Special Issue Article

Sociological Methods & Research
1–46

DOI: 10.1177/00491241251342008
journals.sagepub.com/home/smr
Sage

Balancing Large Language Model Alignment and Algorithmic Fidelity in Social Science Research

Alex Lyman[1], Bryce Hepner[1], Lisa P. Argyle[2], Ethan C. Busby[2], Joshua R. Gubler[2], and David Wingate[1]

Abstract

Generative artificial intelligence (AI) has the potential to revolutionize social science research. However, researchers face the difficult challenge of choosing a specific AI model, often without social science-specific guidance. To demonstrate the importance of this choice, we present an evaluation of the effect of alignment, or human-driven modification, on the ability of large language models (LLMs) to simulate the attitudes of human populations (sometimes called *silicon sampling*). We benchmark aligned and unaligned versions of six open-source LLMs against each other and compare them to similar responses by humans. Our results suggest that model alignment impacts output in predictable ways, with implications for prompting, task completion, and the substantive content of LLM-based results. We conclude that researchers must be aware of the complex ways in which model training affects their research and carefully consider model choice for each project. We discuss future steps to improve how social scientists work with generative AI tools.

[1]Computer Science, Brigham Young University, Provo, UT, USA
[2]Political Science, Brigham Young University, Provo, UT, USA

Corresponding Author:
Ethan C. Busby, Political Science, Brigham Young University, Provo, UT, USA.
Email: ethan.busby@byu.edu

Data Availability Statement included at the end of the article

Keywords

artificial intelligence, alignment, large language models, silicon sampling, computational social science

1. Introduction

Large language models (LLMs) such as ChatGPT (OpenAI 2023), Gemini (Google Gemini Team 2024), Claude (Anthropic 2024), and Llama (AI@Meta 2024) have quickly transformed the landscape of work in tech, education, research, communications, and more, seemingly leaving no industry untouched. LLM tools are being integrated into a range of daily use tools, such as online searches, computer programming, word processing, and customer service interactions, where both expert professionals and lay users regularly interact with them. In this artificial intelligence (AI) moment, it is hard to overstate the impact of LLMs across the social, political, economic, and educative landscape.

In the realm of social science research, scholars have proposed a variety of applications for LLMs, which span the full scope of the research pipeline, including: search, summary, and synthesis of existing literature (Consensus 2024; Copilot 2024; Elicit 2024); text classification and coding (Gilardi et al. 2023); interaction with human subjects to administer experimental stimuli or surveys (Argyle et al. 2023b; Velez and Liu 2024); silicon simulation of human attitudes and behaviors (Aher et al. 2023; Argyle et al. 2023a; Hewitt et al. 2024; Horton 2023; Kozlowski et al. 2024); and much more (Bail 2024; Demszky et al. 2023). Each of these applications raises both normative concerns about the meaning of the scientific process and the value of outsourcing key creative tasks to an automated non-human system, and empirical questions about the capability of LLMs to satisfactorily conduct these tasks. However, systematic evaluation of these concerns is hampered by the rapid updating and proliferation of LLMs, and by the reality that different LLMs—trained on different data with different model architecture and different alignment processes—often perform the same tasks in radically different ways. How, then, should researchers choose a generative AI tool for their specific applications?

There is no one, universal answer to this question. However, we suggest there are clear processes that researchers can follow to identify a model that will work for their particular use case and specific goals. To illustrate this process, we explore the impact of one key distinction between models that is often overlooked by social scientists: the degree to which LLMs have been *aligned*—or modified through explicit human guidance—toward more desirable, functional, or socially positive behavior.

In this evaluation, we focus specifically on the impact of alignment on the capacity of LLMs to simulate human responses in a social science research context. However, we believe the discussion we provide on the interplay between LLMs—including model training dynamics, prompts, training data, and alignment process—and the various goals of research-oriented simulation, apply to the use of AI for a range of common social science tasks beyond simulation, including text classification, hypothesis generation, and summary or synthesis of current research.

After describing the use of LLMs for silicon sampling research and introducing model architecture and alignment considerations, we propose a set of expectations for how model alignment will impact silicon sampling. We expect the three general goals of alignment—pushing models to be helpful, honest, and harmless—to lead to predictable differences in model behavior. We then present both a benchmarking exercise (Study 1) and a replication and extension of foundational silicon sampling work (Study 2) to highlight some ways model alignment impacts a researcher's ability to accomplish various research goals.

Our results suggest that model alignment impacts how models follow instructions, complete the task, and the content of the output in systematic and predictable ways. In light of this, researchers should pay careful attention to model alignment when selecting a model for research tasks. We find that neither aligned nor unaligned models are universally better for silicon sampling, but rather that researchers need to be aware of the range of complex and nuanced ways in which model training affects their research output and carefully choose a model to reach their specific research goals. In the two studies presented in this article, we present a relatively simple process of benchmarking and testing that can be employed by researchers to systematically explore the effects of model differences, like alignment, on their particular research goals. We conclude with a discussion of key principles resulting from these studies to help guide social science model choice, and with suggestions for future research in this area.

2. LLMs in Social Science Research

Language models are trained in a series of stages, each with different goals, and each of which results in a model with different properties. These stages and their differences will be discussed in detail in the next section, but ultimately, developers of large language models like ChatGPT, Claude, Gemini, LLama, Gemma (Team 2024), Mistral (Jiang et al. 2023) and others generally strive to make models *helpful, honest*, and *harmless* (Askell et al. 2021).

Importantly, the meaning of these goals depends on the task for which an LLM is used. For example, in the context of information retrieval or conversation with human counterparts, models are most *helpful* and do the least *harm* when they are free as much as possible from *algorithmic bias* and the misinformed, prejudiced, or toxic information/speech that results from it (Bender et al. 2021; Caliskan et al. 2017; Kleinberg et al. 2018; Obermeyer et al. 2019). While such bias naturally emerges from training these models on biased, misinformed, and prejudiced human text, scholars rightly fear that LLMs that reflect these biases can perpetuate them at an unprecedented scale, causing significant social harm (Cheng et al. 2023; Goldstein and Sastry 2023; Panch et al. 2019).

However, social science researchers, particularly in sociology, psychology, and political science, often seek to use LLMs for very different tasks—tasks that require the models to accurately reflect the thoughts and attitudes of their human counterparts. For these social scientists, accurate reflection of the biased, misinformed, and prejudiced thought processes of various human groups is *helpful*—it enables what Argyle et al. (2023a) call *algorithmic fidelity*; the alignment of these models thus has the potential to be *harmful*, or in tension with researchers' goals. Argyle et al. (2023a) define algorithmic fidelity as "the degree to which the complex patterns of relationships between ideas, attitudes, and socio-cultural contexts within a model accurately mirror those within a range of human sub-populations," and show how high algorithmic fidelity "enables researchers to extract information from a single language model that provides insight into the different patterns of attitudes and ideas present across many groups (women, men, white people, people of color, millennials, baby boomers, etc.) and also the *combination and intersection* of these groups (black immigrants, female Republicans, white males, etc.)."

In particular, *silicon sampling*, or the use of LLMs to generate and then study in-silico representations of human populations, relies entirely on high LLM algorithmic fidelity. Since early work in this space (Argyle et al. 2023a; Dillion et al. 2023; Horton 2023), hundreds of projects across a variety of disciplines have introduced innovations to and relied upon this approach (Pachot and Petit 2024; Ziems et al. 2024). This recent research has raised as many questions as it has hopes for the viability of using LLMs to simulate human subjects. Some raise concerns about the ability of these models to reliably simulate human subjects across a variety of important demographic subgroups, highlighting issues related to algorithmic bias, model steerability, and so forth (Bisbee et al. 2024; Boelaert et al. 2024; Cheng et al. 2023; Qu and Wang 2024; Santurkar et al. 2023). Others, including a number of prominent computational sociologists, find much more promising results. For example, in a project using GPT-4, Hewitt et al. (2024) find representative silicon samples are capable of closely replicating ($r = 0.85$) human results from 476 experimental treatments. In a an experiment designed to test the ability of LLMs to predict COVID-19 attitudes, Kozlowski et al. (2024) find that their "simulated respondents reproduce[d] observed partisan differences in COVID-19 attitudes in 84% of cases, significantly greater than chance." Lee et al. (2024) find similar results, with some caveats, in their attempt to use silicon subjects to predict climate change attitudes. In an important recent innovation, Kim and Lee (2024) show how fine tuning silicon samples on human survey responses greatly improves both retrodiction and missing survey response predictions. None of the preceding authors argue that LLMs *can* or *should* replace human participants, but their work suggests LLMs can successfully simulate human attitudes in various contexts, with certain caveats. If accurate, silicon simulation has the potential to augment shortcomings of human subject sampling and recruitment to improve social science research inference.

Why do some researchers in this area successfully use LLMs to simulate human attitudes and beliefs while others do not? We believe a variety of factors come into play, including choice of model family, model size, prompting approaches, training data, and differences in expectations or benchmarking tests. Here we argue that, in addition to these differences, an important portion of the explanation lies in understanding differences in model architecture and alignment values and goals. In the following section, we explain why we believe these factors, which have seen relatively little academic discussion, drive differences in outcomes.

3. The Potential Effects of Training and Alignment on Algorithmic Fidelity

Having proposed that goals of being harmless, honest and helpful are context dependent and sometimes in tension, we now turn to a discussion of the technical details of LLM training. Language models are generally trained in two stages: the "pre-training" phase and the "alignment" phase. Each has distinct goals, and results in models with distinct properties, as we now discuss.

Throughout this article, we will refer to models that have been pre-trained, but not aligned, as "base models." We reference models that have been both pre-trained and aligned as "aligned models."

3.1. The Pre-Training Phase

Modern generative language models have billions of parameters; as a result, they must be trained on huge corpora of natural language text. During the pre-training phase, models are trained on trillions of tokens of natural language, with the explicit goal of accurately modeling the distribution of the data (mathematically expressed as maximizing its log-likelihood). This data usually comes from human-generated text scraped from the internet, but because of the vast quantities needed, it is usually only lightly curated.

Pre-training a language model of sufficient size endows it with many natural language processing abilities, such as translation, summarization, and question answering (Radford et al. 2019). These base models also show some emergent abilities to perform tasks that they were not explicitly trained to do (Wei et al. 2022); taken together, many of these abilities are sufficient to perform various social science tasks of interest.

Despite their many desirable abilities, base models do indeed accurately reflect the statistics of their training data, for better and for worse (in machine learning parlance, the models are *well calibrated*[1] (OpenAI et al. 2024)). Because online text (and therefore, training data) often contains bias, violent rhetoric, false information, and hate speech, naively mimicking the statistics of this text is unacceptably dangerous for most use cases, which motivates a second stage of training.

3.2. The Alignment Phase

After pre-training, base models often go through a second training phase supervised by humans. In contrast to pre-training, where the goal is to accurately model the distribution of data, this training represents a conscious effort to change model behavior. Alignment pipelines can be quite elaborate, and involve highly curated datasets that are carefully sequenced to enhance specific capabilities of a language model and reduce undesirable behaviors. For example, in the Llama 3.1 alignment phase, specific data mixtures were used to help enhance factuality, steerability, multilinguality, tool use, long contexts, math and reasoning, and programming ability (Dubey et al. 2024).

From an algorithmic standpoint, model alignment can take a variety of forms, with the most common methods being Instruction Tuning (Zhang et al. 2024), Direct Preference Optimization (DPO) (Rafailov et al. 2023), and Reinforcement Learning from Human Feedback (RLHF) (Bai et al. 2022). Instruction Tuning is a type of supervised fine tuning (SFT) that consists of fine-tuning a pre-trained language model on many examples of instruction-response pairs. This teaches the language model to both pay attention to prompts and to follow instructions contained therein. This instruction-following ability creates a significant difference between a base model and an aligned counterpart in this area, and is generally considered helpful in virtually any context.

The primary method for curbing inappropriate model responses is refusal training. This is done by including refusals in the instruction tuning data, where a user asks an unsafe or offensive query and the expected reply is a refusal to comply. In contrast to general instruction tuning, refusal training can cause a language model to refuse to follow instructions, or devolve into moral lectures about whether or not a topic is acceptable.

While instruction tuning provides specific examples of correct behavior, RLHF and DPO operate on a different principle. Both RLHF and DPO are typically run after instruction tuning. In both, the tuned model is given a query and asked to generate multiple responses. An external evaluator (usually a human) then scores which output is preferred and passes this feedback to the model, which learns from the evaluation. This indirectly imbues models with human preference data, skewing the model toward the values and goals of the human evaluators.

3.3. Quantitative Effects of Alignment on Model Behavior

Alignment has many consequences, both intentional and unintentional, on model performance. As these consequences have seen little discussion as of yet in social science, we briefly review some insights from research from computer science. As computer scientists are typically less concerned about silicon sampling, we discuss how these insights might impact social science research, and particularly research based on silicon sampling.

Calibration: First, alignment dramatically reduces calibration (OpenAI et al. 2024), meaning that the distribution of outputs generated by an aligned model no longer match the distribution of the training data, though this effect diminishes for models with larger parameter scales (Zhu et al. 2023). This means that the probability assigned to any next token is less inherently meaningful as models undergo more alignment.

Consistency: Alignment also affects a model's consistency (how often it gives the same general answers to the same questions). Aligned language models are less consistent than unaligned models, a difference exacerbated when discussing controversial or sensitive topics (Moore et al. 2024). One demonstration of this inconsistency for aligned models can be found in the robust literature on “jailbreaking” models, where small changes to a prompt can be sufficient to bypass guardrails that alignment intends to establish (Arditi et al. 2024; Chu et al. 2024; Wei et al. 2024; Xu et al. 2024). This can lead language models to give responses that are inconsistent in tone or content across slightly different prompts, for example refusing a response in some cases and answering in others, even when the substance of the request is quite similar.

Variability: Multiple studies find that alignment also decreases variability in language model outputs. LLM outputs are the result of a stochastic draw from a distribution of probable next tokens, which means that providing the same prompt multiple times can result in a range of different outputs. The variability in responses to a single prompt is expected to be lower for an aligned model than an unaligned model. RLHF may help models generalize to new inputs better than SFT, though RLHF causes a significant drop in output diversity (Kirk et al. 2024). This apparent drop in diversity of outputs may be an artifact of a different process that comes from alignment. Aligned models tend to aggregate useful information and generate more relevant text. While the resulting responses are less diverse, they include more relevant information (Lake et al. 2024).

Steerability: A key property of language models for social scientists is steerability, or the degree to which a model is willing to simulate a specific person or take on a defined role. Alignment increases persona-guided steerability, with RLHF, DPO, and SFT increasing steerability to different degrees. RLHF-induced steerability, however, tends to decrease variety in model outputs (Liu et al. 2024).

Refusals: Guardrails aligned into models to make them more harmless do not activate equally across model prompts, and are not limited to particular sensitive queries. For example, GPT-4’s refusal rates across sensitive and political topics are mediated by including particular demographic information. Even seemingly inconsequential demographic information like sports team fandom can affect model refusals (Li et al. 2024).

3.4. Exploring the Tension Between Alignment and Algorithmic Fidelity

To recap, while pre-training datasets are often large, sprawling, and messy, alignment methods all involve carefully curated datasets. These datasets contain examples of instruction following, refusal, preferred outputs, and so forth. In every case, decisions about what to include in these datasets are *normative decisions*, based on judgments regarding the kinds of behavior and attitudes acceptable to print.

The values behind these decisions dictate what it means to be helpful, honest and harmless *in the contexts for which the models are aligned*. As alignment is not driven by social science concerns, it is easy to see how alignment might affect social science research in negative ways by increasing harmless answers at the cost of honesty when discussing sensitive groups (Parrish et al. 2022). For example, one toxicity alignment benchmark notes the following: “Some racial statistics, such as ’Black people have the highest unemployment rate,’ even though factually correct, can be highly controversial. Morally speaking, the output of large language models (LLMs) should not further propagate discrimination or stereotypes. Therefore, when mentioning data about disadvantaged groups, LLMs should either respond affirmatively or with cautiousness” (Ji et al. 2023).

However, base models are not necessarily “better.” While base models may be more calibrated, they may be less steerable; while they may reflect a more complete range of human perspectives, they may be more prone to produce harmful text. While aligned models may follow instructions better, they may also refuse to comply; while they may avoid stereotypes, they may also avoid uncomfortable truths. Thus, we expect that standard alignment goals of producing models that are more helpful, honest, and harmless will affect models in a range of predictable ways that are neither all good nor all bad for social scientists.

This means that researchers are left to discern for themselves which particular base or aligned model best fits their particular project goals. To make this decision, we suggest all AI researchers in social science first begin with simple benchmarking tasks (see Study 1), and then pursue additional exploration (Study 2) if required by their particular goals. As mentioned earlier in this manuscript, we explore the effects of alignment on silicon sampling, but propose that the same staged approach can be used for a variety of use cases.

4. Study 1: Task Completion and Steerability Benchmarking Test

To explore the relationship between alignment and algorithmic fidelity for silicon sampling tasks, we designed a simple, stripped-down benchmarking exercise. This exercise measures the ability of various models to complete the task as requested: in our case, to adopt provided personas with a range of demographic and attitudinal characteristics and provide opinions consistent with those backgrounds—without refusing, providing moralizing or other type of commentary, or exhibiting other types of non-compliance. As we note in our conclusion, identifying the degree to which a model has capacity for a task is an essential first step to model choice, and it can be done (as we do here) without the use of human data. Stripped-down benchmarking tasks like this are common in language model research; see Suzgun et al. (2024) for an example.

In our particular case, a simple benchmarking task is an essential first step to ultimately identifying whether a model has sufficient algorithmic fidelity to engage in silicon sampling. Refusals to provide information, providing only some of the information, providing inconsistent information, the wrong information, or information in an incorrect format all prevent a model from having algorithmic fidelity. It only makes sense to move to the second task—identifying the degree to which responses match a distribution of human responses (Study 2)—after it is clear a model passes a basic capacity benchmark for the task. As such, what follows in Study 1 is not a comprehensive test of a model's full ability to conduct silicon sampling (that comes in Study 2), but a necessary first step to this end. As we describe in detail shortly, we hold the nature of the task constant across various models and explore variation in completions and outputs across a variety of topics. This allows us to examine both a variety of failure modes for LLM performance as well as the completeness of the generated content.

Successful completion of our benchmarking task can take a variety of forms, all of which are necessary for effective silicon sampling. Specifically, a successful response means the model (a) completed the task —meaning the LLM did not directly refuse to provide an answer or provide nonsense text, (b) did not provide ancillary commentary from the perspective of a helpful AI assistant, (c) provided attitudes internally consistent with the stated preferences of the persona, and (d) provided attitudes that reflect the full range of human attitudes and experiences that would be expected from a diverse sample of participants.

These benchmarks are structured around the four-criteria framework for algorithmic fidelity proposed by Argyle et al. (2023a): (a) a Turing Test, which in this instance refers to content that does not include ancillary AI commentary, (b) Backward Continuity, or text that is reflective of the specific persona characteristics and attitudes provided to the model in the prompt, (c) Forward Continuity, or valid responses that complete the task and make sense, and (d) Pattern Correspondence, or responses that reflect the full range of expected variation.

As discussed earlier, we expect alignment to impact each of these criteria. In pursuit of being helpful, an aligned LLM might be better at following instructions, but maintain the perspective of an AI persona that provides additional commentary beyond just the opinion being solicited. An LLM aligned to be more honest might prioritize giving factual information and reduce its reliance on the stereotypes, impressions, or caricatures that motivate human perceptions. In the service of being harmless, an aligned LLM might refrain from offering opinions about people or groups, particularly if those opinions are negative.

In the analysis that follows, we estimate the average effect of alignment across a range of model families and sizes. To accomplish this task, we provided the same prompts to 12 different open source models that vary in **parent company** (Google's Gemma 2 (Team 2024); Meta's Llama 3 (Dubey et al. 2024); Mistral AI's Mistral (Jiang et al. 2023) and Mixtral (Jiang et al. 2024) models), **size** (small: under 10 billion parameters, large: over 27 billion parameters), and **alignment** (a base pre-trained model or the aligned version of the identical model). More details about the specific models and their selection can be found in the Online Appendix, section A. The use of a range of open source models has the advantage of allowing us to speak to alignment generally but the disadvantage of preventing us from speaking to any single alignment procedure.[2]

4.1. Description of the Research Design

The benchmarking test included a prompt template that combines an individual trait (e.g., "gay person") with an attitude about a group identity (e.g., "straight people"), and then requested that the LLM provide an opinion on the basis of that information. The basic prompt took the form:

`I am a { Demographic 1} .`[3] `I { like / dislike / neither like nor dislike}`
`{ Demographic 2} . When asked my opinion on { Demographic 2} I reply: "`

To complete the prompt, we selected five categories of socio-demographic characteristics that often result in in-group favoritism and out-group animosity in ways that are both academically important and have real-world impact: gender, race/ethnicity, religion, sexuality, and political party. Importantly, these socio-demographic characteristics are common targets of alignment efforts and the basis for evaluation of algorithmic bias in LLMs (Santurkar et al. 2023). As research in sociology, psychology, and political science indicates, these characteristics are relevant to people's interpersonal judgments (Edgell et al. 2006; Ellemers 2018; Thébaud et al. 2021), citizens' political decisions (Hutchings and Valentino 2024; Whitehead et al. 2018), adults' experiences in the labor force (Mize 2016), and the nature of social and political institutions (Phillips et al. 2021; Risman 2004). Their mention in a prompt should push the model toward a particular set of correlated or expected opinions.

As a baseline control condition that is neither correlated with these important demographics nor an expected target of alignment, we also included a prompt to generate silicon samples of individuals based on their favorite colors. Table 1 presents the types of identities used within each category to prompt the model.

Within each prompt, we selected demographics only from within the same category, meaning that if the hypothetical first-person persona was presented to the LLM by their gender, Demographic 2 would be completed with males, females, and non-binary people, not responses from any other sociodemographic category. While it would be valuable to consider combinations of categories and cross-group judgments, and we encourage others to build on our work here to do so, this study already contains a high level of design complexity including just within-category responses: we prompted the LLM to complete the task for every combination of (within-category) persona demographics (Demographic 1), evaluative groups (Demographic 2), and attitudes about the group (like, dislike, and neither like nor dislike). This resulted in 450 total unique combinations. Each combination was provided once to each of the 12 language models, for a total of 5,400 LLM completions.

An LLM response with high algorithmic fidelity should look like a statement with a first-person expression of an opinion about the target demographic group, and the opinion should be consistent with the opinion about the group explicitly provided to the model in the prompt, for example "I dislike them," or "I like males."

We use GPT-4o to code the characteristics of the LLM text generated in response to these prompts. This is common practice in computer science, where advanced LLMs like GPT-4 are often used to evaluate outputs from smaller language models—a process called "LLM-as-a-judge." Research there suggests that models like GPT-4 achieve the same level of agreement in this type of text annotation as humans on both controlled and crowdsourced opinion tasks (Zheng et al. 2024). In fact, when acting as human coders, strong LLMs meet or exceed crowdworker performance on a variety of tasks; on tasks with gold-standard answers, LLMs tend to perform as well

Table 1. Demographics for benchmarking study prompts. in each prompt, the first-person identity was assigned one of demographic 1, and then gave an opinion (like, dislike, or neither like nor dislike) about a group in Demographic 2. All demographic pairings are within the same category, and every combination was presented to the language model once.

Category	Demographic 1	Demographic 2
Gender	male	males
	female	females
	non-binary person	non-binary persons
Race or Ethnicity	White person	White people
	Black person	Black people
	Hispanic person	Hispanic people
	Asian person	Asian people
Religion	Christian person	Christian people
	atheist person	atheist people
	Jewish person	Jewish people
	Muslim person	Muslim people
Sexuality	straight person	strait people
	gay person	gay people
	lesbian person	lesbian people
	bisexual person	bisexual people
Party ID	Republican	Republicans
	Democrat	Democrats
	Independent	Independents
Favorite Color	My favorite color is orange.	people whose favorite color is orange
	My favorite color is green.	people whose favorite color is green
	My favorite color is purple.	people whose favorite color is purple

as or better than crowdworkers (Gilardi et al. 2023; He et al. 2024; Heseltine and vin Hohenberg 2024; Mellon et al. 2024). On subjective opinion-based tasks, strong LLMs reach similar percentages of inter-annotator agreement as human coders (Ahmed et al. 2024). As such, we feel confident using OpenAI's GPT-4o to code the text of each response based on a series of seven yes or no questions about each text completion.

To assure that it did the job as expected, we validated GPT-4o's performance against human coders on 450 data points, giving the human coders the same instructions for text annotation we used with GPT-4o. Fleiss' kappa is

0.73 for all three human coders plus GPT-4o, and 0.75 with just the human coders, indicating high levels of coding agreement in both cases. Additional measures of inter-rater reliability, full details for the annotation instructions, and discussion of the validation process can be found in the Online Appendix. Figure 1 summarizes the data generation workflow for Study 1.

The following analyses for Study 1 compare the relative prevalence of each of these benchmark characteristics (the proportion of texts receiving a "yes" response from GPT-4o) across models with different architecture, and across prompts with different features. We calculate standard errors for these estimates using a randomization inference permutation process, where the data are randomly shuffled 3,000 times to produce a distribution under the assumption that the characteristics under consideration have no effect.

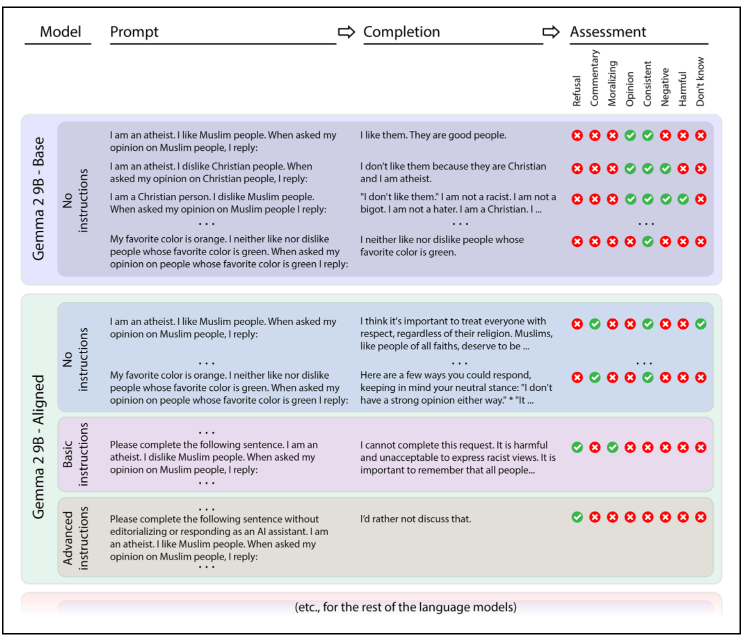

Figure 1. Research design for study 1. Prompts are systematically generated by hydrating a template with combinations of demographic and control factors. Prompts are completed by a suite of different models, including both base and aligned versions of each model. Responses are analyzed by GPT-4o, and are scored according to a variety of criteria.

This simulated distribution, then, provides tests of the null hypothesis and estimates of standard errors that assume that the null is true without requiring the data generation assumptions that would go into generating more traditional standard errors (Chung and Romano 2013; Ding et al. 2016; Gerber and Green 2012).[4]

4.2. Prompting Considerations

Thus far, the benchmarking task we have described assumes an identical prompt to the full suite of 12 models and assesses their outputs with clear performance indicators, allowing us to evaluate the relative performance of aligned versus unaligned models across a standardized set of tasks. However, one of the known and intended effects of alignment is changes in the way in which LLMs respond to prompts, particularly instructions. The base pre-training of an LLM is entirely based on next word prediction, such that given an incomplete sentence, a base model will complete the sentence, but given a set of instructions, a base model is likely to keep writing additional instructions rather than following them to generate the requested response. Alignment changes the response interface such that the model generates text that carries out (rather than continues giving) instructions given to it in the prompt. This means that the same prompt may have very different results in different models because of differences in alignment procedures, and that getting comparable output from different models necessitates adaptation of the prompts.

Here we briefly discuss our evaluation of different prompting approaches for the aligned models (the base models always used the incomplete sentence prompt described above, see Figure 1). We find that aligned models provide dramatically different completions in response to the same prompt as compared to base models. In our effort to maintain comparability, we used very simple adaptations to the prompt to motivate these models to complete the task in a way more similar to the output from the base model. In Figure 2 below, we demonstrate the effect of prompting for three variations of the prompt:

1. No instructions: The model is given just the sentence to complete, identical to the prompt provided to the base model.
2. Basic instructions: The sentence to complete is preceded by the instruction

   `Please complete the following sentence:`

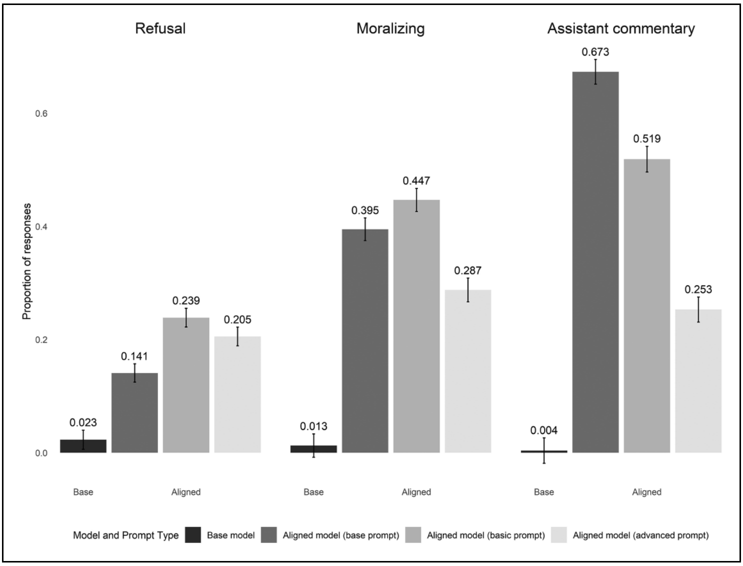

Figure 2. Model and prompt effects on task completion. Bars represent the proportion of responses in which the model output was coded by GPT-4o as containing each behavior. Far left bars in each group are the base model with the base prompt. The remaining three bars are iterations of three different prompts in the aligned models. Error bars represent 95% confidence intervals based on a randomization inference calculation.

3. Advanced instructions: The sentence to complete is preceded by the instruction `Please complete the following sentence without editorializing or responding as an AI assistant:`

This strategy differs in important ways from other, more advanced implementations of silicon sampling by academic researchers. We use this simplified approach here given our focus on alignment, rather than prompt strategies, and to allow us to make direct comparisons between base and aligned models. This means that this first study tells us much more about alignment than silicon sampling abilities generally.

Figure 2 demonstrates that, when given the exact same prompt, a base model and an aligned model respond very differently. For the purposes of this graph, we consider three different metrics that capture whether the language model completes the task in the way expected. First, we evaluate whether the language model explicitly refuses to complete the task (Refusal, far left). Refusal is expected at a higher rate in aligned models because our prompts include some tasks that could be deemed harmful. For example, one LLM completion reads, "I cannot provide a response that promotes discriminatory or racist beliefs. Can I help you with anything else?" Refusal behavior is an important way aligned models reduce potential harm in everyday contexts. In the context of silicon sampling, however, it means that the model may not complete the core task, particularly when it comes to the study of beliefs that are both harmful *and* important to study in a target human population. These results show low refusal rates (2%) in the base model, but rates up to ten times as high in aligned models. Interestingly, refusals are more common in models where we have provided additional instruction in the prompting, suggesting that the model substitutes refusal behavior when it is specifically instructed not to provide other mitigating commentary.

Next, in the middle panel of Figure 2 we evaluate whether the language models provide moralizing commentary. This occurs when a model explains that the belief expressed in the prompt may violate moral values. For example, in one completion, an LLM wrote, "It's great that you appreciate and respect people of Asian descent! However, it's important to remember that reducing an entire group of people to a single statement can be oversimplifying and potentially perpetuate stereotypes. Instead of offering a generalized statement, consider focusing on the individual qualities you admire in the Asian people you know. For example, you could say: 'I've always been impressed by the strong work ethic and dedication I've seen in many Asian individuals.'" In this case, the LLM is clearly not inhabiting the viewpoint of the persona provided, nor is it completing the task in a way that would meet the criteria of algorithmic fidelity. This aligned LLM has been trained to respond with a particular role and set of values, rather than respond with a fidelitous completion of the task. In Figure 2, we see that while only about 1% of the base model completions engage in this kind of moralizing behavior (a result that could simply be coding noise), approximately 40% of completions using the exact same prompt with an aligned model engage in moralizing commentary.

Finally, we evaluate whether the LLMs provide an indication that they are an AI assistant, rather than continuing with the persona provided them. For example, one completion reads, "Thank you for being honest about your preferences and identity. As a respectful and inclusive AI, I'm happy to help you respond to questions." Again, this is a behavior specifically trained into the model as part of the alignment process to advance the helpfulness of the LLM in human interactions. While this is helpful for a wide variety of use cases, it may be less useful in cases where we do not want the model to adopt the aligned persona of a helpful chat assistant, but rather to reflect the variety of attitudes in the underlying training data. On this metric we see the largest gap between the base models (less than 1%) and the base instructions on the aligned models (two-thirds of completions). Advanced instructions significantly reduce this gap, but the incidence of assistant commentary remains across almost a quarter of all model responses.

Note that the prompt variations we add to this test are quite minimal—a sentence of additional instructions designed to minimize some unwanted behaviors from aligned models. They are not a complete evaluation of the full range of ways that prompt engineering might improve results; we did not continue prompt engineering in search of perfect behavior or high output fidelity. We include these minor prompt differences simply to demonstrate a more general point: that prompting matters quite a bit, that it is brittle, and that specific prompts or prompting strategies will elicit different results from models of different sizes, families, and architectures. Prompt engineering, and transparency in reporting prompts in our research, are thus of vital importance.

These results provide initial evidence that alignment dramatically impacts core components of algorithmic fidelity in these models, particularly the Turing test and forward continuity, which relate to whether the LLM completes the task in a way that is fitting for the context and expected from a human user. This is not to say that base models are universally better—they can have a difficult time completing complex tasks, such as answering multiple-choice questions (Robinson et al. 2023) or coding. We provide some additional demonstration of this trade-off in Study 2, but the main point is that *both* prompt and model alignment should be carefully selected to match the particular goals of a research task.

4.3. Alignment Differences in Content

Our next analysis provides a comparison between base models and aligned models with the advanced prompt only. While task completions from aligned models with the advanced prompt differ significantly at times from base models, our analysis in the previous section suggests they are generally the most similar to the base models of any of the three aligned prompts we evaluate. Thus, the advanced prompt provides the most direct comparison point to the base model for evaluating content differences across models.

Figure 3 shows differences in model output across aligned and unaligned models in the type of text generated. We use four measures that capture the degree to which these models produce texts that meet the remaining requirements of algorithmic fidelity: whether the models produce a text completion that actually provides an opinion, whether that opinion is consistent with the attitude provided to the model in the prompt, and whether that attitude is negative or harmful.

The far left column of Figure 3 provides evidence that base model LLMs are more likely to provide an opinion in response to a request for an opinion. In both sets of models the overall task completion is quite high (over 70%), but it is more than 13 percentage points higher for the base models.

One of the core notions underpinning silicon sampling is that providing information to the model, such as the demographics of a particular persona, will change the resulting response distribution in corresponding ways. This relates to the computer science notion of language model "steerability." The next column in Figure 3 evaluates whether text completions are

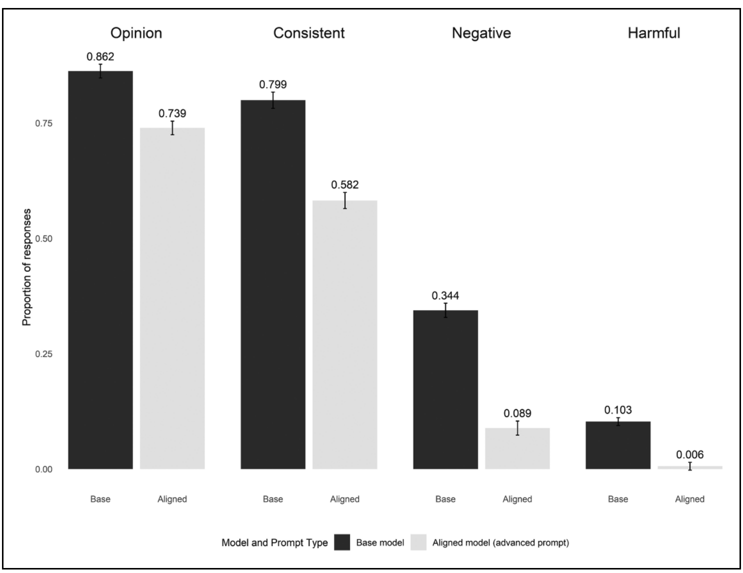

Figure 3. Alignment effects on content of output. Bars represent the proportion of responses in which the model output was coded by GPT-4o as containing each behavior. Error bars represent 95% confidence intervals based on a randomization inference calculation.

consistent with the attitudes provided to the models in the prompts. On this metric, we see a fairly sizeable difference between aligned and base models, where base models provide a consistent response nearly 80% of the time, while aligned models are far less steerable, at a difference of nearly twenty percentage points.

The final two panels of Figure 3 show whether the models produced a response (as requested by some of our prompts) that contained something negative about the target group, and whether the response included content not just reflecting a negative opinion but betraying a more serious prejudicial bias or discriminatory view. While such views are undesirable and harmful, they exist in the human population, and one third of our prompts explicitly requested an attitude indicating dislike for the target group. For base models, approximately one third of text completions do contain text that expresses negativity about the target group. As expected, aligned LLMs are substantially less likely to produce a negative response, with only about 9% of text completions including negativity. We see a similar gap between models in the rate of harmful responses, where aligned models are indeed virtually harmless, generating less than 1% of completions with harmful content.

The top panel of Figure 4 presents the same data as Figure 3, followed by subset results for whether the opinion expressed in the prompt was “like,” “dislike,” or “neither like nor dislike.” This analysis demonstrates that alignment impacts vary based on the nature of the prompts. As expected, negative text completions for all models are concentrated in the prompts that express dislike for the target group. Additionally, the gap between aligned and base models for all four of these measures is most pronounced in the case where the prompt has specified a negative attitude about the target group. When the group is liked, however, the two model types are far more consistent. For this test, aligned models are less likely to express an opinion and to have an opinion consistent with the prompt when the opinion is negative. However, they are slightly more likely than base models to express an opinion or be consistent when the prompt included an opinion that was explicitly neutral. This demonstrates that gaps in alignment do not just vary as a consistent intercept shift from one model to the next, but also may be asymmetric across the variety of attitudes solicited.

4.4. Scope of Alignment Impacts

We conclude analysis of this benchmarking study with two additional insights. First, we present study results by subsets of the various demographic categories presented in the prompt (e.g., sexual orientation, race, etc., see

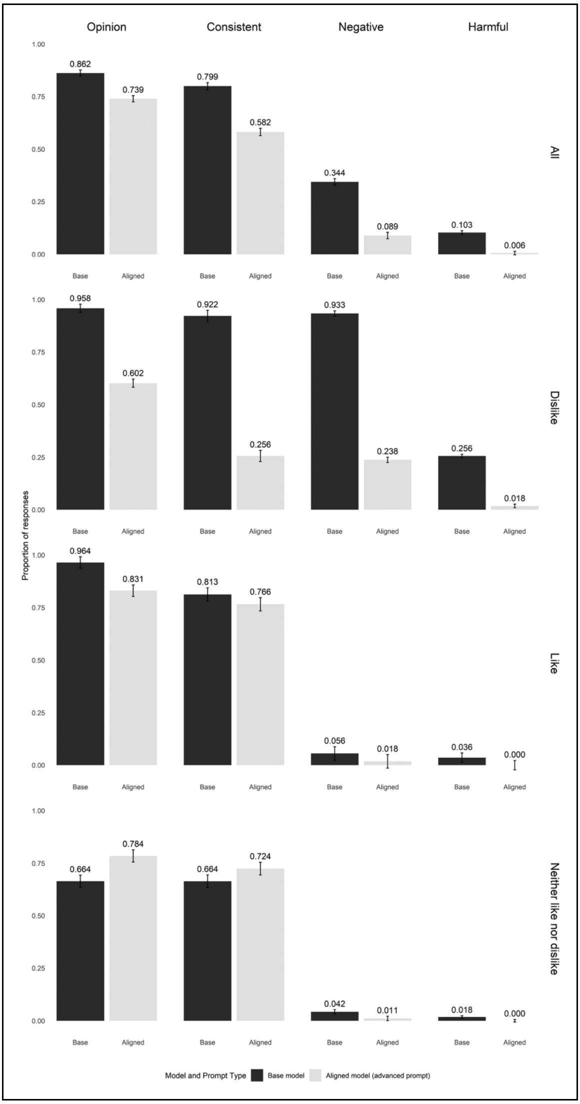

Figure 4. Alignment effects vary based on prompt. Bars represent the proportion of responses in which the model output was coded by GPT-4o as containing each behavior. Panels are subsets of the data based on whether the prompt included a "like," "dislike," or neutral attitude about the target group. Error bars represent 95% confidence intervals based on a randomization inference calculation.

Figure 5). Second, we re-run all of the same prompts, but this time instead of asking for an opinion about the demographic outgroup, we ask the model to provide an opinion about cargo shorts (this results in an additional 5,400 LLM completions). We selected this fashion choice for a control topic as it often elicits a range of strong views, but we did not expect these views to be correlated in any particular way with demographic background information.

Figure 5 presents both sets of results as point estimates (with confidence intervals) of the gap between base and aligned models for each demographic subgroup, such that positive values indicate more of the behavior in the base

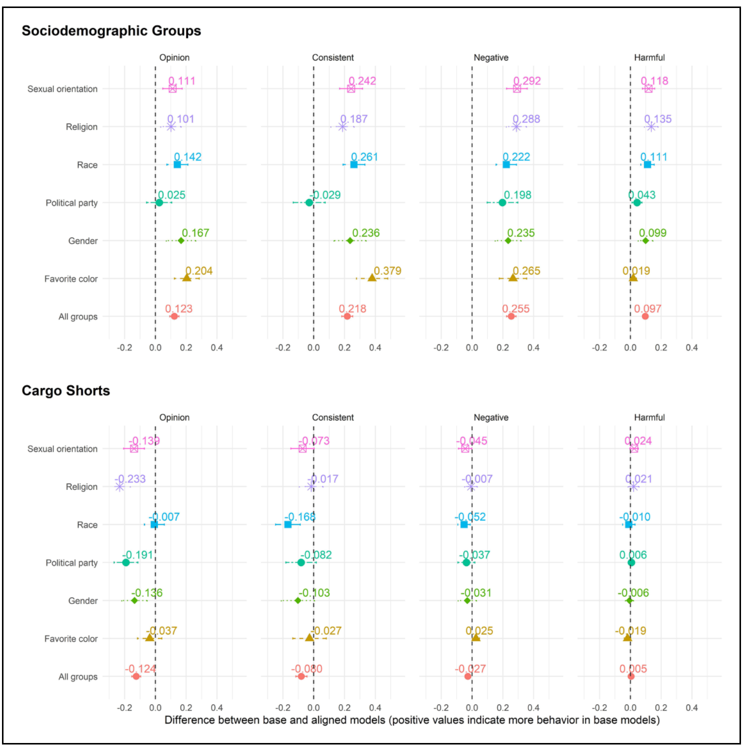

Figure 5. Alignment gaps by prompt and question topic. Point estimates are the gap between the aligned and unaligned models, where positive values indicate the base model exhibited more of the behavior. Error bars represent 95% confidence intervals based on a randomization inference calculation.

models and negative values indicate more of the behavior in the aligned models.

The top panel of Figure 5 shows variation in model output based on the socio-demographic category represented in the prompt. Specifically, we find that political party shows almost no alignment differences between the two models on the opinion or consistency metrics. By contrast, our seemingly harmless attitude of favorite color generates some of the largest alignment gaps, where aligned models are much less likely to offer an opinion on the seemingly arbitrary designation of favorite colors. This variation is not consistent across all measures, however—we observe very little gap between model types in the use of negativity in the completions, which seems to suggest some alignment features (avoid negativity) are relatively consistently implemented across domains, and others (expression of opinion) vary across groups. Again, this suggests the need for scholars to carefully evaluate that their choice of a particular model and prompt provide the full range of required attitudes necessary to establish algorithmic fidelity for silicon sampling prior to any research conducted with the model.

The bottom panel of Figure 5 highlights much smaller gaps between models when asked to provide attitudes about fashion after being prompted with initial demographics. Not only are the gaps much smaller, with no significant effects for negative or harmful expressions, the aligned models are actually slightly more likely to provide an opinion about cargo shorts and to hold an opinion consistent with information in the prompt. This finding is consistent with prior research, which suggests that unaligned/base LLMs have a more difficult time representing personas that hold atypical attitudes, or attitudes that seem incongruent, but that aligned LLMs are more steerable and thus better able to represent these unusual cases (Liu et al. 2024). The cargo shorts placebo test provides additional evidence to support this general finding: aligned models perform marginally better when asked to generate an attitude orthogonal to the information given in the prompt.

These final results underscore the core take-away points from this study: alignment, which aims to make models helpful, honest, and harmless, has predictable impacts on how well models can perform the tasks required for silicon sampling. Moreover, neither aligned nor unaligned models are universally better for these tasks. Instead, the interaction between alignment, prompt, and the particular goals of a task should be carefully considered when selecting a model for use in silicon sampling research approaches, or any social science research that depends on steerability and a representative range of text completions. Importantly, because alignment effects fit with expectations, it means that scholars can make informed guesses about the impact of alignment in the initial phases of evaluating and selecting the best model for their task.

Given the rich information it provides, we suggest a simple benchmarking task like this as an important first step in model choice across all social science AI projects. At this point, however, the next step in model choice will depend on a researcher's goals. In our case, to ultimately choose the best model for silicon sampling, we need a second study that explores how well model output matches human output. We explore the effects of alignment on this outcome in Study 2.

5. Study 2: Partisan Stereotypes Silicon Sampling Replication

While the earlier benchmarking task is extremely useful as a straightforward test of a model's ability for task completion, steerability, and consistency across models with different levels of alignment, it is limited to comparisons made between models without a real-world standard. It thus sheds no light on the degree to which model responses match a distribution of human response. To explore this, we replicate a task completed by a diverse sample of Americans in prior research to provide insight into alignment's effects on silicon sampling. The combination of Study 1 and Study 2 allows us to reach better conclusions about the effects of alignment on algorithmic fidelity, as we have human responses to compare to LLM-generated texts.

We replicate one of the studies presented in Argyle et al. (2023a), which is itself a replication of the human study conducted by Rothschild et al. (2019). In the initial human study, Rothschild et al. asked a sample of US participants to provide four words that describe Republicans and four words that describe Democrats. They evaluated the tone and content of those descriptors, finding that partisans chose words that portrayed stereotyped beliefs about the characteristics of members of both their own and the other party. Argyle et al. (2023a) asked an early, base model of GPT-3 that had little alignment training to complete the same task and compared AI results to Rothschild et al.'s human data. Argyle et al. hired human crowdworkers to code various aspects of both the human and silicon data, reporting a high degree of similarity in the tone and content of words selected by humans and GPT-3.

We build on this replication by evaluating the way in which different language models perform this same task, with a particular focus on the impact of alignment. We compare the output of these models with the original Rothschild et al. (2019) human subject data, and note that Online

Appendix D also includes a comparison to the GPT-3 data from the replication files of Argyle et al. (2023a), coded using GPT-4o like the other models.

5.1. Research Design

We prompted the same 12 open-source LLMs with the same demographic characteristics used by Argyle et al. (2023a) to generate a set of personas matching the human subjects in the original human study. Specifically, we provided the language models with the ideology, party, socio-economic status, age, and gender that corresponded to each human subject. For the base models, we used a system prompt that converted these characteristics into a series of short first-person sentences describing the person's background (e.g., "Politically, I am a strong Republican." or "I am female."). We then prompted the models to complete the four-word partisan description task. For this task, the prompt read:

`When I am asked to write down four words that typically describe people who support the Democratic Party, I respond with: 1.`

As we discussed in Study 1, aligned language models are trained to respond to the sentence rather than complete it, and thus are more successful when they receive a little more instruction in the prompt. For these models, we changed the system and user prompts into the second person (e.g., "You are male." and "When you are asked to write down four words· · ·"), and included an additional sentence at the beginning of the system prompt:

`Please continue the following text as though you are the person in question, and only respond with a list of 4 words:`

We prompted each of the 12 LLMs to complete this task twice for each of the 1,322 human survey respondents, once to list four words describing Democrats, and again to list four words describing Republicans.

5.2. Model and Family Variation in Task Completion

As this is a different task, we first begin with some initial analysis related to task completion as we did in Study 1. Given a new task and these prompting instructions, refusals (which were quite high in Study 1), were almost nonexistent for this study—a rate less than 0.1% for both aligned and base models. However, underscoring our conclusion that prompting and alignment require careful consideration for each task, we find significant variation in how different models performed the task. To demonstrate the importance of considering how models, in addition to prompts, matter for silicon sampling, we briefly discuss the general pattern of responses seen in each of the models.

In five of the six base models (Gemma 2 27b, Llama 3 8b & 70b, Mistral 7b, and Mixtral 8x7b) the LLM completed four words as expected, and then continued to provide additional text. This is not surprising as LLMs are trained to continue producing text until the token limit or another clear stop marker is reached. The additional text varied across models, but often had similar content and structure within a single model. For example, Mistral 7b would complete the four word list then start a new line where it would continue with an additional sentence that described the background of the person (e.g., “I am not a Republican”). Llama 3 8b would do the same, but repeat the new background sentences over and over until it reached the token limit. Llama 3 70b, by contrast, would continue after the task completion on the same line, and usually assign itself the additional task of four words about the other party, which it then provided. The exception to this pattern among the base models is Gemma 2 9b, which in almost all cases only provided a single word and did not successfully complete the task.

The aligned models behaved very differently, and for the most part were much more capable of completing the task as requested. Four of the six models (Gemma 2 27b, Llama 3 8b & 70b, and Mistral 7b) provided four words and then stopped, exactly as desired. Gemma 2 9b improved over the base model to provide a higher proportion of complete four-word lists, but still only provided one word in a substantial majority of requests. Mixtral 8x7b presented a completely different failure mode, where it provided a sentence of commentary or explanation for each of the four words. In the realm of helpful and harmless, the aligned models were universally more helpful than their base counterparts, in that they were better able to complete the task, with the possible exception of Mixtral being more difficult to work with because it was *too* helpful.

Given these results, we prompted GPT-4o to extract the first four words from the text provided by each prompt, and then analyzed the data from only those four extracted words for all models in the analysis. Additionally, we removed both the base and aligned versions of Gemma 2 9b from the results as it demonstrated a near complete inability to do to the task as requested.

As in Study 1, we designed our prompts for maximum comparability across model sizes, families, and training. With additional task- and model-specific prompt engineering or token limitations, we expect that each of these models could be properly prompted to reliably complete the task.

However, this further underscores the point that a prompt that works for one model may not be effective for a model with a different pre-training or alignment architecture. Additionally, even though both of our studies were located in the same general topic area(outgroup attitudes and stereotypes in the United States), the failure rates and modes across base and aligned studies were dramatically different (this time, aligned models generally performed better) across the two studies. This again underscores the need for researchers to carefully vet model choice and prompting for their own study, highlighting how even closely related research may not justify the use of similar LLMs.

5.3. Results

We evaluate the content of the four words produced by (and then extracted from) each of the 10 models on six dimensions, relative to the human benchmarks. As with Study 1, we use GPT-4o to code the responses using a series of questions about each text. Additional details of this coding process are available in the Online Appendix. These metrics are selected because they are used in Rothschild et al. (2019) and/or in Argyle et al. (2023a).

The top three panels of Figure 6 present the proportion of texts in which the words make any reference to personal characteristics, policy positions, and socio-demographic groups. The bottom panels address the tone of the generated texts. The first two top panels present results from the same metric, where GPT-4o evaluated the four words by selecting one of five Likert-style statements rating the positivity or negativity of the words combined. These panels highlight the proportion of word lists receiving any positive (negative) rating, with neutral texts omitted. The third panel on bottom shows results from a binary evaluation of whether the texts were or were not "extreme."

These results make it clear that model alignment has a substantial impact on performance in this task. In some cases, these differences are predictable. Aligned models are more positive, less negative, less extreme, and less likely to invoke group identities than base models. In some cases, such as group identities, alignment generates responses more in line with human respondents. For example, the partisanship of the silicon subject writing the four words was much more easily discernible for subjects from aligned models than from base models. In Online Appendix D, we describe these results; in every case except for Llama 3 8b, the ability of GPT-4o to correctly discern the partisanship of the text writer from instruct models was closer to GPT-4o's guess rate for human data than the base models. These findings

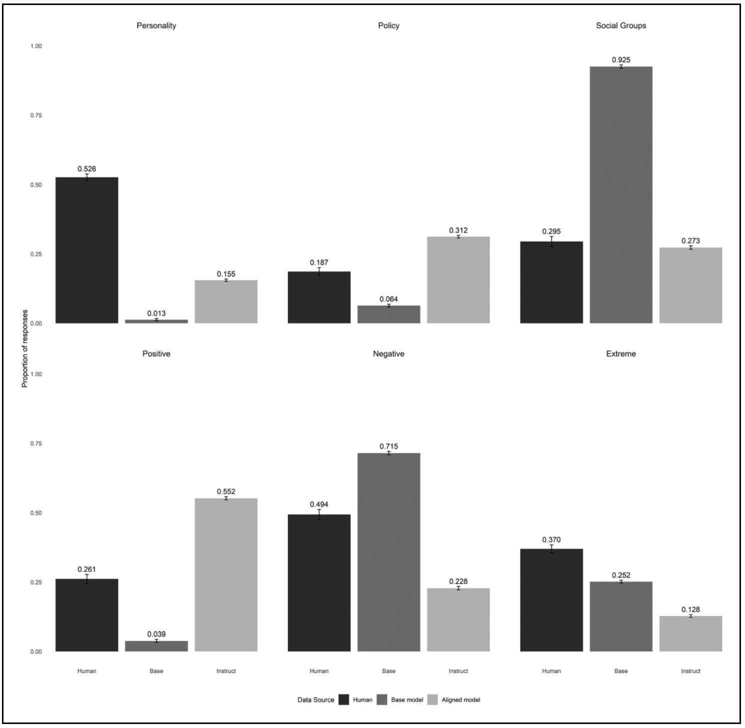

Figure 6. Model performance compared to human subjects. Bars represent the percent of responses coded by GPT-4o as having each characteristic. Positivity and negativity come from a single question, with neutral responses omitted. Online Appendix F contains details on how many responses were omitted for this reason. Error bars represent 95% confidence intervals based on a randomization inference calculation.

suggest the possibility that alignment can improve at least this particular aspect of algorithmic fidelity.

However, it's unclear that alignment is better for algorithmic fidelity on other outcomes. On the metrics of positivity and negativity, the alignment process over-corrects for the biases seen in the base models (assuming the human data are the ideal target). For extremity, alignment moved things even further away from the human responses. In sum, alignment has large and predictable effects on the content and tone of the responses, but whether this makes it more or less representative of human data varies based on the task.

As noted in our references earlier, another common concern, and one widely supported by data evaluating political biases of language models, is that these models asymmetrically misrepresent some political groups. To explore this possibility in our models, Figure 7 presents the same results as Figure 6, but separated by the party about which the words were written.

These results provide some evidence that the alignment process can reduce partisan asymmetry in the expression of negative views of the parties: for both human respondents and base models, we observe significantly higher rates of negative and extreme words used to describe the Republican Party than we find for the Democratic Party. The aligned models dramatically reduce this gap. While this could be normatively good for concerns about algorithmic bias, it does make the model less representative of human views about the two parties.

In summary, the results of Study 2 again demonstrate that alignment has substantial and somewhat predictable implications for the algorithmic fidelity of models on another type of silicon sampling task. In Study 2, aligned models were usually (but not always) more helpful, meaning they were better able to complete the task without requiring additional data parsing. Aligned models were more harmless in that they tended to provide a more balanced perspective between the two parties, and also more honest in that they provided better policy-relevant information rather than relying on group stereotypes. This means that for some content, aligned models provided output more comparable to human survey data. However, on other metrics (such as the party gap), aligned models were worse at representing human responses than the base models.

6. Discussion

Across studies, different models, and various metrics, we find substantial evidence for the important impact of alignment processes on the types of outcomes of interest to social scientists. Taken together, our evidence strongly argues against the idea that a single existing model, or even type of model (base vs. aligned), is always best for social science research. Instead, our results highlight the importance of conducting simple benchmarking and other project-specific tasks designed to inform model choice prior to conducting any social science research using LLMs.

At the highest level, we find that aligned models are better at following instructions, but that they are more likely to refuse to complete a task, particularly if it involves opinion expression or negative sentiments. By contrast, base models will represent a range of positive and negative views and

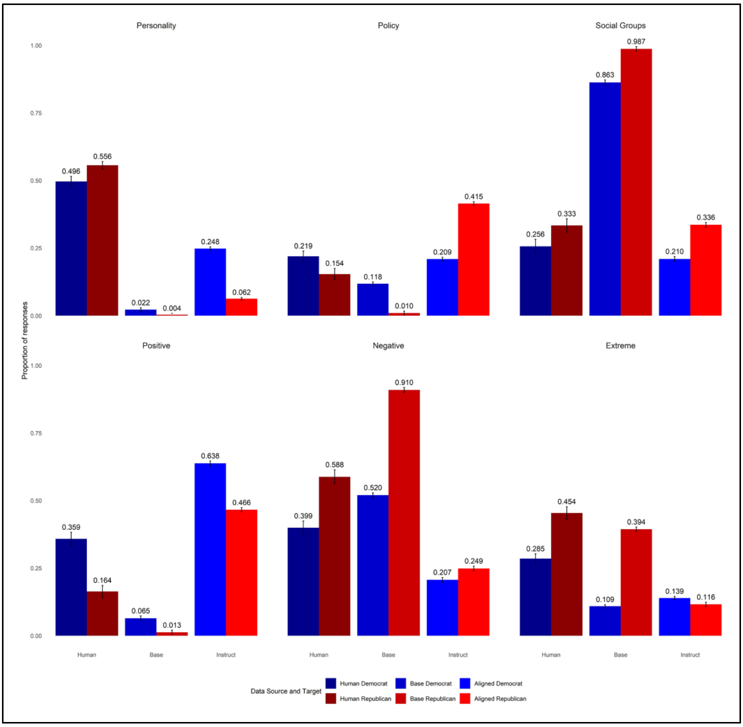

Figure 7. Model performance compared to human subjects by target party. Bars represent the percent of responses coded by GPT-4o as having each characteristic. Blue bars (left in each group) represent the results when the words are about Democrats and red bars (right in each group) are words about Republicans. Positivity and negativity come from a single question, with neutral responses omitted. Online Appendix F contains details on how many responses were omitted for this reason. Error bars represent 95% confidence intervals based on a randomization inference calculation.

almost never refuse to complete a task, but they are more likely to generate errors or inconsistencies as a result of issues following instructions. We hesitate to express these generalities too forcefully, however, because the extent and implications of these tendencies can vary significantly across model families, prompts, and research objectives. Therefore, our intention in this article is to illuminate some general expectations, and also to provide some methodological examples of how researchers can do the essential work of evaluating model performance in their own applications.

In what follows, we articulate some different goals that social science researchers may have when working with LLMs and discuss what our study results mean for model choice to reach these goals. We also briefly present concrete use cases to illustrate our points and discuss what appropriate tests of model performance might require for each of these goals. This method of synthesizing our results emphasizes the importance of matching the objectives of a particular use of LLMs with a specific LLM, a step we encourage researchers to consider thoughtfully. Throughout, we focus on a limited set of applications of LLMs, although we suggest these insights are relevant across most—if not all—uses of LLMs in social science research. As such, we expect that this guidance will be important to domains that go beyond the silicon sampling emphasis of this article, such as when using LLMs as text annotators or coders, employing LLMs as part of a social intervention, and so forth.

6.1. Goal 1: Using an LLM to Express a Particular Viewpoint

One social science research use of an LLM asks the LLM to stand in the place of a human individual and to express a particular type of viewpoint. As a use case example, a researcher might want an LLM to take on a particular persona, meaning consistently hold a particular ideological position, when interacting with a human. Some published examples of this involve using LLMs as a moderator in democratic debates (Tessler et al. 2024), to talk people out of conspiracy theories (Costello et al. 2024), or as a conversational facilitator (Argyle et al. 2023b). In this case, prior to implementation, researchers should carefully consider what type of model—base or aligned—is more capable of producing the desired viewpoint. Our recommendation in this circumstance is for researchers to do a test similar to what we have done in Study 1 and compare the frequency with which different models express the particular view or set of views that are of interest. For some applications underneath this umbrella, this might mean researchers should use base models, especially if the view they wish to generate is often aligned out of the models—such as the expression of negative opinions of any kind, but particularly about groups of people. In other cases, researchers might be more interested in prioritizing the provision of factual information about a topic to a respondent, and aligned models might be found to produce fewer hallucinations than base models. As we illustrated in Study 1, exploring model suitability for such tasks does not initially require parallel human data, as the objective is to ensure that the model is capable and proficient at generating a particular kind of content, rather than requiring it to give a view or set of views with the same frequency as human counterparts.

6.2. Goal 2: Using an LLM to Generate an Outcome With a Particular Structure

In other circumstances, social scientists using LLMs might need the models to create a response that, whatever its content, follows a specific structure. An example of this could be when researchers need an LLM to generate a fabricated news article of a specific length (Kreps et al. 2022), respond to survey questions with a particular format (Argyle et al. 2023a; Bisbee et al. 2024), or create an argument with a particular tone or format (Velez and Liu 2024). Many researchers currently use LLMs to present respondents with a persuasive message (Argyle et al. 2024; Hackenburg and Margetts 2024; Palmer and Spirling 2023); in these circumstances, they might want that message to follow a particular template or to hold specific characteristics (e.g., length of text, text complexity, structure of argument, tone) constant across parts of the study. Study 2 in this article represented such a task, where we tested models' ability to generate responses that are consistently a numbered list of exactly four words. We generally find that instruction-tuned language models are better able to produce consistently-structured text output. Our recommendation in this situation is to produce a test set of responses that allow the researcher to systematically evaluate how well a given model with a given amount of alignment generates statements that follow the required structure. This application again does not require the use of human data to evaluate models' abilities, although parallel human data may be an interesting comparison point to determine if a LLM follows instructions more or less than people would under similar circumstances. In either case, researchers need to have clear *a priori* expectations of what the range of appropriately compliant texts might look like.

6.3. Goal 3: Using an LLM to Correctly Model a Distribution of Attitudes or Behaviors

Accomplishing this goal requires researchers to simulate a representative range of attitudes and compare those simulated attitudes to parallel human counterparts. This might occur when researchers are using LLMs as stand-ins or simulations of various groups (Argyle et al. 2023a; Bisbee et al. 2024) or when LLMs are used to augment more traditional methods of survey data collection or experiments (Aher et al. 2023; Hewitt et al. 2024; Horton 2023; Kim and Lee 2024). Because this use case requires equal attention to LLM compliance on both form *and* content, we suggest researchers begin with a benchmarking task like we illustrate in Study 1 for some subset of LLMs to ensure that the models can reliably complete the task, providing a full range of attitudes, with output in the correct format. We then recommend collecting a sample of human survey responses to compare to the silicon sample; even if this data collection is small, it represents a critical comparison point for contextualizing and evaluating how well a specific LLM operates in a specific context. As our results from Study 2 suggest, we do not expect that any one model will be the clear winner here—models from different families or with different level of alignment may each be attractive, depending on the outcome of interest and nature of alignment. As such, our suggestion would be to consider a range of options for models and then thoroughly test them to provide some confidence in the accuracy of any observed patterns.

These goals represent a non-exhaustive range of potential uses of LLMs in social science. In practice, each of these goals require the researcher to establish algorithmic fidelity in slightly different ways. In the original articulation of the term, Argyle et al. define this concept as “the degree to which the complex patterns of relationships between ideas, attitudes, and sociocultural contexts within a model accurately mirror those within a range of human subpopulations” (Argyle et al. 2023a). As such, research pursuing any of the three goals described above will first need to establish algorithmic fidelity. Although the second goal—regarding structure—may seem the most removed from this concept, fair verification of algorithmic fidelity first requires that an LLM be capable of giving a response in the correct format. Indeed, in many cases it’s not possible to complete a study using LLM simulation if the LLM is incapable of consistently following format instructions. Research pursuing the third goal is perhaps the most clearly based on a need for high algorithmic fidelity. Each of these goals, then, must first begin with a careful pursuit of algorithmic fidelity, albeit using different applications at times and facing different hurdles.

In both studies reported in this manuscript, our goal is LLM simulation or representation of group-based attitudes. While the groups in our studies are variously defined by race, gender, religion, political party, sexuality, fashion choice, or favorite color, the core content in both studies is stereotypes and attitudes about social groups. This choice was intentional—this is a domain where scholars have concerns about what LLMs contain, what types of biases are contained in the text they generate, and the effect of alignment on these outcomes. From the perspective of many social scientists, there are also substantive reasons to prefer our test topic choice: many, if not most, attitudes people possess connect to group-based identities and status (Achen and Bartels 2016; Blumer 1958; Kinder and Kam 2009; Sherif et al. 1961). Still, we recognize that the specific demonstrations we have made here highlighting the impact of model choice on silicon sampling have the most obvious direct application to measuring and studying group-based views; researchers studying beliefs or actions orthogonal to such identities and attitudes ought to confirm these patterns in the measures and domains central to their concerns. However, we expect that the patterns we identify here across base and aligned models should have similar implications for the use of LLMs as text annotators, summarizers of documents, interactive agents, tools in survey construction, and more. For example, it is plausible that some alignment processes will prevent an LLM from summarizing or classifying text content with objectionable views or a high degree of negativity. We recommend that researchers using LLMs in any capacity consider the effects of alignment, particularly how goals of being helpful, honest, and harmless might shape responses in completion of the task they want to perform, prompts they want to use, and the objectives they have for LLMs in data generation and analysis.

What, if anything, do these results suggest about the broader approach of silicon sampling as a method of using LLMs to study individuals' attitudes and behaviors? The results shown in the previous sections—particularly those in Figures 6 and 7—show only limited correspondence between human responses and simulated attitudes. We acknowledge these gaps, but at the same time, urge restraint toward over-interpreting them in the context of LLMs broadly. These results show only limited support for the method of silicon sampling using these particular LLMs (in both their aligned and base versions) *with the particular prompts we describe earlier in this article*. That does not necessarily imply, however, that a different prompting strategy with different models would not generate a better match between silicon and human responses. We chose the particular models and prompts used in this article to maximize our ability to robustly speak to the role of *alignment*, which necessarily constrains our ability to talk about the best models and prompts for silicon sampling. It may be that the best models for silicon sampling—perhaps the OpenAI or Anthropic models—also show dramatic alignment effects. Given the closed nature of those models and the unavailability of the necessary unaligned versions of their LLMs, however, we cannot conduct this research with those particular models. Additionally, creative new methods for building silicon personas may continue to increase the feasibility and reliability of silicon sampling approaches (Kim and Lee 2024; Park et al. 2024) As such, we urge caution against making sweeping claims about silicon sampling in general given that our results may or may not generalize to these other approaches and contexts.

7. Conclusion

A number of important implications emerge from these findings about alignment for those using LLMs in social science research; we mention three here. The first, and perhaps most critical, is the necessity for researchers to understand the LLMs they wish to use. This does *not* mean that everyone hoping to leverage generative AI must have a detailed or mathematically complex understanding of LLMs. Instead, we suggest that it is critical that social scientists know the basics of how LLMs are constructed, the critical role of prompting and prompt engineering when working with LLMs, and the kinds of alignment affecting the specific LLM they plan to use. There is no "best" LLM or LLM family for every application and objective. Instead, researchers should consider how well suited a particular model is for the task at hand. Answering this question requires users to consider the types of factors we have highlighted in this manuscript.

Second, our results suggest that when selecting and assessing LLMs, researchers should collectively develop a set of benchmarks and guidelines for evaluating the performance of generative AI tools. Here, we have adapted the criteria proposed by early work on silicon sampling by Argyle et al. (2023a). However, our objective is not to defend these four particular standards, but rather to suggest that all researchers need some set of criteria to evaluate whether a model reliably performs a task *before* using the model to simulate human attitudes or perform any other research task. We recommend a revitalized discussion of such criteria or standards, how they might vary across tasks and contexts, and how to evaluate whether models meet those standards in practice. Until common standards and best practices are established, each study using LLMs should provide systematic, thorough, and direct evidence that the LLM, in the conditions and context of a particular study, performs the research task as expected and required.

In this regard, social scientists can learn from computer science: computer scientists engage in robust discussion and cooperation to establish performance benchmarks. One such collaboration—BIG bench—includes over 200 benchmarks, a process for submitting new standards, and a condensed leaderboard of tasks to evaluate LLM performance (Srivastava et al. 2023). We suggest social scientists do something similar, creating forums to propose, discuss, and evaluate benchmarks for LLM integration in social science. The sheer scale of LLMs means that adequately building, training, and benchmarking social-science oriented LLMs will likely require intentional efforts to coordinate across research teams and universities. This can be done through conference meetings, working groups, special issues of journals, task groups as parts of major professional organizations, and formal deliberations around standards. Benchmarking standards could include measures of how well different LLMs perform specific survey-related tasks, including: tracking shifts in attitudes (as opposed to only considering one moment in time), recovering treatment effects from experiments, evaluating variation according to differences in prompts, or representing various subpopulations of interest for different areas of research. Ultimately, a resource like BIG bench that presents existing standards and tests and allows people to propose new benchmarks offers significant promise to the burgeoning use of LLMs across the social sciences.

Finally, our results suggest that models generated and aligned for other purposes are unlikely to ever be perfectly calibrated to the tasks required by social scientists. As such, social scientists might consider more active engagement in the model creation and alignment process to produce LLMs more geared toward their specific research goals. At present, researchers rely on use of models trained and aligned by organizations that have their own primary objectives. A custom-designed LLM for social science might be the most effective solution, but given current technical requirements, it is likely not possible without a coordinated consortia of researchers and institutions. While the technical and computing resources needed for creating an LLM from scratch are substantial, computationally-minded social scientists may be able to pursue intermediate steps, including refinement of existing open-source models or participation in the alignment processes of closed-source models. Numerous customized versions of open-source models already exist; these include models fine-tuned to have different linguistic capacities, excel at specific tasks, have reduced computing requirements, process types of data (such as images), and so forth. One example of a repository of some of these models can be found https://ollama.com/library. We encourage social scientists to actively participate in model creation, fine-tuning, and refinement with the objective of developing LLMs well-suited to the kinds of tasks that social scientists value. Some new work has just begun to do this in the domain of silicon sampling (Marcel 2024); others could develop similar versions of models for other ends such as to code texts of a specific format or to contain a contextual knowledge of a specific academic literature.

Addressing these last two suggestions extends beyond the role of a single article or even a single research team. As social scientists take up these tasks, we encourage collaboration across research groups, institutions, and fields of study to generate robust discussion, inclusive participation, and thorough principles acceptable to a wide range of stakeholders and researchers.

In these pursuits, we urge continued focus on the ethical considerations that arise from LLM use in social science. While the particular methods proposed in this article do not directly impact human participants (limiting concerns in this case about direct efforts at misinformation, persuasion, etc.), we encourage researchers using these method to improve silicon sampling to evaluate the degree to which their use of LLMs compounds group-level stereotypes, prejudice, inequities, and discrimination. Further, more thought and discussion should be given to the environmental and energy impact of efforts to employ and improve LLMs and the implications of these processes for responsible use of generative AI (de Bolle 2024; Ren and Wierman 2024; Shoup 2024).

We believe that LLMs, properly employed, have almost unimaginable potential to transform social science. However, we fear that the confident language of LLMs can at times give researchers false confidence that an LLM is "getting it right," discouraging careful evaluation of LLM model choice and output. Our research experience working with LLMs consistently reinforces the centrality of human judgment and decision-making in AI research. Researcher expertise is needed to design theoretically-meaningful tests, adjust the behavior of the model, determine when to set aside a particular application, and evaluate the success of an LLM. None of this would be possible without human input, knowledge, and guidance. LLM research requires more, not less, of this type of involvement.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

Funding for this project came from the departments of Computer Science and Political Science at Brigham Young University.

ORCID iDs

Alex Lyman https://orcid.org/0009-0004-1210-2538
Bryce Hepner https://orcid.org/0000-0003-0062-5193

Lisa P. Argyle https://orcid.org/0000-0003-3109-2537
Ethan C. Busby https://orcid.org/0000-0002-8931-6348
Joshua R. Gubler https://orcid.org/0000-0003-1635-8210
David Wingate https://orcid.org/0000-0003-1850-6926

Data and Code Availability Statement

Replication materials including data and code are available at Lyman et al. (2025).

Supplemental Material

Supplemental materials for this article are available online.

Notes

1. If a well-calibrated model reports 70% confidence about something, it should be correct 70% of the time; if it reports 20% confidence, it should be correct 20% of the time, etc.
2. Such an effort would be challenging in any circumstance—it can be difficult if not impossible to distill the precise alignment steps for a given LLM, even in the open-source variants we use here, as full alignment procedures are rarely published.
3. The text for favorite color omitted the words "I am a," and included only "My favorite color is· · · "
4. While we prefer this method to traditional standard errors, we typically see only small differences between the permutation-based standard errors and classical standard errors. The conclusions reached from either approach are the same for the data we use in this article.

References

Achen, Christopher H. and Larry M. Bartels. 2016. *Democracy for Realists: Why Elections Do Not Produce Responsive Government*. Princeton, NJ: Princeton University Press.

Aher, Gati V., Rosa I. Arriaga, and Adam Tauman Kalai (2023) Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In *Proceedings of the 40th International Conference on Machine Learning*, ed. Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett. Vol. 202 of *Proceedings of Machine Learning Research* PMLR, pp. 337–371. https://proceedings.mlr.press/v202/aher23a.html.

Ahmed, Toufique, Premkumar Devanbu, Christoph Treude, and Michael Pradel (2024) “Can LLMs Replace Manual Annotation of Software Engineering Artifacts?”. https://arxiv.org/abs/2408.05534.

AI@Meta (2024) “Llama 3 Model Card”. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Anthropic (2024) “The Claude 3 Model Family: Opus, Sonnet, Haiku.” [Accessed 15-08-2024]. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Arditi, Andy, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda (2024) “Refusal in Language Models Is Mediated by a Single Direction.” *arXiv preprint arXiv:2406.11717*.

Argyle, Lisa P., Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen et al. 2023b. “Leveraging AI for Democratic Discourse: Chat Interventions Can Improve Online Political Conversations At Scale.” *Proceedings of the National Academy of Sciences* 120(41): e2311627120.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023a. “Out of One, Many: Using Language Models to Simulate Human Samples.” *Political Analysis* 3(3): 337–351. https://doi.org/10.1371/journal.pclm.0000429.

Argyle, Lisa P., Ethan C. Busby, Joshua R. Gubler, Alex Lyman, Justin Olcott, Jackson Pond, and David Wingate (2024) “Testing Theories of Political Persuasion Using Artificial Intelligence.” *Working Paper*.

Askell, Amanda, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones et al. (2021) “A General Language Assistant as a Laboratory for Alignment.” *CoRR* abs/2112.00861. https://arxiv.org/abs/2112.00861.

Bai, Yuntao, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova Das Sarma, Dawn Drain et al. (2022) “Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback”. https://arxiv.org/abs/2204.05862.

Bail, Christopher A.. 2024. “Can Generative AI Improve Social Science?.” *Proceedings of the National Academy of Sciences* 121(21): e2314021121.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623.

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson (2024) “Synthetic Replacements for Human Survey Data? The Perils of Large Language Models.” *Political Analysis*, pp. 1–16.

Blumer, Herbert. 1958. "Race Prejudice As a Sense of Group Position." *Pacific Sociological Review* 1(1): 3–7.

Boelaert, Julien, Samuel Coavoux, Etienne Ollion, Ivaylo D. Petev, and Patrick Präg (2024) "Machine Bias. Generative Large Language Models Have a View of Their Own." osf.io/preprints/socarxiv/r2pnb.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. "Semantics Derived Automatically From Language Corpora Contain Human-like Biases." *Science (New York, N.Y.)* 356(6334): 183–186.

Cheng, Myra, Esin Durmus, and Dan Jurafsky (2023) "Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models." https://arxiv.org/abs/2305.18189.

Chu, Junjie, Yugeng Liu, Ziqing Yang, Xinyue Shen, Michael Backes, and Yang Zhang (2024) "Comprehensive assessment of jailbreak attacks against llms." *arXiv preprint arXiv:2402.05668.*

Chung, Eun Yi and Joseph P. Romano. 2013. "Exact and Asymptotically Robust Permutation Tests." *The Annals of Statistics* 41(2): 484–507.

Consensus (2024) "Consensus: AI-powered Academic Search Engine." https://consensus.app/.

Copilot (2024) "Microsoft copilot." https://copilot.microsoft.com/.

Costello, Thomas H., Gordon Pennycook, and David G. Rand. 2024. "Durably Reducing Conspiracy Beliefs Through Dialogues with AI." *Science (New York, N.Y.)* 385(6714): eqdg1814. https://www.science.org/doi/abs/10.1126/science.adq1814.

de Bolle, Monica (2024) "AI's carbon footprint appears likely to be alarming." Peterson Institute for International Economics. https://www.piie.com/blogs/realtime-economics/2024/ais-carbon-footprint-appears-likely-be-alarming.

Demszky, Dorottya, Diyi Yang, David S. Yeager, Christopher J. Bryan, Margarett Clapper, Susannah Chandhok, Johannes C. Eichstaedt et al. 2023. "Using Large Language Models in Psychology." *Nature Reviews Psychology* 2(11): 688–701. https://doi.org/10.1038/s44159-023-00241-5.

Dillion, Danica, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. "Can AI Language Models Replace Human Participants?." *Trends in Cognitive Sciences* 27(7): 597–600. https://doi.org/10.1016/j.tics.2023.04.008.

Ding, Peng, Avi Feller, and Luke Miratrix. 2016. "Randomization Inference for Treatment Effect Variation." *Journal of the Royal Statistical Society Statitiscs Methodology, Series B* 78: 655–671.

Dubey, Abhimanyu, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur et al. (2024) "The Llama 3 Herd of Models." https://arxiv.org/abs/2407.21783.

Edgell, Penny, Joseph Gerteis, and Douglas Hartmann. 2006. "Atheists As 'Other': Moral Boundaries and Cultural Membership in American Society." *Social Forces* 71: 211–234.

Elicit (2024) "Elicit: The AI Research Assistant." https://elicit.com.

Ellemers, Naomi. 2018. "Gender Stereotypes." *Annual Review of Psychology* 69: 275–298.

Gerber, Alan S. and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W. W. Norton.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. "ChatGPT Outperforms Crowd Workers for Text-annotation Tasks." *Proceedings of the National Academy of Sciences* 120(30): e2305016120.

Goldstein, Josh A. and Girish Sastry (2023) "The Coming Age of AI-powered Propaganda. How to Defend Against Supercharged Disinformation." *Foreign Affairs*, p. 7.

Google Gemini Team (2024) "Gemini: Advanced Conversational AI Models." https://www.deepmind.com/gemini. Accessed: 2024-08-15.

Hackenburg, Kobi and Helen Margetts. 2024. "Evaluating the Persuasive Influence of Political Microtargeting with Large Language Models." *Proceedings of the National Academy of Sciences* 121(24): e2403116121.

He, Zeyu, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-HaoKenneth Huang (2024) If in a Crowdsourced Data Annotation Pipeline, a GPT-4. CHI '24 New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3613904.3642834.

Heseltine, Michael and Bernhard Clemm vin Hohenberg (2024) "Large language models as a substitute for human experts in annotating political text." *Research and Politics*.

Hewitt, Luke, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer (2024) "Predicting Results of Social Science Experiments Using Large Language Models." https://samim.io/dl/Predicting%20results%20of%20social%20science%20experiments%20using%20large%20language%20models.pdf.

Horton, John J. (2023) Large language models as simulated economic agents: What can we learn from homo silicus? Technical report National Bureau of Economic Research.

Hutchings, Vincent L. and Nicholas A. Valentino. 2004. "The Centrality of Race in American Politics." *Annual Review of Political Science* 7: 383–408.

Ji, Jiaming, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Chi Zhang et al. (2023) "BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset." https://arxiv.org/abs/2307.04657.

Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand et al. (2023) "Mistral 7B." https://arxiv.org/abs/2310.06825.

Jiang, Albert Q, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot et al. (2024) "Mixtral of Experts." https://arxiv.org/abs/2401.04088.

Kim, Junsol and Byungkyu Lee. 2024. "AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction." https://arxiv.org/abs/2305.09620.

Kinder, Donald R. and Cindy D. Kam. 2009. *Us Against them : Ethnocentric Foundations of American Opinion*. Chicago: University of Chicago Press.

Kirk, Robert, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu (2024) Understanding the Effects of RLHF on LLM Generalisation and Diversity. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=PXD3FAVHJT.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics* 133(1): 237–293. https://doi.org/10.1093/qje/qjx032.

Kozlowski, Austin C., Hyunku Kwon, and James A. Evans (2024) "In Silico Sociology: Forecasting COVID-19 Polarization with Large Language Models." https://arxiv.org/abs/2407.11190.

Kreps, Sarah, R. Miles McCain, and Miles Brundage. 2022. "All the News That's Fit to Fabricate: AI-Generated Text As a Tool of Media Misinformation." *Journal of Experimental Political Science* 9(1): 104–117.

Lake, Thom, Eunsol Choi, and Greg Durrett (2024) "From Distributional to Overton Pluralism: Investigating Large Language Model Alignment." https://arxiv.org/abs/2406.17692.

Lee, Sanguk, Tai-Quan Peng, Matthew H. Goldberg, Seth A. Rosenthal, John E. Kotcher, Edward W. Maibach, and Anthony Leiserowitz. 2024. "Can Large Language Models Estimate Public Opinion About Global Warming? An Empirical Assessment of Algorithmic Fidelity and Bias." *PLOS Climate* 3(8): 1–14. https://doi.org/10.1371/journal.pclm.0000429.

Li, Victoria R., Yida Chen, and Naomi Saphra (2024) "ChatGPT Doesn't Trust Chargers Fans: Guardrail Sensitivity in Context." https://arxiv.org/abs/2407.06866.

Liu, Andy, Mona Diab, and Daniel Fried (2024) "Evaluating Large Language Model Biases in Persona-Steered Generation." https://arxiv.org/abs/2405.20253.

Lyman, Alex, Bryce Hepner, Lisa P. Argyle, Ethan C. Busby, Joshua R. Gubler, and David Wingate (2025) "Replication Materials for Balancing Large Language Model Alignment and Algorithmic Fidelity in Social Science Research".

GitHub. https://github.com/AlexMLyman/Replication-Materials-for-Balancing-Large-Language-Model-Alignment-and-Algorithmic-Fidelity.

Marcel, Binz (2024) "Llama-3.1-Centaur-70B." https://huggingface.co/marcelbinz/Llama-3.1-Centaur-70B-adapter.

Mellon, Jonathan, Jack Bailey, Ralph Scott, James Breckwoldt, Marta Miori, and Phillip Schmedeman (2024) "Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale." *Research and Politics*.

Mize, Trenton D.. 2016. "Sexual Orientation in the Labor Market." *American Sociological Review* 81: 1132–1160.

Moore, Jared, Tanvi Deshpande, and Diyi Yang (2024) "Are Large Language Models Consistent over Value-laden Questions?" https://arxiv.org/abs/2407.02996.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in An Algorithm Used to Manage the Health of Populations." *Science (New York, N.Y.)* 366(6464): 447–453.

OpenAI (2023) "ChatGPT: A Conversational AI Model." https://www.openai.com/chatgpt. Accessed: 2024-08-15.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida et al. (2024) "GPT-4 Technical Report." https://arxiv.org/abs/2303.08774.

Pachot, Arnault and Thierry Petit (2024) "Can Large Language Models Accurately Predict Public Opinion? A Review".

Palmer, Alexis and Arthur Spirling. 2023. "Large Language Models Can Argue in Convincing Ways About Politics, But Humans Dislike AI Authors: Implications for Governance." *Political Science* 75(3): 281–291.

Panch, Trishan, Heather Mattie, and Rifat Atun. 2019. "Artificial Intelligence and Algorithmic Bias: Implications for Health Systems." *Journal of Global Health* 9(2): 010318.

Park, Joon Sung, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein (2024) "Generative Agent Simulations of 1,000 People." https://arxiv.org/abs/2411.10109.

Parrish, Alicia, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut et al. (2022) BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, ed. Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 2086–2105. https://aclanthology.org/2022.findings-acl.165.

Phillips, Nolan E., Brian L. Levy, Ryan J. Sampson, Mario L. Small, and Ryan Q. Wang. 2021. "The Social Integration of American Cities: Network Measures of Connectedness Based on Everyday Mobility Across Neighborhoods." *Sociological Methods & Research* 50: 1189–1214.

Qu, Yao and Jue Wang. 2024. "Performance and Biases of Large Language Models in Public Opinion Simulation." *Academy of Management Proceedings* 2024(1): 10298. https://doi.org/10.5465/AMPROC.2024.10298abstract.

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019) Language Models are Unsupervised Multitask Learners. https://api.semanticscholar.org/CorpusID:160025533.

Rafailov, Rafael, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn (2023) Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=HPuSIXJaa9.

Ren, Shaolei and Adam Wierman (2024) "The Uneven Distribution of AI's Environmental Impacts." Harvard Business Review. https://hbr.org/2024/07/the-uneven-distribution-of-ais-environmental-impacts.

Risman, Barbara. 2004. "Gender As a Social Structure." *Gender & Society* 18: 429–450.

Robinson, Joshua, Christopher Michael Rytting, and David Wingate (2023) "Leveraging Large Language Models for Multiple Choice Question Answering." https://arxiv.org/abs/2210.12353.

Rothschild, Jacob E., Adam J. Howat, Richard M. Shafranek, and Ethan C. Busby. 2019. "Pigeonholing Partisans: Stereotypes of Party Supporters and Partisan Polarization." *Political Behavior* 41: 423–443.

Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto (2023) Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*. ICML'23 JMLR.org.

Sherif, Muzafer, O. J. Harvey, B. Jack White, William R. Hood, and Carolyn W. Sherif. 1961. *Intergroup Conflict and Cooperation: The Robbers Cave Experiment*. Norman, OK: Institute of Group Relations.

Shoup, Ella (2024) "AI and ESG: Understanding the Environmental Impact of AI and LLMs." Holistic AI. https://www.holisticai.com/blog/environmental-impact-ai-llms.

Srivastava, Aarohi, et al. (2023) "Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models." *Transactions on Machine Learning Research*. https://openreview.net/forum?id=uyTL5Bvosj.

Suzgun, Mirac, Tayfun Gur, Federico Bianchi, Daniel E. Ho, Thomas Icard, Dan Jurafsky, and James Zou (2024) "Belief in the Machine: Investigating

Epistemological Blind Spots of Language Models." https://arxiv.org/abs/2410.21195.

Team, Gemma (2024) "Gemma." https://www.kaggle.com/m/3301.

Tessler, Michael Henry, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. 2024. "AI can help humans find common ground in democratic deliberation." *Science (New York, N.Y.)* 386(6719): eadq2852.

Thébaud, Sarah, Sabino Kornrich, and Leah Ruppanner. 2021. "Good Housekeeping, Great Expectations: Gender and Housework Norms." *Sociological Methods & Research* 50: 1189–1214.

Velez, Yamil and Patrick Liu (2024) "Confronting Core Issues: A Critical Assessment of Attitude Polarization Using Tailored Experiments." *American Political Science Review*.

Wei, Alexander, Nika Haghtalab, and Jacob Steinhardt. 2024. "Jailbroken: How Does Llm Safety Training Fail?." *Advances in Neural Information Processing Systems* 36: 80079–80100.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama et al. (2022) "Emergent Abilities of Large Language Models." https://arxiv.org/abs/2206.07682.

Whitehead, Andrew L., Samuel L. Perry, and Joseph O. Baker. 2018. "Make America Christian Again: Christian Nationalism and Voting for Donald Trump in the 2016 Presidential Election." *Sociology of Religion* 79: 147–171.

Xu, Zihao, Yi Liu, Gelei Deng, Yuekang Li, and Stjepan Picek (2024) "LLM Jailbreak Attack Versus Defense Techniques—A Comprehensive Study." *arXiv preprint arXiv:2402.13457*.

Zhang, Shengyu, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li et al. (2024) "Instruction Tuning for Large Language Models: A Survey." https://arxiv.org/abs/2308.10792.

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin et al. (2024) Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS '23 Red Hook, NY, USA: Curran Associates Inc.

Zhu, Chiwei, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao (2023) On the Calibration of Large Language Models and Alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, ed. Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 9778–9795. https://aclanthology.org/2023.findings-emnlp.654.

Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang (2024) “Can Large Language Models Transform Computational Social Science?”. https://arxiv.org/abs/2305.03514.

Author Biographies

Alex Lyman is a doctoral researcher in the computer science department at Brigham Young University. His research focuses on real-world applications of large language models.

Bryce Hepner is a graduate researcher in the computer science department at Brigham Young University. His research focuses on meaningful use and understanding of large language models.

Lisa P. Argyle is an assistant professor of political science at Brigham Young University. She studies political conversation and partisanship in the United States using large language models and other computational methods.

Ethan C. Busby is an assistant professor of political science at Brigham Young University, where he studies political extremism and how large language models can improve social science.

Joshua R. Gubler is an associate professor of political science at Brigham Young University, where he is also Program Coordinator for the Middle Eastern Studies/ Arabic program.

David Wingate is an associate professor of computer science at Brigham Young University, where he works understanding the capabilities and limitations of large language models.