, S., Narayanan, A., Nussberger, A.-M., Yasseri, T., Nickl, P., Almaatouq, A., Hahn, U., Kurvers, R. H. J. M., Leavy, S., Rahwan, I., Siddarth, D., Siu, A., Woolley, A. W., Wulff, D. U., and Hertwig, R. How large language models can reshape collective intelligence. *Nature Human Behaviour*, 8(9):1643–1655, 2024. doi: 10.1038/s41562-024-01959-9. URL https://doi.org/10.1038/s41562-024-01959-9.

Candès, E. and Recht, B. Exact matrix completion via convex optimization. *Commun. ACM*, 55(6):111–119, June 2012. ISSN 0001-0782. doi: 10.1145/2184319.2184343. URL https://doi.org/10.1145/2184319.2184343.

Carroll, M., Chan, A., Ashton, H., and Krueger, D. Characterizing Manipulation from AI Systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400703812. doi: 10.1145/3617694.3623226. URL https://doi.org/10.1145/3617694.3623226.

Carroll, M., Foote, D., Siththaranjan, A., Russell, S., and Dragan, A. AI Alignment with Changing and Influenceable Reward Functions. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Chen, M., Sun, L., Li, T., sunhaoze, ZhouYijie, Zhu, C., Wang, H., Pan, J. Z., Zhang, W., Chen, H., Yang, F., Zhou, Z., and Chen, W. ReSearch: Learning to Reason with Search for LLMs via Reinforcement Learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=OuGAwwAT8G.

Chen, M., Wang, X., Chen, K., and Koudas, N. Generative Engine Optimization: How to Dominate AI Search. *arXiv preprint arXiv:2509.08919*, 2025b.

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., and Jurafsky, D. Sycophantic ai decreases prosocial intentions and promotes dependence. *Science*, 391(6792):eaec8352, 2026a. doi: 10.1126/science.aec8352. URL https://www.science.org/doi/abs/10.1126/science.aec8352.

Cheng, M., Yu, S., Lee, C., Khadpe, P., Ibrahim, L., and Jurafsky, D. ELEPHANT: Measuring and understanding social sycophancy in LLMs. In *The Fourteenth International Conference on Learning Representations*,

2026b. URL https://openreview.net/forum?id=igbRHKEiAs.

Conitzer, V., Freedman, R., Heitzig, J., Holliday, W. H., Jacobs, B. M., Lambert, N., Mosse, M., Pacuit, E., Russell, S., Schoelkopf, H., Tewolde, E., and Zwicker, W. S. Position: Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9346–9360. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/conitzer24a.html.

Cui, Z., Li, N., and Zhou, H. A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science*, 5(8):627–634, 2025. doi: 10.1038/s43588-025-00840-7. URL https://doi.org/10.1038/s43588-025-00840-7.

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?, 2025. URL https://arxiv.org/abs/2510.05444.

Edelman, J., Zhi-Xuan, T., Lowe, R., Klingefjord, O., Wang-Mascianica, V., Franklin, M., Kearns, R. O., Hain, E., Sarkar, A., Bakker, M., Barez, F., Duvenaud, D., Foerster, J., Gabriel, I., Gubbels, J., Goodman, B., Haupt, A., Heitzig, J., Jara-Ettinger, J., Kasirzadeh, A., Kirkpatrick, J. R., Koh, A., Knox, W. B., Koralus, P., Lehman, J., Levine, S., Marro, S., Revel, M., Shorin, T., Sutherland, M., Tessler, M. H., Vendrov, I., and Wilken-Smith, J. Full-stack alignment: Co-aligning ai and institutions with thick models of value, 2025. URL https://arxiv.org/abs/2512.03399.

Feng, J., Huang, S., Qu, X., Zhang, G., Qin, Y., Zhong, B., Jiang, C., Chi, J., and Zhong, W. ReTool: Reinforcement Learning for Strategic Tool Use in LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=tRk1nofSmz.

Feng, S., Sorensen, T., Liu, Y., Fisher, J., Park, C. Y., Choi, Y., and Tsvetkov, Y. Modular Pluralism: Pluralistic Alignment via Multi-LLM Collaboration. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4151–4171, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.240. URL https://aclanthology.org/2024.emnlp-main.240/.

Fisher, J., Appel, R. E., Park, C. Y., Potter, Y., Jiang, L., Sorensen, T., Feng, S., Tsvetkov, Y., Roberts, M., Pan, J., Song, D., and Choi, Y. Position: Political neutrality in AI is impossible — but here is how to approximate it. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=H72JEXAPwo.

Fishkin, J. S. *Democracy When the People Are Thinking: Revitalizing Our Politics Through Public Deliberation*. Oxford University Press, 2018.

Franklin, M., Tomašev, N., Jacobs, J., Leibo, J. Z., and Osindero, S. AI Agent Traps. 2026.

Freedman, R. Adaptive Pluralistic Alignment: A pipeline for dynamic artificial democracy, 2026. URL https://arxiv.org/abs/2605.01642.

Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., and Summerfield, C. The levers of political persuasion with conversational artificial intelligence. *Science*, 390(6777):eaea3884, 2025. doi: 10.1126/science.aea3884. URL https://www.science.org/doi/abs/10.1126/science.aea3884.

Hewitt, L., Ashokkumar, A., Ghezae, I., and Willer, R. Predicting Results of Social Science Experiments using Large Language Models. *Preprint*, 2024.

Hewitt, L., Dale, M. K., and de Font-Reaulx, P. DeliberationBench: A Normative Benchmark for the Influence of Large Language Models on Users' Views, 2026. URL https://arxiv.org/abs/2603.10018.

Hu, T., Baumann, J., Lupo, L., Collier, N., Hovy, D., and Röttger, P. SimBench: Benchmarking the Ability of Large Language Models to Simulate Human Behaviors. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=PL51SpN6ZJ.

Kim, S. and Kim, J. Spring: Continual llm personalization via selective parametric adaptation and retrieval-interpolated generation. *arXiv preprint arXiv:2601.09974*, 2026.

Kirk, H. R., Vidgen, B., Röttger, P., and Hale, S. A. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6(4):383–392, 2024a.

Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., and Hale, S. A. The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 105236–105344. Curran Associates, Inc., 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/be2e1b68b44f2419e19f6c35a1b8cf35-Paper-Datasets_and_Benchmarks_Track.pdf.

Király, F. J., Theran, L., and Tomioka, R. The algebraic combinatorial approach for low-rank matrix completion. 16:1391–1436, 2015. ISSN 1532-4435.

Kolluri, A., Wu, S., Park, J. S., and Bernstein, M. S. Finetuning LLMs for Human Behavior Prediction in Social Science Experiments. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30096–30111, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1530. URL https://aclanthology.org/2025.emnlp-main.1530/.

Kong, C., Fan, Y., Wan, X., Jiang, F., and Wang, B. PlatoLM: Teaching LLMs in Multi-Round Dialogue via a User Simulator. 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. 7841–7863, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.424. URL https://aclanthology.org/2024.acl-long.424/.

Kraiczy, S., Milli, S., Berker, R. E., Bose, A., Amos, B., Watson-Daniels, J., Nickel, M., Elkind, E., and Procaccia, A. D. Social Choice Foundations for Simulation-Augmented Generation. 2026.

Krsteski, S., Russo, G., Chang, S., West, R., and Gligorić, K. Valid Survey Simulations with Limited Human Data: The Roles of Prompting, Fine-Tuning, and Rectification, 2025. URL https://arxiv.org/abs/2510.11408.

Landemore, H. Beyond the fact of disagreement? the epistemic turn in deliberative democracy. *Social Epistemology*, 31(3):277–295, 2017. doi: 10.1080/02691728.2017.1317868. URL https://doi.org/10.1080/02691728.2017.1317868.

Leike, J. A proposal for importing society's values. https://aligned.substack.com/p/a-proposal-for-importing-societys-values, March 2023. Accessed: 2026-05-05.

Levine, N., Duvenaud, D., and Radford, A. Introducing talkie: a 13B vintage language model from 1930. April 2026. URL https://talkie-lm.com/introducing-talkie.

Luettgau, L., Cheung, V., Dubois, M., Juechems, K., Bergs, J., Symes, L., Davidson, H., O'Dell, B., Kirk, H. R., Rollwage, M., and Summerfield, C. People readily follow personal advice from ai but it does not improve their well-being, 2026. URL https://arxiv.org/abs/2511.15352.

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., and Zhang, Y. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3776–3786, 2025. doi: 10.1109/TASLPRO.2025.3606231.

Magister, L. C., Metcalf, K., Zhang, Y., and Ter Hoeve, M. On the Way to LLM Personalization: Learning to Remember User Conversations. In Jia, R., Wallace, E., Huang, Y., Pimentel, T., Maini, P., Dankers, V., Wei, J., and Lesci, P. (eds.), *Proceedings of the First Workshop on Large Language Model Memorization (L2M2)*, pp. 61–77, Vienna, Austria, August 2025. Association for Computational Linguistics. ISBN 979-8-89176-278-7. doi: 10.18653/v1/2025.l2m2-1.5. URL https://aclanthology.org/2025.l2m2-1.5/.

Meka, R., Jain, P., and Dhillon, I. Matrix completion from power-law distributed samples. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL https://papers.nips.cc/paper/2009/hash/1fc214004c9481e4c8073e85323bfd4b-Abstract.html.

Miyazaki, S. and Hall, A. B. Why Do AI Models Tell Left-Wing Voters to Support the Communist Party? AI Voting Advice in Japan's 2026 General Election. 2026.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Nickel, M. No free delivery service: Epistemic limits of passive data collection in complex social systems. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 37*, volume 37, pp. 102279–102306. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-3248. URL http://dx.doi.org/10.52202/079017-3248.

Ovadya, A. 'Generative CI' through Collective Response Systems, 2023. URL https://arxiv.org/abs/2302.00672.

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., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., and Bernstein, M. S. Generative Agent Simulations of 1,000 People. *arXiv preprint arXiv:2411.10109*, 2024.

Pimentel-Alarcón, D. L., Boston, N., and Nowak, R. D. A characterization of deterministic sampling patterns for low-rank matrix completion. 10:623–636, 2016. ISSN 1932-4553.

Qiu, T., Zhang, Y., Huang, X., Li, J. X., Ji, J., and Yang, Y. Progressgym: Alignment with a millennium of moral progress. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=0mRouJElbZ.

Qiu, T., He, Z., Chugh, T., and Kleiman-Weiner, M. The lock-in hypothesis: Stagnation by algorithm. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=mE1M626qOo.

Qu, Y. and Wang, J. Performance and biases of Large Language Models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1–13, 2024.

Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T. S., and Barzilay, R. Conformal Language Modeling. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pzUhfQ74c5.

Shaib, C., Li, M., Joseph, S., Marshall, I., Li, J. J., and Wallace, B. Summarizing, simplifying, and synthesizing medical evidence using GPT-3 (with varying success). In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1387–1407, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.119. URL https://aclanthology.org/2023.acl-short.119/.

Sharma, M., McCain, M., Douglas, R., and Duvenaud, D. Who's in Charge? Disempowerment Patterns in Real-World LLM Usage. *arXiv preprint arXiv:2601.19062*, 2026.

Shen, J. H., Carter, S., Dargan, R., Gillotte, J., Handa, K., Hong, J., Huang, S., Jagadish, K., Kearney, M., Levinstein, B., Linthicum, R., McCain, M., Millar, T., Julapalli, M., Price, S., Stern, M., Saunders, D., Tamkin, A., Vallone, A., Clark, J., Pollack, S., Eaton, J., Ganguli, D., and Durmus, E. How People Ask Claude for Personal Guidance. https://www.anthropic.com/research/claude-personal-guidance, 2025. Anthropic Research Blog.

Sorensen, T., Jiang, L., Hwang, J. D., Levine, S., Pyatkin, V., West, P., Dziri, N., Lu, X., Rao, K., Bhagavatula, C., Sap, M., Tasioulas, J., and Choi, Y. Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024a. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i18.29970. URL https://doi.org/10.1609/aaai.v38i18.29970.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., and Choi, Y. Position: A Roadmap to Pluralistic Alignment. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024b.

Su, J., Luo, J., Wang, H., and Cheng, L. API Is Enough: Conformal Prediction for Large Language Models Without Logit-Access. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 979–995, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.54. URL https://aclanthology.org/2024.findings-emnlp.54/.

Suh, J., Jahanparast, E., Moon, S., Kang, M., and Chang, S. Language Model Fine-Tuning on Scaled Survey Data for Predicting Distributions of Public Opinions. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21147–21170, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1028. URL https://aclanthology.org/2025.acl-long.1028/.

Suh, J., Raj, A., Kang, M., and Chang, S. Quantifying the Utility of User Simulators for Building Collaborative LLM Assistants, 2026. URL https://arxiv.org/abs/2605.09808.

Sun, S., Lee, E., Nan, D., Zhao, X., Lee, W., Jansen, B. J., and Kim, J. H. Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information. *arXiv preprint arXiv:2402.18144*, 2024.

Waight, H., Yang, E., Yuan, Y., Messing, S., Roberts, M. E., Stewart, B. M., and Tucker, J. A. State media control influences large language models. *Nature*, 2026. doi: 10.1038/s41586-026-10506-7. URL https://doi.org/10.1038/s41586-026-10506-7.

Wang, A., Morgenstern, J., and Dickerson, J. P. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, pp. 1–12, 2025a.

Wang, K., Li, X., Yang, S., Zhou, L., Jiang, F., and Li, H. Know You First and Be You Better: Modeling Human-Like User Simulators via Implicit Profiles. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21082–21107, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1025. URL https://aclanthology.org/2025.acl-long.1025/.

Williams, M., Carroll, M., Narang, A., Weisser, C., Murphy, B., and Dragan, A. On Targeted Manipulation and Deception when Optimizing LLMs for User Feedback. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Wf2ndb8nhf.

Williams-Ceci, S., Jakesch, M., Bhat, A., Kadoma, K., Zalmanson, L., and Naaman, M. Biased ai writing assistants shift users' attitudes on societal issues. *Science Advances*, 12(11):eadw5578, 2026. doi: 10.1126/sciadv.adw5578. URL https://www.science.org/doi/abs/10.1126/sciadv.adw5578.

Wu, S., Choi, E., Khatua, A., Wang, Z., He-Yueya, J., Weerasooriya, T. C., Wei, W., Yang, D., Leskovec, J., and Zou, J. Humanlm: Simulating users with state alignment beats response imitation. 2026. URL https://humanlm.stanford.edu/.

Yang, B., Xu, L., Li, Y., Liu, K., Jiang, X., and Yan, Z. Sensorpersona: An llm-empowered system for continual persona extraction from longitudinal mobile sensor streams. *arXiv preprint arXiv:2604.06204*, 2026.

Yao, S., Shinn, N., Razavi, P., and Narasimhan, K. R. {τ}-bench: A benchmark for \underline{T}ool-\underline{A}gent-\underline{U}ser interaction in real-world domains. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=roNSXZpUDN.

Ye, A., Huang, J. Y., Guo, A., Novick, R., Broderick, T., and Gordon, M. L. Navigating the Conceptual Multiverse. *arXiv preprint arXiv:2604.17815*, 2026.

Zhang, L. H., Milli, S., Jusko, K., Smith, J., Amos, B., Revel, M., Kussman, J., Titus, L., Radharapu, B., Yu, J., et al. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. *arXiv preprint arXiv:2507.09650*, 2025.

Zhou, X., Zhu, H., Mathur, L., Zhang, R., Yu, H., Qi, Z., Morency, L.-P., Bisk, Y., Fried, D., Neubig, G., and Sap, M. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mM7VurbA4r.

Zhou, X., Sun, W., Ma, Q., Xie, Y., Liu, J., Du, W., Welleck, S., Yang, Y., Neubig, G., Wu, S. T., and Sap, M. Mind the Sim2Real Gap in User Simulation for Agentic Tasks, 2026. URL https://arxiv.org/abs/2603.11245.