AI Simulations of Audience Attitudes and Policy Preferences: “Silicon Sampling” Guidance for Communications Practitioners

John Wihbey and Samantha D’Alonzo[1]

Working Paper

Last updated: October 23, 2025

AIMES Lab
Northeastern University

AI-Media
Strategies Lab

[1] John Wihbey, Associate Professor [j.wihbey@northeastern.edu]; Samantha D’Alonzo, Doctoral Researcher

Title:

AI Simulations of Audience Attitudes and Policy Preferences: “Silicon Sampling” Guidance for Communications Practitioners

Abstract

This working paper reviews and translates a broad array of academic research on "silicon sampling"—using Large Language Models (LLMs) to simulate public opinion—and offers guidance for practitioners, particularly those in communications and media industries, conducting message testing and exploratory audience-feedback research. Findings show LLMs are effective complements for preliminary tasks like refining surveys but are generally not reliable substitutes for human respondents, especially in policy settings. The models struggle to capture nuanced opinions and often stereotype groups due to training data bias and internal safety filters. Therefore, the most prudent approach is a hybrid pipeline that uses AI to improve research design while maintaining human samples as the gold standard for data. As the technology evolves, practitioners must remain vigilant about these core limitations. Responsible deployment requires transparency and robust validation of AI findings against human benchmarks. Based on the translational literature review we perform here, we offer a decision framework that can guide research integrity while leveraging the benefits of AI.

Note: We would like to thank Christoph Riedl of Northeastern University; Kaicheng Yang of Binghamton University (SUNY); Christopher Bail of Duke University and Chenyan Jia and Min Ge of Northeastern University for their feedback. We acknowledge support from the John S. and James L. Knight Foundation, as part of Northeastern’s Institute for Information, the Internet, and Democracy.

Introduction

As the capabilities of large language models (LLMs), such as OpenAI's ChatGPT, Google's Gemini, and Anthropic's Claude, have improved, many professionals have begun experimenting with leveraging generative AI technologies as a test bed for scoping potential audience/public response. Public relations, advocacy, and communications practitioners have begun experimenting with using LLMs throughout the survey pipeline, using them for everything from synthetic focus groups and pilots to pre-testing framings and simulation machines of the general public and its policy preferences. Figuring out how to appropriately apply LLMs in the survey pipeline can be a daunting task, especially because the technological landscape is evolving so rapidly and LLMs can complete most survey-related tasks cheaply and quickly. In this paper, we argue that just because LLMs can complete most survey-related tasks does not mean they should. In fact, we explicitly caution against using LLMs as substitutes for human respondents during survey administration, or "silicon sampling," due to a number of limitations in current models. While pure substitution may not be appropriate, we do highlight other steps in the survey pipeline where LLMs can be used in exploratory ways as complements, so long as practitioners consider a number of factors that can influence LLM output. We also walk through which subdomains under the umbrella of "public opinion" that LLMs may be best suited for.

In this paper, we first walk through the classical survey pipeline, from design to administration to analysis, and explore what types of tasks LLMs are well-suited for within that pipeline. We then walk through how LLM design decisions and user prompting impact LLM outputs. Lastly, we provide a set of questions, visualized in Figure 1, that practitioners can use to guide their design process and evaluate their own work, as well as others' work in this space. Given that this is an emerging area of practice, we seek to point to strategies that have analytical validity and may have efficacy, as well to provide cautionary guidance where appropriate.

To develop these guidelines, we reviewed findings from about 30 academic papers, particularly focusing on well-cited pieces of literature in the space, which are summarized in

Table 2. We searched databases such as Google Scholar using the keywords "silicon sampling" and related terms. We prioritized recent studies that captured newer trends among LLM development and that explicitly performed an experiment to evaluate LLMs as human substitutes. We supplemented these articles with highly-cited, more general works about LLMs, focusing, when possible, on literature directly published by AI companies. We also relied on additional, more general papers throughout this review, but only papers that conducted an experiment explicitly evaluating LLMs in contexts related to silicon sampling were included in **Table 2**.

While this paper provides general guidance for practitioners, it has notable limitations. First, LLMs continue to develop and improve across a variety of areas and benchmarks, so our literature review and recommendations represent a point in time. Second, much of our guidance presumes users will not have large engineering teams to tune and test highly customized models, meaning that we are focused mostly on using the models "as is," without significant modifications.

Where (or Where Not) to Incorporate LLMs in a Survey Workflow

LLMs as Survey Designers

LLMs can serve as a helpful starting point for practitioners when designing surveys across domains. At their most basic, LLMs can be used to edit human-written survey questions, either by proofreading initial drafts of questions or providing guidance on wording. Practitioners can simply prompt LLMs with drafts of questions and a description of the goal of the survey and get feedback in a conversational way [1]. LLMs as editors have been helpful in other domains, such as scientific writing [2], and can help survey designers catch errors or tighten up ambiguous question or answer phrasing. Successes in non-survey domains can be extended to this use case, as certain models have shown an ability to identify key components of survey questions and provide constructive feedback on their phrasing [3]. Anecdotally LLM-generated feedback on survey vagueness has been consistent with accepted academic practices, and LLMs may be able to identify ambiguous or leading answer choices [4].

When prompting LLMs for feedback, practitioners should be aware of known biases and tendencies within AI models. LLMs have been known to display sycophantic behaviors, like excessively agreeing with or flattering users [5]. To get the maximum value out of LLMs at this stage, practitioners should actively try to use objective statements in prompts or explicitly prompt for critiques by asking things like “Why might this question be difficult for survey takers to understand?” Practitioners can also ask questions like “What is the point of this question?” to explore if the question is effectively conveying what the practitioner hopes.

Although an LLM’s ability to write effective survey questions has not been studied systematically, there are a number of ways practitioners can leverage LLMs during the question writing process, so long as they sanity check the outputs. For very early stage exploratory tasks in consumer product research, such as A/B message testing, LLMs can be used to get signals on a potential product or service. LLMs seem to be able to give at least some basic insights on the magnitude and direction of human preferences in consumer product settings [6]. LLMs are also able to summarize information that is likely established in their pretraining dataset [7], [8], so practitioners could prompt LLMs to summarize existing digitized survey data if they believe the survey is in the LLM’s training dataset. Guidance for thinking through this can be found in later sections. These initial LLM-provided signals or summaries can be good starting points for resource-constrained practitioners when deciding on which areas to focus on or questions to prioritize.

LLMs may also be helpful in adapting questions written for one cultural context or language to another [9], and LLMs in this role have shown promise. One study found that GPT-4 was able to generate questions from English to German as well as conventional translation methods [10]. Another study found that human survey takers found LLM-adopted questions across cultural contexts were slightly clearer and less biased than traditional questions [9]. More generally, research has shown that LLMs can highlight phrases that may be interpreted differently by different cultures [3]. Preemptively identifying confusing phrases can save practitioners valuable time and money during later stages of the survey process.

LLMs as Pilot-Testers

Related to using LLMs as survey designers, practitioners can also use LLMs as exploratory pilot-testers to finalize their surveys and save time and money in early-stage iteration. The effectiveness of this may depend on what domain practitioners are interested in. For reasons that will be outlined in the section “LLMs as Survey Takers,” pilot-testing will likely be more effective (although not guaranteed and still requiring human validation) on non-controversial, non-political topics where practitioners can count on some priors (reasonably well formed beliefs) already within the distribution of public opinion.

To conduct meaningful pilot-tests, practitioners can instruct LLMs to take on different personas with different attributes and then practitioners can administer their survey to the LLM-personas. There are, to our knowledge, no well-defined guidelines yet on how to decide how many, or which, personas to include or how to prompt for them. Discussions in later sections will help practitioners understand how (or how not) to conduct persona-based pilot-testing; but at a high level, practitioners should keep in mind that many models struggle to accurately represent specific groups. Practitioners should also keep in mind that if they administer pilot-testing surveys to an LLM with no additional demographic details, the LLM will give answers that reflect its default settings, which in many cases are biased toward, for example, certain US viewpoints.

In general, this process can generate a cost- and time-efficient initial data set that practitioners can analyze to assess their initial hypotheses before running full-scale human surveys [11]. Specifically, initial data could help practitioners understand which sub-populations to focus on when running the survey for humans [4] or which questions have low consensus and thus should be prioritized. In keeping with general survey best practices, practitioners should ensure their pilot-test questions meet a minimum standard of accepted quality (e.g., they do not have middle options, “I don’t know” as an answer choice, or asymmetrically worded answer choices). This is especially important when pilot-testing with LLMs because LLMs do not answer biased questions in the same way humans do [12] and so pilot-tested results from biased questions can be misleading for practitioners. The intention at this stage is not to generate a finalized dataset, but rather simply to refine potentially confusing survey questions and prioritize which questions to ask and which audiences to target.

LLMs as Survey Takers

Despite these many caveats about pilot-testing, it may be tempting to use LLMs as *full-on substitutes* for humans in survey research, especially as representative surveys become harder to conduct. While some research has highlighted the potential of LLMs in this role [13], the research literature to date suggests caution in terms of using LLMs as pure human substitutes, *especially* for topics where practitioners expect there to be high variance among human answers (divisive topics), for political topics, or for topics that involve multiple levels of thinking.

The evidence suggests that any research that aims to use LLMs as human substitutes should leverage expert validation before relying on LLM-simulated opinions, especially for making consequential decisions involving public communications, product rollouts, policy framing, or other real-world media or messaging tasks. Although similar in appearance, pilot-testing should not be mistaken for real audience data and cannot support firm conclusions.

Inaccurate Distribution of Answers

LLMs are unreliable as human substitutes because many models consistently fail to accurately capture the distribution of human responses across certain topics [7], [14], [15], [16], [17], [18], generally presenting a narrower range of opinions than can be found in real data [6], [19], [20], [21]. Many models tend to collapse opinion differences across divisive topics, like race or religion [21], but do not collapse opinion differences for common current political topics (e.g., gun rights, immigration, transgender rights, abortion, etc.). In fact, various research papers have actually shown that certain models tend to overemphasize the ideological differences for such topics and present political groups as more polarized than they are in practice [7], [21], [22]. This somewhat inconsistent distribution collapse or expansion makes it difficult to rely on LLMs for capturing the full range of public opinion across topics and thus limits the usefulness of LLMS as human substitutes.

When using LLMs for pilot-testing, practitioners should avoid political and divisive topics, where these distribution problems are most pronounced. For all pilot-testing, validate LLM outputs against non-LLM sources. When creating personas, be particularly cautious about including political attributes, as some research shows this dramatically alters LLM responses in unpredictable ways [21].

Stereotyping Certain Groups

Not only do LLMs fail to capture the full range of human opinion, some models also tend to perform particularly poorly, either exaggerating differences for the group or failing to capture within-group variation, for Independents [17], non-Hispanic Black Americans [23], [24], conservatives, nonbinary individuals, and people of Middle-Eastern or Hispanic background [24]. Other research has found that when prompted to consider a certain country's perspective, responses from tested models do shift but in a way that may reflect cultural stereotypes rather than a nuanced understanding of the population of interest [25]. Additionally, when querying in non-English languages, answers from certain models do not necessarily reflect the opinions of the non-English language speakers [25]. Some research suggests that certain models cannot predict the opinions of any demographic group consistently [6]. If this is true, it makes using LLMs to gauge opinions of different sociodemographic groups particularly unreliable, as we would not even be able to anticipate where errors may arise. As previously mentioned, practitioners trying to build personas for early stage-pilot testing should keep this stereotyping in mind and think of the early-stage data as providing directional signals at best. Practitioners should always sanity check the outputs with non-LLM data.

Non-Human Cognitive Patterns

While we may anthropomorphize LLMs, they have not been shown to actually mimic human behavior or reasoning. Tested LLMs don't make the same errors as humans - they fail to mimic meta-cognitive signals like uncertainty, fatigue, or emotional response during certain tasks [26] and avoid cognitive errors that are expected of humans on certain psychological tasks [27]. Some tested models are more goal-directed, showing less exploratory behaviors relative to humans when browsing websites [26]. They also tend to be hyper-accurate, giving perfect estimates for obscure quantities [7], [28], such as the melting temperature of aluminum, despite the fact that humans are unlikely to know such information. These non-human attributes mean LLMs cannot be relied on to capture public opinion in domains where practitioners expect there should be public uncertainty, like science, or where human reasoning processes are of interest to the practitioner.

Augmenting LLMs & Future Techniques to Monitor

Although there is not yet a clear consensus, two emerging techniques, fine-tuning and Retrieval-Augmented-Generation (RAG), may make LLMs more effective as human substitutes in the future. Fine-tuning may allow practitioners to extend survey findings to new subpopulations or domains, while RAG could increase the reliability of LLMs by grounding responses in a specific corpus. Both techniques require additional, external datasets, so they may only be helpful in situations where the workflow bottleneck is *not* collecting human samples but rather extending or analyzing existing human survey data.

Fine-tuning involves taking a trained model and using new, labeled data to calibrate the model further for a specified task[1]. Typically, fine-tuning involves actually changing weights on open-sourced models. OpenAI, which has primarily closed-source models, does offer an abstracted ability to fine-tune models. RAG involves supplementing LLMs' knowledge base with internet access of a new external database, which does not necessarily need to be labeled. When prompted, LLMs will supplement the provided prompt

[1] Fine-tuning is *different* from few-shot prompting - providing examples of desired "call" and 'responses' to LLMs in their context windows, which will be discussed more later.

with information gained from a search of the provided external database. In general, this process may help models overcome limitations for knowledge-intensive tasks [8].

Among the limited published papers on fine-tuning so far, outcomes have been generally positive, with a few negative studies mixed in. One study found that fine-tuning certain models with additional public opinion survey data help the models better generalize opinions for unseen populations, question waves, and topics [29]. Another found that fine-tuning LLMs on behavioral data substantially improved their ability to generate human-like actions in simulated online shopping tasks [30]. There have been even fewer studies, to date, about using RAG techniques to supplement silicon sampling. One study did find that incorporating RAG improved the accuracy of simulated political opinions [8]. Another study, however, found that significant differences between human and tested model performance on economic games persisted despite fine-tuning and RAG [31]. Findings for these techniques are mixed, and both techniques are more resource-intensive than out-of-the-box use, making them impractical for many practitioners. We recommend waiting for more research to come out on fine-tuning and RAG for silicon sampling purposes, before using either technique for extending or analyzing human survey data.

LLMs as Survey Interpreters

LLMs may be helpful when analyzing human survey data after collection, at least for certain types of data. One potential use case is having LLMs annotate or classify existing open-ended data to simplify later analysis tasks for humans. So far, there have been mixed results on this front. Some studies have shown LLMs can do such tasks accurately [32], while others have shown models struggle with text-processing tasks [33] or perform barely better than random guessing at classification style tasks [34]. Practitioners who use LLMs in this way should validate the outputs. Practitioners should be hesitant using LLMs to identify or explain causal relationships in data through prompting methods alone have yielded mixed results [49] and unexpected failure modes [50].

In general, practitioners should be cautious about privacy considerations when sharing human survey data with LLMs, especially when using interfaces such as ChatGPT and especially when survey data contains personally identifiable information (PII) or sensitive responses. If practitioners have serious privacy concerns, they will need to deploy open-source models on their personal servers, although this requires significant additional resources and may not be feasible for many practitioners.

How (or How Not) to Incorporate LLMs in a Survey Workflow

After practitioners have decided where they would like to insert an LLM into their survey workflow, they then have to figure out exactly how to go about doing that. There are three main open questions at this stage (1) Which model to use? (2) How to adapt it and prompt it such that it can complete the task? and (3) How to interpret or validate the outputs? We explore these questions below.

Model Selection

It may be tempting to think of each model version, from each AI company, as relatively interchangeable – settling for the first model you try, or just using the one with the best name recognition. In reality, different design decisions mean that LLMs perform very differently across discrete tasks. Therefore, we recommend understanding key components of model design and how different designs impact downstream LLM outputs. With this basic understanding, practitioners should have the starting knowledge and vocabulary to navigate effectively the evolving LLM landscape. Even if practitioners do not have flexibility in which model to use, due to cost considerations or other constraints, a basic understanding of model decisions will still give practitioners a competitive edge in interpreting and utilizing LLM outputs.

The following section discusses different model attributes to consider when choosing a model. **Figure 1** summarizes the design characteristics of models made available by OpenAI, Anthropic, Meta, Google, and Mistral AI. **Figure 1** also includes references to academic studies that referenced each model. **Table 2** provides high-level summaries of each academic paper. Due to the rapid evolution of the space and the relative lag of academic publishing, models in **Figure 1** are no longer state-of-the-art. We still include

Figure 1 and **Table 2** to help practitioners understand what types of experiments have been conducted. With that being said, practitioners should be cautious not to overextend the cited experiments. Various research papers have suggested that results can vary across combinations of topic, model, and model-checkpoint. Some research has even suggested that the same model can produce different results at different points in time [7]. Therefore, as a general rule of thumb, it is always recommended to start the model selection process by performing simple benchmarking and project-specific tasks [15].

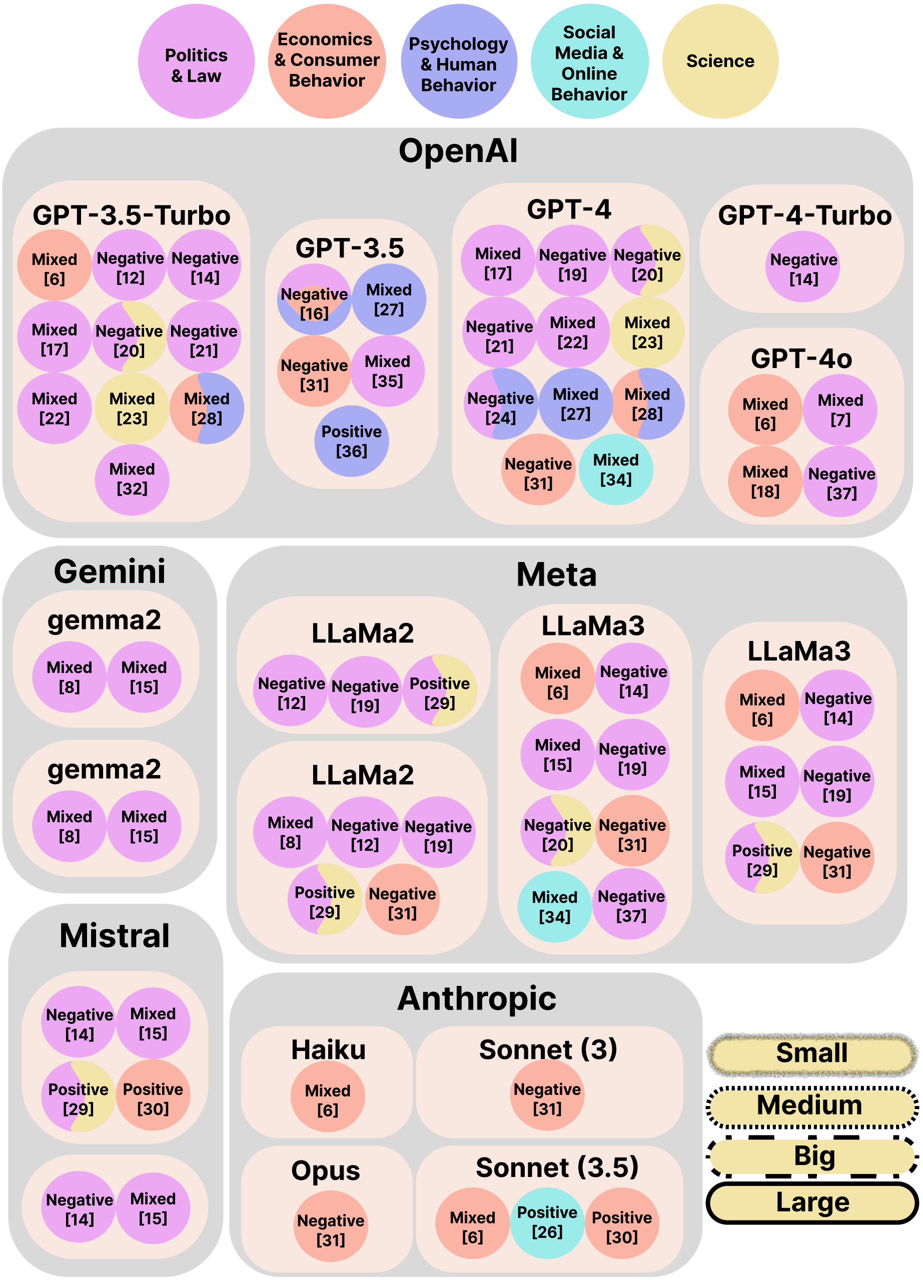


Figure 1: A visual summary of academic research for a select number of models. Each circle within the model boundary represents a cited paper that included that model. The color of each circle represents the domain — Politics & Law, Economics & Consumer Behavior, Psychology & Human Behavior, Social Media & Online Behavior, Science — of the paper; papers that covered multiple domains are represented with multiple colors. The borders of each model represent the estimated model size. Size classifications are based on estimated numbers of parameters and break down as follows: Small ($\leq 10B$), Medium ($> 10B, \leq 100B$), Big ($> 100B, \leq 1T$), and Large ($> 1T$).

Model Size (Parameters)

Model size refers to the number of “parameters” or “weights” within a given model. In general, larger models – those with more than a few billion parameters – generally tend to be more accurate and versatile than smaller models [38], as they may be able to retain more knowledge from training data than models with fewer parameters (and therefore perform better on tasks where the factual knowledge needed is in their training data set [39].) Although it is not recommended to use either large or small models as human substitutes for survey responses – it is strongly recommended to avoid using small (< 10B parameters) models for this task and certain complementary tasks. Small models struggle to follow lengthy, multi-step instructions [40], fail to illuminate opinion differences between sociodemographic groups [33], and, when presented with multiple-choice survey style questions, tend to select answers labeled “A” [19]. Although larger language models offer benefits, such models also tend to be more costly, energy-intensive, and slower to run.

Model Alignment

Model alignment refers to the methods, such as Reinforcement Learning from Human Feedback (RLHF) [41] or Constitutional AI [42], undertaken by companies to ensure that models align with human values, goals, and preferences. While the specific alignment methods used vary across companies and models, a general goal is to produce harmless LLMs. While a noble goal for broader LLM use, this harmless alignment may have undesirable consequences for silicon sampling – such as refusal to perform a certain task or inhabit certain viewpoints [15]. Additionally, aligned LLMs have exhibited a tendency to provide innocuous answers for certain sensitive topics, even when being prompted with demographic information that, in reality, should alter the expressed opinion [17], and aligned models may be less likely to produce negative responses compared to base models [15], [18].

As previously alluded to, a crucial takeaway is that model alignment impacts the distribution of LLM outputs, typically narrowing or altering the range of opinions expressed in a way that does not accurately reflect general human opinion. Some research has hypothesized that aligned models may better represent training data or

scientists than the general public. As a result, models may present the illusion of consensus where, in reality, public opinion is more nuanced, such as on issues of climate change [23] or misrepresent public opinions on topics such as recent Supreme Court rulings [37].

As previously mentioned, in order to maximize the efficacy of using LLMs as human surrogates, it is important to understand, at least at a high level, what each AI company focuses on during alignment and how that may impact downstream tasks. However, the exact alignment procedures used for each model are not publicized. Below, in **Table 1**, we provide starting references for each major AI Company's stated alignment processes.

AI Company	**Alignment Procedures**
OpenAI (ChatGPT)	InstructGPT, Reinforcement Learning from Human Feedback
Anthropic (Claude)	Constitutional AI, Reinforcement Learning from AI Feedback
Google DeepMind (Gemini)	Reinforcement Learning from Human Feedback, Recursive Reward Modeling, Debate
Meta (LLaMa)	Supervised Fine-Tuning, Rejection Samplin Direct Preference Optimization,

Table 1: References, where possible, to the stated alignment procedures of four major companies

Training Data

Practitioners should try to consider the scope - particularly the temporal scope[2] - of the training dataset used by models. Knowing the model's cutoff date is essential for

[2] After the model has been trained, there is a cutoff date such that no new information is included in the models pretraining data.

understanding which events the model has prior data on. Research has shown that LLMs do a poor job extending the knowledge in their training, especially for policy issues where models tend to overgeneralize their understanding of ideological differences to new policy issues outside their temporal scope [22].

In addition to thinking through what is temporally relevant, practitioners should think about what *types* of data may be present and how that will impact the LLMs ability to provide accurate answers on certain topics. Although no company has made its exact training datasets available, general knowledge of training datasets and ongoing research has highlighted a few trends. Whether it is because of the training data or some other model design specification, many language models have been shown to have default settings that do not represent the general public. The default settings of many models appear to reflect publics in the United States [7] or European or South America countries [25]. Some iterations of ChatGPT in particular have shown a pro-environmental, left-libertarianism default [35]. Many of the stereotypes and misrepresentations discussed previously may be because of skewed training data, so practitioners should consider whether or not their use case is likely to fall within the scope of an LLM's training corpora before relying on LLMs for assistance.

Reasoning vs. Non-Reasoning Models

Relatively new to the LLM scene are "reasoning models" or models that are specialized to answer questions that require complex, multi-step reasoning, such as math proofs, complex decision making, and chain-of-thought reasoning [43]. Reasoning models are typically more costly than non-reasoning models, and there has been some research indicating that "reasoning" models display limited ability to execute generalizable reasoning and experience accuracy collapse once problems get too complex [44]. As previously mentioned, research has shown that LLM reasoning is not akin to human reasoning. Because current silicon sampling involves direct prompting for opinion, rather than complex problem-solving, non-reasoning models are likely sufficient for most practitioners' purposes.

Model Specifications

Even after practitioners have selected a model, there are still decisions to make - the most important being how to prompt the model efficiently. In general, with many knobs to turn and a rapidly evolving landscape, we recommend that practitioners keep detailed records of what model, model version, model temperature, and prompt practitioners used and on what date they accessed the model.

Prompt Engineering

Various research [6], [21], [31], [33] suggests that small variations in prompt wording, as well as choosing to include or exclude certain demographic information, can lead to large, sometimes unexpected variations in LLM output. With that in mind, perhaps the best advice is to experiment with different prompts to ensure that results are as robust as possible. One concrete thing to keep in mind is the size of the context window[3]. Practitioners should ensure that relevant instructions fall inside the context window of the model they are using.

Another decision practitioners must make when it comes to prompting is whether they will be doing "zero-shot" or "few-shot" prompting. In zero-shot prompting, no examples are provided to the LLM. In few-shot prompting, practitioners give LLMs examples of queries and expected responses. While few-shot prompting may be helpful to help models understand appropriate formatting for examples or learn more about a given task, some research has shown that LLMs tend to over-attend to examples provided [31], meaning few-shot prompting may have undesired consequences and lead to a narrowing of expected outputs around provided examples. Perhaps the best advice is to experiment with different prompt phrasings to ensure prompts are robust and experiment with including or excluding certain information to understand how it may influence the outcomes.

[3] The number of tokens that a model can accept as its input prompt

Model Temperature

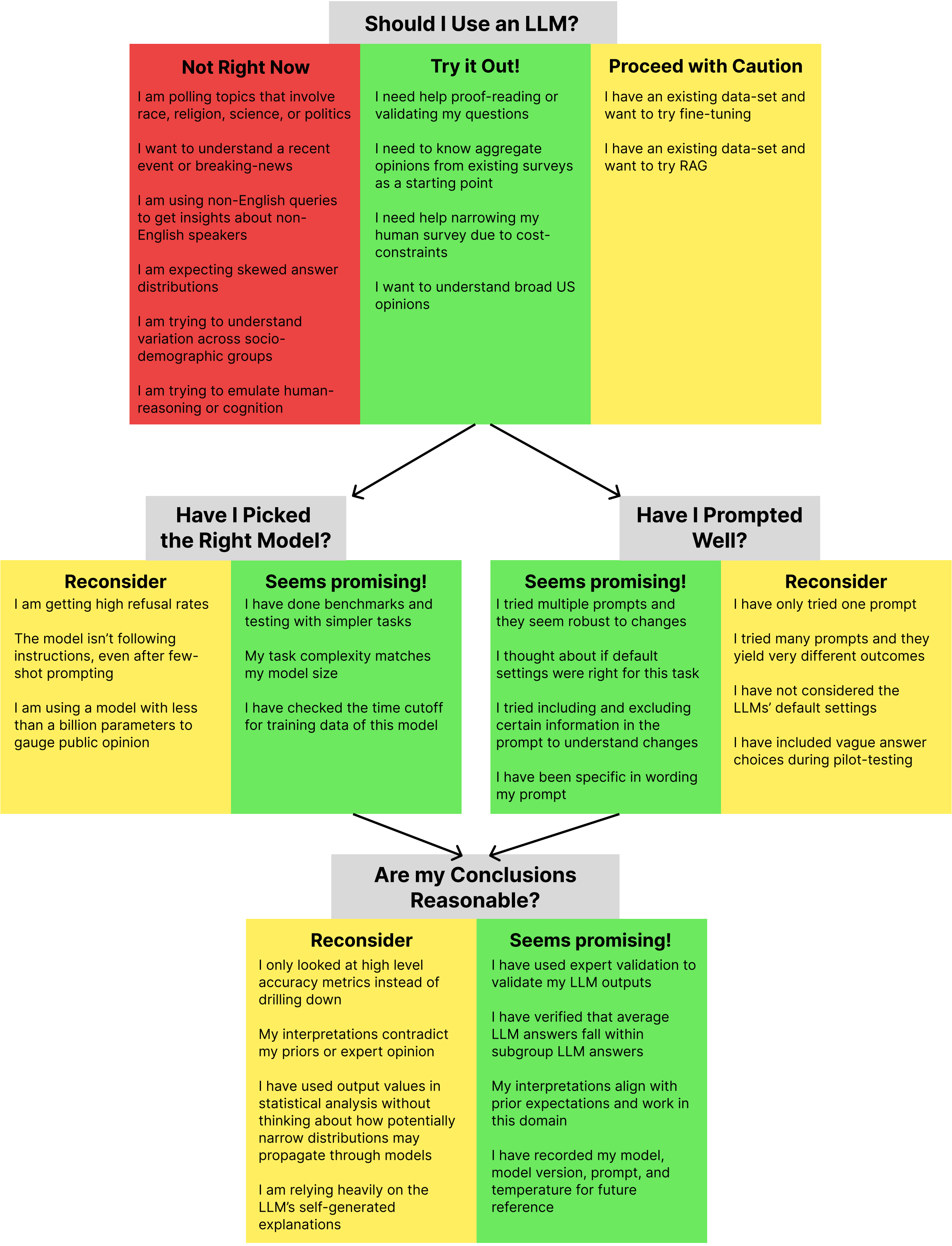
A less important parameter that practitioners should have awareness of is model temperature. Temperature is a tunable LLM parameter (typically in the settings) meant to control the randomness of an LLM's output. Theoretically, low temperatures make LLM outputs more deterministic, leading to more repetitive, less diverse, output. High temperatures, theoretically, therefore lead to more "creative" and unconventional outputs. That being said, research suggests that in practice the connection between "creativity" and temperature is weak [45], and that "higher" or more notionally creative temperatures lead to less accuracy and incoherent text [46]. Though practitioners may see much written about temperature, it is best to avoid relying too heavily on it as a parameter of interest in current LLM iterations.

Unreliable Self-Explanations

Finally, an important thing to keep in mind when using LLMs is that their self-explanations, although seemingly helpful, especially for survey research, do not actually reflect the actual reasoning process of the model [47]. Practitioners should be wary of relying heavily on these self-reported explanations throughout other stages of the survey research pipeline as well.

Practitioner's Decision Tree Guidance

Taking all of these insights together, we summarize results and operationalize them as practical guidance for practitioners: **Figure 2** provides a summary of what types of questions practitioners should consider when incorporating LLMs into survey workflows. The multiple-step decision process recommended here can, we hope, provide practitioners with more careful guidance, weighing the risks and opportunities associated with using generative AI models to make estimates about potential audience responses and public opinion, whether attitudes, preferences, or intended behaviors.



Should I Use an LLM?

Not Right Now

I am polling topics that involve race, religion, science, or politics

I want to understand a recent event or breaking-news

I am using non-English queries to get insights about non-English speakers

I am expecting skewed answer distributions

I am trying to understand variation across socio-demographic groups

I am trying to emulate human-reasoning or cognition

Try it Out!

I need help proof-reading or validating my questions

I need to know aggregate opinions from existing surveys as a starting point

I need help narrowing my human survey due to cost-constraints

I want to understand broad US opinions

Proceed with Caution

I have an existing data-set and want to try fine-tuning

I have an existing data-set and want to try RAG

Have I Picked the Right Model?

Reconsider

I am getting high refusal rates

The model isn't following instructions, even after few-shot prompting

I am using a model with less than a billion parameters to gauge public opinion

Seems promising!

I have done benchmarks and testing with simpler tasks

My task complexity matches my model size

I have checked the time cutoff for training data of this model

Have I Prompted Well?

Seems promising!

I tried multiple prompts and they seem robust to changes

I thought about if default settings were right for this task

I tried including and excluding certain information in the prompt to understand changes

I have been specific in wording my prompt

Reconsider

I have only tried one prompt

I tried many prompts and they yield very different outcomes

I have not considered the LLMs' default settings

I have included vague answer choices during pilot-testing

Are my Conclusions Reasonable?

Reconsider

I only looked at high level accuracy metrics instead of drilling down

My interpretations contradict my priors or expert opinion

I have used output values in statistical analysis without thinking about how potentially narrow distributions may propagate through models

I am relying heavily on the LLM's self-generated explanations

Seems promising!

I have used expert validation to validate my LLM outputs

I have verified that average LLM answers fall within subgroup LLM answers

My interpretations align with prior expectations and work in this domain

I have recorded my model, model version, prompt, and temperature for future reference

Figure 2: A decision tree with questions practitioners should ask themselves at each step of the LLM-integration process for survey research.

Conclusion

The new world of LLMs is filled with promises about what the technology can do. Certainly, there are new capabilities available now that were not available a few years ago. Still, as this field evolves it is important for practitioners to stay current, as new frontier models offer new capabilities.

For communications professionals, research into using AI for opinion research offers a clear takeaway: these tools are powerful supplements, but not substitutes, for human respondents. LLMs excel in the early stages of research—they are great for pre-testing surveys, refining question wording, and exploring initial messaging concepts. However, they are not yet reliable for capturing or estimating theoretical public opinion on their own. Current models consistently struggle to reflect the full diversity of human thought. They tend to stereotype certain demographic groups, and their reasoning processes are fundamentally different from that of humans. Safety filters potentially deepen these problems by narrowing response options; inherent biases in model training data means an LLM may not be able to represent the audiences you need to understand.

Therefore, the most effective approach is to use LLMs to strengthen survey design while relying on human samples as the gold standard for final data. Certainly, using LLMs as a kind "red team" or catalyst for internal scenario planning may be useful. But the outputs may not be anywhere close to scientific, potentially leading practitioners astray. As AI technology evolves, practitioners must remain vigilant about these core challenges, especially the tendency for models to oversimplify complex opinions or misrepresent marginalized voices. To use these tools responsibly, the best advice is to maintain transparency about one's methods and build in robust processes to validate AI findings against real human feedback. By treating LLMs as powerful but imperfect assistants that enhance—rather than replace—human insight, communications and media professionals doing message and product testing can leverage their benefits while protecting the integrity of one's practical research.

Title	Ref	Method	Domain	Dataset	Task Format	Results
AgentA/B Automated and Scalable Web A/B Testing and Interactive LLM Agents	[26]	Zero-Shot Prompting with Demographics	Social Media & Online Behavior	Custom Collected Dataset	Predicting next web action	**Positive** - Humans and AI Agents showed similar outcomes when using websites, but both took different paths to get there - with humans more exploratory and the AI Agents more goal-directed
Artificially Intelligent Opinion Polling	[32]	Zero-Shot Prompting	Politics & Law (US)	Custom Collected Dataset	Classification, Social Media Data	**Mixed** - Models tested were better suited annotating existing, unstructured sample data. Lack of explainability makes LLMs unsuitable for silicon sampling
Balancing Large Language Alignment and Algorithmic Fidelity in Social Science Research	[15]	Zero-Shot Prompting with Demographics	Politics & Law (US)	Existing Academic Study	Open-Ended	**Mixed** - Significant variations in performance across model family, prompt, and research objectives. Alignment processes heavily impact model ability to complete tasks related to in-group/out-group sentiment
Better Aligned with Survey Respondents or Training Data? Unveiling Political Leanings of LLMs on U.S. Supreme Court Cases	[37]	Zero-Shot Prompting	Politics & Law (US)	SCOPE	Multiple Choice	**Negative** - Tested models exhibited strong alignment with training corpora but did not align with human opinions regarding the Supreme Court
Beyond Believability: Accurate Human Behavior Simulation with Fine-Tuned LLMs	[30]	Few-Shot Prompting, Fine-Tuning	Economics & Consumer Behavior	Custom Collected Dataset	Predicting next web action	**Positive** - Incorporating domain-specific, context-aware human behavioral data via fine-tuning improved accuracy for online shopping behavioral simulations
Can AI Language Models replace human participants?	[36]	Few-Shot Prompting	Psychology & Human Behavior	Existing Academic Study	Likert-Style	**Positive** - Strong alignment between GPT models and human moral judgments
Can Large Language Models Capture Public Opinion about Global Warming? An Empirical Assessment of Algorithmic Fidelity and Bias	[23]	Zero-Shot Prompting with Demographics & Issue-Related Covariates	Science	Custom Collected Dataset	Multiple Choice	**Mixed** - Including demographic and issue-related covariates significantly enhances model accuracy when probing global warming; models performed poorly for non-Hispanic Black Americans
CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations	[24]	Zero-Shot Prompting with Demographics	Politics & Law (US), Psychology & Human Behavior	Custom Collected Dataset	Open-Ended, Question-Answering Interview	**Negative** - GPT models failed to capture multidimensionality of certain groups and perpetuated stereotypes

Continued on next page

Table 3 Continued from previous page

Title	Ref	Method	Domain	Dataset	Task Format	Results
Diminished Diversity-of-Thought in a Standard Language Model	[16]	Zero-Shot Prompting, Zero-Shot Prompting with Demographics	Politics & Law (US), Economics & Consumer Behavior, Psychology & Human Behavior	Existing Academic Study	Multiple Choice, Open-Ended	**Negative** - LLMs were only able to replicate about a third of the original findings and gave uniform answers for certain questions where humans had high variation (a “correct answer” effect) and were very sensitive to changes in order answers
Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design	[12]	Zero-Shot Prompting	Politics & Law (US)	Existing Academic Study	Multiple Choice	**Negative** - Tested models do not exhibit human-like behavior across a range of biased questions
Human-Like Intuitive Behavior and Reasoning Biases emerged in Large Language Models but disappeared in ChatGPT	[27]	Zero-Shot Prompting	Psychology & Human Behavior	Custom Collected Dataset	Semantic Illusion and Cognitive Reasoning Tests	**Mixed** - Tested GPT models outperform humans on these psychological tasks, avoiding the cognitive traps embedded in the tasks, unlike earlier models which fall for the traps
Ideology and Policy Preferences in Synthetic Data: The Potential of LLMs for Public Opinion Analysis	[7]	Zero-Shot Prompting, Zero-Shot Prompting with Demographics	Politics & Law (South Korea)	Existing Academic Study	Multiple Choice	**Mixed** - Tested LLMs can replicate key survey patterns, including demographic and ideological patterns, but also tend to overemphasize ideological differences on contentious issues
Language Model Fine-Tuning on Scaled Survey Data for Predicting Distributions of Public Opinions	[29]	Zero-Shot Prompting with Demographics, Few-Shot Prompting with Demographics, Fine-Tuning	Politics & Law, Science	General Social Survey (2022)	Multiple Choice	**Positive** - Fine-tuning LLMs with public opinion survey data improves their ability to predict human response distributions and generalize to unseen populations, question waves, and topics
Large Language Models as Simulated Economic Agents: What can We Learn from Homo Silicus?	[48]	Zero-Shot Prompting	Economics & Consumer Behavior	Existing Academic Research	Likert-Style, Multiple Choice	**Positive** - Tested GPT models were able to recover findings from experiments with actual humans relatively cheaply
Machine Bias. How Do Generative Language Models Answer Opinion Polls?	[14]	Zero-Shot Prompting with Demographics	Politics & Law (Multiple Countries)	World Values Survey	Multiple Choice	**Negative** - Tested models cannot replace research subjects for opinion/attitudinal research and have a ‘machine bias’ that randomly varies across topics
On LLM Augmented AB Experimentation	[34]	Zero-Shot Prompting, Few-Shot Prompting, Fine-Tuning	Social Media & Online Behavior	Existing Academic Research	A/B Testing	**Mixed** - In certain situations, models perform only slightly higher than random guessing. The most promising method involved fine-tuning LLMs to produce engaging headlines and then using the fine-tuned LLM for ratings

Continued on next page

Table 3 Continued from previous page

Title	**Ref**	**Method**	**Domain**	**Dataset**	**Task Format**	**Results**
Out of One, Many: Using Language Models to Simulate Human Samples	[13]	Zero-Shot Prompting with Demographics	Politics & Law (US)	ANES (2016/2020), Existing Academic Study	Open-Ended, Multiple Choice	**Positive** - Tested GPT models showed similarity with human samples that was deeper than surface level, capturing nuanced, multifaceted human opinions
Questioning the Survey Responses of Large Language Models	[19]	Zero-Shot Prompting	Politics & Law (US)	American Community Survey (2019)	Multiple Choice	**Negative** - Tested models are biased towards selecting answers labeled with letter 'A' and, when adjusting for this bias, trend toward uniformly random survey responses
Random Silicon Sample: Simulating Human Sub-Population Opinion Using Large Language Model Based on Group-Level Demographic Information	[17]	Zero-Shot Prompting with Demographics	Politics & Law (US)	ANES (2020)	Multiple Choice	**Mixed** - Tested models can replicate responses similar to U.S. public opinion polls; replicability depends on demographic group and topic. Models also show bias towards 'harmless' responses when discussing sensitive topics
Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting	[33]	Zero-Shot Prompting with Demographics	Social Media & Online Behavior	Existing Academic Study	Text Classification	**Mixed** - Sociodemographic prompting is not robust; prompt formulation and model choices lead to large variance in answers. Adding such information helps in some tasks, but not others. More than half of the labels are incorrectly classified
Should you use LLMs to simulate opinions? Quality checks for early-stage deliberation	[20]	Zero-Shot Neutral Prompting & Zero-Shot Prompting with Demographics	Politics & Law, Science	Custom Collected Dataset	Likert-Style	**Negative** - No models tested passed all quality control checks, meaning they do not pass a minimum threshold for silicon sampling reliability
Simulating Human Opinions with Large Language Models: Opportunities and Challenges for Personalized Survey Data Modeling	[18]	Zero-Shot Prompting with Demographics	Economics & Consumer Behavior	Custom Collected Dataset	Binary Outcomes, Likert-Style	**Mixed** - Tested models were able to approximate aggregate rankings but produced overly positive results and reduced variance compared to real data. No sociodemographic characteristic predicted accuracy
Synthesizing Public Opinions with LLMs: Role Creation, Impacts, and the Future to eDemocracy	[8]	Zero-Shot Prompting, RAG	Politics & Law (US)	Cooperative Election Study (2021)	Likert-Style	**Mixed** - Prompting alone led to high adherence on questions across models; RAG framework further improved adherence, but evaluative dataset may have already been present in LLM training data

Continued on next page

Table 3 Continued from previous page

Title	**Ref**	**Method**	**Domain**	**Dataset**	**Task Format**	**Results**
Synthetic Replacements for Human Survey Data? The Perils of Large Language Models	[21]	Zero-Shot Prompting with Demographics	Politics & Law (US)	ANES	Feeling Thermometer	**Negative** - Tested models showed less variation than human answers. Models exaggerated outgroup apathy in politics but underestimated outgroup apathy in race/religion. Responses were not repeatable across three different times
Take Caution in Using LLMs as Human Surrogates	[31]	Zero-Shot Prompting, Few-Shot Prompting, RAG, Fine-Tuning	Economics & Consumer Behavior	Custom Collected Dataset	Open-Ended	**Negative** - Significant differences between humans and tested models in reasoning depth, response distributions, and sensitivity to game framing, even with advanced techniques like RAG or Fine-Tuning
The Political Ideology of Conversational AI: Converging Evidence on ChatGPT's pro-environmental, left-libertarian orientation	[35]	Zero-Shot Prompting	Politics & Law (Germany, Netherlands)	Existing Academic Study	Likert-Style	**Mixed** - Using Wahl-O-Mat, StemWijzer, and Political Compass, tested GPT models exhibited a pro-environmental, left-libertarian political orientation
Towards Measuring the Representation of Subjective Global Opinions in Language Models	[25]	Zero-Shot Prompting with Country Information	Politics & Law (Global)	Pew Global Attitudes Survey, World Values Survey	Multiple Choice	**Negative** - By default, models tend to mimic USA/European/South American countries. Responses shift to reflect the country's perspective when prompted, but may reflect cultural stereotypes
Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies	[28]	Zero-Shot Prompting with (Implied) Demographics	Economics & Consumer Behavior, Psychology & Human Behavior	Existing Academic Study	Open-Ended, Multiple Choice	**Mixed** - Models were able to recreate economic, psycholinguistic, and social psychology experiments but had a hyper-accuracy distortion
Using LLMs for Market Research	[6]	Zero-Shot Neutral Prompting & Zero-Shot Prompting with Demographics	Economics & Consumer Behavior	Existing Academic Study	Multiple Choice, Open-Ended	**Mixed** - Models can reflect some willingness-to-pay differences between customer groups, but fail to reflect other differences and do not reflect any group particularly well

Table 3: A summary of academic papers that have explored the feasibility of silicon sampling through experiments.

References

[1] T. S. Behrend and R. N. Landers, “Participant Interactions with Artificial Intelligence: Using Large Language Models to Generate Research Materials for Surveys and Experiments,” *J. Bus. Psychol.*, June 2025, doi: 10.1007/s10869-025-10035-6.

[2] M. Khalifa and M. Albadawy, “Using artificial intelligence in academic writing and research: An essential productivity tool,” *Comput. Methods Programs Biomed. Update*, vol. 5, p. 100145, 2024, doi: 10.1016/j.cmpbup.2024.100145.

[3] M. Sarstedt, S. J. Adler, L. Rau, and B. Schmitt, “Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines,” *Psychol. Mark.*, vol. 41, no. 6, pp. 1254–1270, 2024, doi: 10.1002/mar.21982.

[4] D. M. Rothschild, J. Brand, H. Schroeder, and J. Wang, “Opportunities and risks of LLMs in survey research,” 2024, *SSRN*. doi: 10.2139/ssrn.5001645.

[5] L. Malmqvist, “Sycophancy in Large Language Models: Causes and Mitigations,” Nov. 22, 2024, *arXiv*: arXiv:2411.15287. doi: 10.48550/arXiv.2411.15287.

[6] J. Brand, A. Israeli, and D. Ngwe, “Using LLMs for Market Research”.

[7] K. Lee, J. Park, S. Choi, and C. Lee, “Ideology and Policy Preferences in Synthetic Data: The Potential of LLMs for Public Opinion Analysis”.

[8] R. Karanjai *et al.*, “Synthesizing Public Opinions with LLMs: Role Creation, Impacts, and the Future to eDemorcacy,” Mar. 31, 2025, *arXiv*: arXiv:2504.00241. doi: 10.48550/arXiv.2504.00241.

[9] D. M. Adhikari, A. Hartland, I. Weber, and V. K. Cannanure, “Exploring LLMs for Automated Generation and Adaptation of Questionnaires,” in *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, July 2025, pp. 1–23. doi: 10.1145/3719160.3736606.

[10] O. Haavisto and R. Welsch, “Questionnaires for Everyone: Streamlining Cross-Cultural Questionnaire Adaptation with GPT-Based Translation Quality Evaluation,” July 30, 2024, *arXiv*: arXiv:2407.20608. doi: 10.48550/arXiv.2407.20608.

[11]P. Hämäläinen, M. Tavast, and A. Kunnari, “Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg Germany: ACM, Apr. 2023, pp. 1–19. doi: 10.1145/3544548.3580688.

[12] L. Tjuatja, V. Chen, T. Wu, A. Talwalkwar, and G. Neubig, “Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design,” *Trans. Assoc. Comput. Linguist.*, vol. 12, pp. 1011–1026, Sept. 2024, doi: 10.1162/tacl_a_00685.

[13] L. P. Argyle, E. C. Busby, N. Fulda, J. Gubler, C. Rytting, and D. Wingate, “Out of One, Many: Using Language Models to Simulate Human Samples,” *Polit. Anal.*, vol. 31, no. 3, pp. 337–351, July 2023, doi: 10.1017/pan.2023.2.

[14] J. Boelaert, S. Coavoux, É. Ollion, I. Petev, and P. Präg, “Machine Bias. How Do Generative Language Models Answer Opinion Polls? <sup/>,” *Sociol. Methods Res.*, vol. 54, no. 3, pp. 1156–1196, Aug. 2025, doi: 10.1177/00491241251330582.

[15] A. Lyman, B. Hepner, L. P. Argyle, E. C. Busby, J. R. Gubler, and D. Wingate, “Balancing Large Language Model Alignment and Algorithmic Fidelity in Social Science Research,” *Sociol. Methods Res.*, vol. 54, no. 3, pp. 1110–1155, Aug. 2025, doi: 10.1177/00491241251342008.

[16] P. S. Park, P. Schoenegger, and C. Zhu, “Diminished diversity-of-thought in a standard large language model,” *Behav. Res. Methods*, vol. 56, no. 6, pp. 5754–5770, Jan. 2024, doi: 10.3758/s13428-023-02307-x.

[17] S. Sun *et al.*, “Random Silicon Sampling: Simulating Human Sub-Population Opinion Using a Large Language Model Based on Group-Level Demographic Information”.

[18] C. Kaiser, J. Kaiser, V. Manewitsch, L. Rau, and R. Schallner, “Simulating Human Opinions with Large Language Models: Opportunities and Challenges for Personalized Survey Data Modeling,” *N. Y. City*, 2025.

[19] R. Dominguez-Olmedo, M. Hardt, and C. Mendler-Dünner, “Questioning the Survey Responses of Large Language Models,” *NeurIPS*, doi: https://doi.org/10.48550/arXiv.2306.07951.

[20] T. Neumann, M. De-Arteaga, and S. Fazelpour, “Should you use LLMs to simulate opinions? Quality checks for early-stage deliberation,” May 05, 2025, *arXiv*: arXiv:2504.08954. doi: 10.48550/arXiv.2504.08954.

[21] J. Bisbee, J. D. Clinton, C. Dorff, B. Kenkel, and J. M. Larson, “Synthetic Replacements for Human Survey Data? The Perils of Large Language Models,” *Polit. Anal.*, vol. 32, no. 4, pp. 401–416, Oct. 2024, doi: 10.1017/pan.2024.5.

[22] N. E. Sanders, A. Ulinich, and B. Schneier, “Demonstrations of the Potential of AI-based Political Issue Polling,” *Harv. Data Sci. Rev.*, vol. 5, no. 4, Oct. 2023, doi: 10.1162/99608f92.1d3cf75d.

[23] S. Lee *et al.*, “Can large language models estimate public opinion about global warming? An empirical assessment of algorithmic fidelity and bias,” *PLOS Clim.*, vol. 3, no. 8, p. e0000429, Aug. 2024, doi: 10.1371/journal.pclm.0000429.

[24] M. Cheng, T. Piccardi, and D. Yang, “CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations,” Oct. 17, 2023, *arXiv*: arXiv:2310.11501. doi: 10.48550/arXiv.2310.11501.

[25] E. Durmus *et al.*, “Towards Measuring the Representation of Subjective Global Opinions in Language Models,” Apr. 12, 2024, *arXiv*: arXiv:2306.16388. doi: 10.48550/arXiv.2306.16388.

[26] D. Wang *et al.*, “AgentA/B: Automated and Scalable Web A/BTesting with Interactive LLM Agents,” Apr. 21, 2025, *arXiv*: arXiv:2504.09723. doi: 10.48550/arXiv.2504.09723.

[27] T. Hagendorff, S. Fabi, and M. Kosinski, “Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT,” *Nat. Comput. Sci.*, vol. 3, no. 10, pp. 833–838, Oct. 2023, doi: 10.1038/s43588-023-00527-x.

[28] G. Aher, R. I. Arriaga, and A. T. Kalai, “Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies,” July 09, 2023, *arXiv*: arXiv:2208.10264. doi: 10.48550/arXiv.2208.10264.

[29] J. Suh, E. Jahanparast, S. Moon, M. Kang, and S. Chang, “Language Model Fine-Tuning on Scaled Survey Data for Predicting Distributions of Public Opinions,” Feb. 24, 2025, *arXiv*: arXiv:2502.16761. doi: 10.48550/arXiv.2502.16761.

[30] Y. Lu *et al.*, “Beyond Believability: Accurate Human Behavior Simulation with Fine-Tuned LLMs,” Mar. 26, 2025, *arXiv*: arXiv:2503.20749. doi: 10.48550/arXiv.2503.20749.

[31] Y. Gao, D. Lee, G. Burtch, and S. Fazelpour, “Take caution in using LLMs as human surrogates,” *Proc. Natl. Acad. Sci.*, vol. 122, no. 24, p. e2501660122, June 2025, doi: 10.1073/pnas.2501660122.

[32] R. Cerina and R. Duch, “Artificially Intelligent Opinion Polling,” Sept. 12, 2023, *arXiv*: arXiv:2309.06029. doi: 10.48550/arXiv.2309.06029.

[33] T. Beck, H. Schuff, A. Lauscher, and I. Gurevych, “Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting,” *Association for Computational Linguistics*, vol. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2589–2615, Mar. 2024, doi: 10.18653/v1/2024.eacl-long.159.

[34] S. Shankar, R. Sinha, and M. Fiterau, “On LLM Augmented AB Experimentation,” presented at the CaLM @ NeurIPS, Oct. 2024.

[35] J. Hartmann, J. Schwenzow, and M. Witte, “The political ideology of conversational AI: Converging evidence on ChatGPT’s pro-environmental, left-libertarian orientation,” Jan. 05, 2023, *arXiv*: arXiv:2301.01768. doi: 10.48550/arXiv.2301.01768.

[36] D. Dillion, N. Tandon, Y. Gu, and K. Gray, “Can AI language models replace human participants?,” *Trends Cogn. Sci.*, vol. 27, no. 7, pp. 597–600, July 2023, doi: 10.1016/j.tics.2023.04.008.

[37] S. Xu, T. Y. S. S. Santosh, Y. Elazar, Q. Vogel, B. Plank, and M. Grabmair, “Better Aligned with Survey Respondents or Training Data? Unveiling Political Leanings of LLMs on U.S. Supreme Court Cases,” June 28, 2025, *arXiv*: arXiv:2502.18282. doi: 10.48550/arXiv.2502.18282.

[38] S. Badshah and H. Sajjad, “Quantifying the Capabilities of LLMs across Scale and Precision,” May 08, 2024, *arXiv*: arXiv:2405.03146. doi: 10.48550/arXiv.2405.03146.

[39] A. Chowdhery *et al.*, “PaLM: Scaling Language Modeling with Pathways,” Oct. 05, 2022, *arXiv*: arXiv:2204.02311. doi: 10.48550/arXiv.2204.02311.

[40] D. Jaroslawicz, B. Whiting, P. Shah, and K. Maamari, “How Many Instructions Can LLMs Follow at Once?,” July 15, 2025, *arXiv*: arXiv:2507.11538. doi: 10.48550/arXiv.2507.11538.

[41] L. Ouyang *et al.*, “Training language models to follow instructions with human feedback,” Mar. 04, 2022, *arXiv*: arXiv:2203.02155. doi: 10.48550/arXiv.2203.02155.

[42] Y. Bai *et al.*, “Constitutional AI: Harmlessness from AI Feedback,” Dec. 15, 2022, *arXiv*: arXiv:2212.08073. doi: 10.48550/arXiv.2212.08073.

[43] S. Raschka, “Understanding Reasoning LLMs.” Accessed: Aug. 27, 2025. [Online]. Available: https://magazine.sebastianraschka.com/p/understanding-reasoning-llms

[44] P. Shojaee, I. Mirzadeh, K. Alizadeh, M. Horton, S. Bengio, and M. Farajtabar, “The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity”.

[45] M. Peeperkorn, T. Kouwenhoven, D. Brown, and A. Jordanous, “Is Temperature the Creativity Parameter of Large Language Models?,” May 01, 2024, *arXiv*: arXiv:2405.00492. doi: 10.48550/arXiv.2405.00492.

[46] M. Renze, “The Effect of Sampling Temperature on Problem Solving in Large Language Models,” in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 7346–7356. doi: 10.18653/v1/2024.findings-emnlp.432.

[47] A. Madsen, S. Chandar, and S. Reddy, “Are self-explanations from Large Language Models faithful?,” May 16, 2024, *arXiv*: arXiv:2401.07927. doi: 10.48550/arXiv.2401.07927.

[48] John J. Horton, “Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?,” Jan. 18, 2023, *arXiv*: arXiv:2301.07543. doi: 10.48550/arXiv.2301.07543.