Sociotechnical Risks of Simulating Humans with Language Model Agents

CAROLYN ZOU, Northwestern University, USA

Large language models are increasingly being used as agents to simulate human participants in behavioral experiments, offering a cost-effective and scalable alternative to traditional human behavioral research. However, the validity and robustness of agent-based behavioral experiments remain largely unexplored. In this paper, we systematically investigate the intrinsic properties of LLMs that threaten the reliability of agent-based simulations, focusing on issues arising from prompt sensitivity, stochasticity, and memorization. We propose a set of auditing methods, namely perturbation and iteration of experimental conditions, to assess the stability of simulated findings. Applying these methods to a selection of prominent studies employing language model agents, we reveal prevalent failure modes that undermine the robustness of their conclusions. Our findings highlight the need for careful and informed evaluation when interpreting the results of simulated behavioral experiments and emphasize the importance of developing methodical procedures for assessing the reliability of these simulations. We discuss the implications of our work for the use of language models in human behavioral research and provide recommendations for future research aimed at developing more robust simulations via auditing along dimensions of breadth and depth. This work contributes to the growing body of literature on the evaluation and use of LLMs in social and behavioral science research, offering best practice recommendations for researchers seeking to leverage these potent tools responsibly and reliably.

1 INTRODUCTION

The ability of large language models (LLMs) to generate believable natural language across contexts has led to their use in domains where human labor would otherwise be required. As a result, many novel applications of LLMs use these systems to produce outputs that increasingly resemble human behavior. In particular, some prior work has focused on the development of language model "agents" by supplying LLMs with capabilities inspired by human cognitive processes; namely, interaction with an external environment and/or other agents, reflection, and memory. [24]

Agential applications of LLMs have also enabled attempts at simulating participants for behavioral experiments. In lieu of human subjects, language models are instructed to behave as those subjects, and practitioners aim to produce realistic experimental outcomes while minimizing cost and potential risk to participants. Prior and ongoing work in this area seeks to establish the legitimacy of such methods by means of replication: a known outcome in the social or behavioral sciences is selected, the original study instruments are administered to language model agents, and the simulated outcome is compared to the original. The successful results of these replications, such as *Can AI language models replace human participants?* [9] and *Using cognitive psychology to understand GPT-3* [3], have led many to conclude that the outputs of language model agents seem realistic, often agree with observed human behavior, and can be interpreted using metaphors of human cognition.

However, the application of language models to behavioral research presents unique challenges that threaten the robustness and validity of the findings. LLMs exhibit a number of intrinsic properties that can lead to unexpected and unreliable behaviors, such as sensitivity to prompt wording, stochasticity in output generation, and memorization of training data. These properties can introduce significant variability and inconsistency in agent-generated behaviors, making it difficult to draw reliable conclusions about the simulated outcomes. Moreover, the opacity of LLMs makes it difficult to interpret the capabilities and limitations of agent-based experiments; these interpretability challenges and the resulting sociotechnical gap may cause practitioners to further misinterpret results. Without sufficient knowledge of how language models and resulting agents generate their outputs, researchers may be prone to misinterpreting

Author's address: Carolyn Zou, cqz@u.northwestern.edu, Northwestern University, Evanston, Illinois, USA.

findings or attributing unwarranted significance to the simulated behaviors. This can lead to erroneous confidence in the validity of agent-based research or the applicability of these methods to the study of human behavior.

In this paper, we aim to systematically investigate the validity of behavioral research conducted with language model agents by identifying and documenting prevalent modes of failure for the robustness of these simulations. We review prior research in this area and conduct a series of audits on prominent agent-based studies to assess the impact of these threats on the reliability of their conclusions. Our work aims to contribute to the growing body of literature on the use and evaluation of simulated behavioral research by:

(1) Identifying key threats to the robustness and validity of behavioral experiments conducted with language model agents, focusing on issues arising from prompt sensitivity, stochasticity, and memorization in LLMs.
(2) Proposing a set of systematic auditing methods, perturbation and iteration, to assess the stability and consistency of simulated behavioral outcomes.
(3) Demonstrating an application of these methods by auditing a selection of prominent studies utilizing language model agents, revealing prevalent failure modes that undermine the reliability of their conclusions.
(4) Discussing the implications of our findings for the use of language model agents in behavioral research and providing recommendations for future work in this area, with an emphasis on developing more deliberate methods for both producing and interpreting agent-produced data.

By procedurally assessing the robustness and validity of agent-based behavioral experiments, we hope to provide researchers with a clearer understanding of the limitations and challenges associated with using LLMs to simulate human behavior, particularly with regards to robustness threats that are intrinsic to architectural features of LLMs. Our work highlights the need for deliberate evaluation when interpreting the results of simulated studies, and emphasizes the importance of developing more rigorous processes for assessing the reliability of work in this emerging field. Through the identification and documentation of prevalent failure modes, we hope to contribute to the development of robust best practices for using language models in social and behavioral science research, ultimately leading to more accurate and reliable conclusions.

2 PRIOR WORK

2.1 Replication with language model agents

The use of language model agents in social and behavioral science research has gained significant attention in recent years. Researchers have explored the potential of language models to replicate human responses across a wide range of tasks, from cognitive reflection tests [11] and decision-making heuristics [30] to moral reasoning [27] and public opinion surveys [8, 26]. These studies have substantiated the promise of language model agents as a cost-effective and efficient means of simulating human behavior, potentially enabling large-scale, high-powered studies that would be infeasible with traditional methods.

2.1.1 Task types. The majority of studies employing language models focus on replicating human behavior in various experimental settings. For instance, [2] used agents to simulate political text, voting behavior, and partisan trait correlations, while [25] and [22] explored the replication of psychology studies and sensory judgments, respectively. This focus on replication is driven by the need to establish the viability of using agents as proxies for human participants, with known ground truth results providing a simple method for assessing parity.

Tasks can be classified as in- or out-of-distribution, referring to whether information about the replication target was present in a language model's training data. Currently, there are no definitive methods to establish whether a given task is in-distribution, but some proxies exist. Given that "language models (mostly) know what they know," [18] a practitioner can query the LLM about the study in question. Alternatively, the date of publication can be compared to the reported "knowledge cutoff" of the language model [14].

Another relevant distinction is whether a task aims to measure attitudinal or behavioral outcomes. Attitudinal outcomes are typically assessed through self-report measures, such as surveys or questionnaires, in which participants are directly asked to express their preferences, opinions, or beliefs. Several studies have employed language models to complete personality assessments [5, 17], rate the ethics of scenarios [9], and predict opinions on political issues [13]. In these cases, the participants' responses are taken at face value as an accurate representation of their attitudes. In contrast, behavioral experiments focus on observing participants' actions within a controlled environment. For example, [12] used LLMs to simulate economic agents in various experimental settings, while [32] explored the generation of faithful synthetic data for computational social science. In these studies, participants are given some information about the context and constraints of the experiment but are not fully aware of the specific measures being studied. Researchers then draw conclusions about the participants' behavior based on their observed actions within the experimental setting.

2.1.2 Input types. The majority of these studies rely on proprietary models developed by OpenAI, particularly GPT-3 and its variants, which are accessed through the OpenAI API. These models have been used to simulate a wide range of tasks, including cognitive reflection tests [11], decision-making heuristics [30], economic experiments [12], and public opinion surveys [26]. However, a growing number of studies are also exploring the use of open-source models, such as those utilizing BERT [8], GPT-Neo [16], and Alpaca [1, 19]. These models offer greater flexibility, as they can be easily fine-tuned on domain-specific data. Additionally, researchers using open-source models have levers of control over architectural factors that are more transparent and more influential than those available for the most commonly used proprietary models. However, their performance often lags behind that of proprietary models on benchmarks and leaderboards such as Chatbot Arena [7].

Language models rely on natural language inputs, or prompts, that contain instructions for completing tasks. Therefore, the success of agents in faithfully replicating human behavior depends on the prompts used to elicit their responses. Most studies employ a zero-shot learning approach, where the model is given a task description or question without any sample responses in the prompt. This allows researchers to assess the model's inherent capabilities and limitations, while minimizing the confounding effects of arbitrarily selected factors in the input, such as examples used for one- or few-shot learning. However, some studies also employ few-shot learning techniques, where the model is provided with a small number of examples before being asked to complete the task. For instance, [9] used a few-shot approach to normalize the model's ethical ratings, while [19] compared the performance of zero-shot, few-shot, and fine-tuned models on a survey imputation task.

Lastly, many studies aim to assess the impact of different "persona" attributes on agents' behavior by prompting language models with a description of a person that it should represent. The most common approach is to include generic gendered names or demographic descriptions of race or age in the prompt. For example, [29] and [5] used gendered names and demographic attributes to explore the impact of these factors on agents' responses to economic experiments and personality assessments, respectively. Other studies provide more detailed personas with occupations [1, 30] or personality traits [17]. Some studies also explore the use of real-world personas, such as politicians [33]. In

these cases, the model is typically conditioned on the persona's name and relevant biographical information, such as party affiliation or voting record.

2.1.3 Criticisms. While the capabilities of language model agents have been explored in numerous contexts, their use has also faced criticism from various perspectives. One key concern is the potential for language models to perpetuate biases. [15] argues that psychology research using language models trained on large, uncontrolled datasets may encode and amplify societal biases, raising methodological implications for results produced this way.

Another point of criticism is the acknowledgment that language models are fundamentally different from human cognition. [20] emphasizes that language modeling processes are based on pattern recognition and statistical associations, rather than true understanding or reasoning. This difference can lead to issues when using techniques like reinforcement learning from human feedback (RLHF) to fine-tune language models, as these methods may shift the models away from any true distribution of human behavior and towards outputs that are optimized for specific reward functions.

Furthermore, some researchers have found that modifications to prompts and hyperparameters can have substantial influence on the outcomes of language model simulations. During the development of this work, concurrent research conducted similar extensions of language model-based simulations: [21] explores the Horizon economic decision making task from [3] with varying degrees of chain-of-thought prompting, hyperparameters, and underlying models, finding that the changes can significantly impact the quality and diversity of generated outputs.

2.2 The replication crisis in the social sciences

Given the focus on replication in much of the aforementioned work, we utilize literature about the "replication crisis" in the social sciences to understand the latent robustness threats. Researchers of the replication crisis assert that many high-profile studies in psychology, economics, and political science fail to replicate when conducted by independent researchers [4, 23]. Several factors have been identified as contributing to this issue. These include publication bias, where journals are more likely to publish positive and novel findings than null results or replications; undesirable research practices, such as p-hacking and selective reporting; and the use of small sample sizes, which can lead to false positives and inflated effect sizes [28]. The use of LLMs in social and behavioral research faces challenges that closely resemble those identified as causes of the replication crisis.

The high sensitivity of language models to the specific wording and formatting of prompts is analogous to the problem of p-hacking in human-subject studies: just as researchers may engage in p-hacking by selectively reporting results or manipulating data until they achieve a desired level of statistical significance, practitioners may engage in "prompt hacking" by modifying prompts until they elicit a desired response from the model. Similarly, the use of small sample sizes in human-subject research, which can lead to underpowered studies and non-replicable findings, is paralleled by the practice of running simulations only once or a small number of times, obscuring the overall distribution of possible outputs and potentially leading to unreliable conclusions. Finally, the publication bias favoring positive and novel results over null findings or replications in traditional research can also arise in agent-based studies, given the potential for prompt hacking and the stochasticity of results. Researchers may be able to tweak prompts or re-run simulations until they obtain a successful outcome, leading to the selective reporting of positive findings and the suppression of null or inconsistent results.

3 THREATS TO ROBUSTNESS

3.1 Features of language models

Inherent technical characteristics of language models can pose challenges for robustness when using agents in behavioral experiments, especially when faulty metaphors of human cognition obscure the unhumanlike nature of these processes. Three prominent examples of this are prompt sensitivity, stochasticity, and memorization.

Prompting. Prompting is the primary method of eliciting desired behaviors from language models. However, these models can be highly sensitive to the specific wording, formatting, and context of the prompts provided. Even small changes in prompting can lead to substantial differences in model outputs. This is problematic for robustness, as it suggests experimental results may not be reproducible if prompts are even trivially modified. It also means the conclusions drawn from an experiment with synthetic participants could be an artifact of the specific prompt used rather than a robust, replicable, and generalizable finding about the phenomenon being studied.

Stochasticity. Stochasticity refers to the inherent randomness in language model outputs. Most language models do not produce fully deterministic results, but rather sample from a probability distribution over possible outputs. This sampling is typically done using techniques like top-k sampling or nucleus sampling, which truncate the probability distribution to only consider the most likely tokens. The specific sampling parameters, such as the value of k or the probability threshold for nucleus sampling, can significantly impact the variability and quality of the generated text.

While some stochasticity can be beneficial for mimicking the natural variance in human behavior, it also makes it difficult to obtain consistent, replicable results from language model agents. Findings may not hold up under repetition of the experiment, raising several concerns: the distribution of outputs may be misaligned with the distribution of human responses, and a single experimental outcome may be statistically improbable when considering the overall distribution.

Memorization. Memorization is a known issue with language models, where the model reproduces verbatim snippets of content from its training data in response to prompts that resemble in-distribution sequences. This occurs because neural language models are trained to minimize the negative log-likelihood of the training data, which incentivizes them to assign high probabilities to sequences that appear frequently in the data. If a simulated experiment involves a prompt that was present in training, causing the language model to return a memorized response, the output could be misinterpreted as the model exhibiting some capability or behavior when it is only repeating previous seen data.

This compromises the validity of conclusions drawn from such an output, and is especially concerning considering the current popularity of replication as a means to validate the use of agents for behavioral experiments; these studies often make claims about generalizable capabilities despite testing on input/output pairs that were public and thus potentially included in the training corpus. Some prior work attempts to mitigate memorization effects through techniques like response rescaling, but the effectiveness of these approaches is still poorly understood.

3.2 Sociotechnical gap

3.2.1 The gulf of process. Human interaction with language models is particularly fraught with challenges arising from the anthropomorphization of systems that produce natural language. This tendency to attribute human-like qualities to language models can lead to misunderstandings and unrealistic expectations about their capabilities and limitations. In addition, many of the most widely used and easily accessible language models tend to respond with signals indicative of intention, such as first-person language or claims to sentience. Even experts are susceptible—research papers in

computer science increasingly ascribe human characteristics to these systems, this is especially prevalent in research related to language models, and downstream media coverage of research amplifies these errors [6].

The process by which generative models produce outputs differs substantially from that of human cognition processes. This is perhaps most evident when thinking about diffusion in image generation models: most create images by starting with random noise and iteratively denoising according to a natural language prompt. However, these generation processes are largely obscured in generative models released for interaction with non-technical end-users, making these gaps less evident. With language models, especially, natural language outputs from generative models can look indistinguishable from recorded natural language data from human subjects. The generation process typically involves a decoder-only transformer architecture, where each token is generated conditioned on the previous tokens. The model learns to assign probabilities to different possible next tokens based on patterns in the training data. This is fundamentally different from how people utilize language, involving meaning representation, reasoning, planning, and other cognitive processes.

There exists a fundamental "gulf of process" between the processes that generative models and humans use to produce language or visual artifacts [31]. For instance, in comparing how language models generate text and how humans produce language, we note that language models are not engaging in anything resembling humanlike reasoning, despite surface-level appearances. Rather, they are performing sequence prediction based on patterns in their training data. This represents a significant disconnect between language models and the human cognition metaphors that are often used to understand them.

The implications of the gulf of process amplify validity concerns around simulated behavioral research, especially at the level of results interpretation by practitioners. When human evaluators anthropomorphize language models and ascribe LLM outputs to human-like thought processes, it becomes easy to over-attribute meaning and intentionality to their behaviors. Experimenters may inaccurately conclude that believable responses indicate human behavioral phenomena, when a more grounded analysis would recognize the output is an artifact of the language modeling process.

Prompting. Prompt sensitivity exacerbates this issue, as small changes to experiment designs that would be trivial for human participants can substantially change LLM behaviors by shifting the model into a different region of its learned probability space. A language model may appear to robustly exhibit some behavior under one specific prompting approach, but this illusion of humanlike consistency fails to hold given minor prompt perturbations.

Stochasticity. The stochastic nature of LLM outputs also widens the process gulf. A single trial is often insufficient to draw robust conclusions, as the model may generate a completely different response if the experiment is repeated. This can lead to inconsistent or contradictory results that don't accurately reflect the underlying phenomenon being studied.

Researchers may also interpret inconsistency in model-generated behaviors as representing meaningful variability, akin to the natural variance in human behavior between individuals and across contexts. In reality, stochasticity in language model outputs is an architectural artifact not grounded in any stable world model or intentional behavioral variation on the part of the AI. The level of variance indicated in repeated simulated outcomes may only be spuriously correlated with the actual distribution of outcomes.

Memorization. Similarly, memorization can lead experimenters to read faithful replication of human-written content as an agent having human-like knowledge or exhibiting human-like behavior. In truth, memorized outputs indicate nothing more than the model's ability to store and reproduce training data under the right circumstances.

Currently, replication is a popular method of validation as it provides a clear ground truth and often open access to the necessary experimental resources. However, this ease of access poses a threat to validity: the natural language artifacts of a replication target, including both its experimental instruments and the outcomes as described in a published paper, may have also been collected for the text corpus upon which a LLM was trained. In a study evaluating the ability of language models to replicate human behavior on a variety of psychology findings, agents showed little to no variation in responses, with a strong preference for the supposedly correct answer [25].

3.2.2 Resulting threats. The process gulf between language models and human cognition, coupled with misaligned mental models, gives rise to a range of sociotechnical issues when language model agents are used as proxies for human participants in behavioral research. These issues can be broadly categorized into problems of over-attribution, false extrapolation, and misplaced trust.

Over-attribution. Over-attribution refers to the tendency of human evaluators to ascribe unwarranted meaning, intentionality, and capability to LLM outputs. The human-like coherence of language model generations, especially in the case of large models, leads people to anthropomorphize these systems and interpret their behaviors as the product of human-like thought processes, knowledge, and reasoning abilities.

In reality, LLMs are engaging in complex pattern matching and sequence prediction based on their training data, not understanding or intentionality. The core architecture of most language models is the transformer, which learns to map input sequences to output sequences using an attention mechanism. The model is trained to predict the next token in a sequence given the previous tokens, without any explicit representation of meaning or reasoning.

This over-attribution can lead researchers to draw erroneous conclusions about the human behavioral phenomena supposedly being exhibited by language model agents in experiments. They may believe they have observed theory of mind reasoning, for example, when in fact the model has simply generated a plausible response based on patterns in its training data. Crucially, in cases where a ground truth for evaluation does not exist, multiple conflicting outcomes can seem plausible which negates the value of simulating outcomes. Over-attribution can also lead to overestimation of the general capabilities of languge models, with people assuming that performance in one domain that resembles human behavior translates to broad, humanlike intelligence.

False extrapolation. False extrapolation arises when people assume that agent behaviors observed under one specific set of experimental conditions will hold true more generally. They may believe that an agent exhibiting some desired behavior under a particular prompt is a robust result that will consistently appear across contexts. In reality, their outputs depend on prompt wording, formatting, and other contextual factors. A small change in experimental design that would be trivial for human participants can substantially alter the behavior of a language model by shifting it into a different region of its learned probability space.

This fragility, combined with inherent stochasticity in model outputs, means that results obtained from simulated experiments may not be reproducible or generalizable. An effect observed in one study could be an artifact of the specific prompts, models, or parameters used rather than a finding about the behavioral phenomenon being studied. False extrapolation can lead to overconfidence in the validity of conclusions drawn from simulated experiments.

Misplaced trust. Misplaced trust is a related issue, where the human-like language abilities of language models can inspire unwarranted trust in their outputs. Researchers may put undue faith in the accuracy, reliability, and truthfulness of simulated responses, overlooking the potential for the model to generate false, inconsistent, or nonsensical information.

For instance, they may assume that a language model prompted to provide an explanation for prior behavior will generate text indicative of a causal factor, when in reality the autoregressive nature of LLMs necessitates that any rationalization is post-hoc. Language models do not maintain a coherent world model or sense of their own "beliefs" or "intents" over time; each response is generated independently based on the immediate prompt. This is especially concerning for proposals to use simulated participants in high-stakes contexts, such as modeling the effects of policy interventions or simulating human behavior in sensitive domains. Overreliance on generated results could thus lead to flawed decision making, substantiated with unreliable evidence.

4 AUDITING METHODS

We present a series of experimental procedures to identify the presence and extent of these threats, including probing the effects of prompt variation on model outputs, measuring the impact of inherent stochasticity on the reliability of these outputs, and determining the extent of memorization by examining the models' responses for potential regurgitation of their training data.

4.1 Perturb

First, we propose systematic **perturbation** of the experimental setup for testing the robustness of LLM-based behavioral studies. This involves making small, controlled changes to the inputs given to the model and observing how these changes affect the outputs.

One core perturbation is prompt variation. By making targeted or systematic changes to the wording, formatting, or structure of the prompts used to elicit responses from the model, we can test how sensitive the results are to the specific prompts used. If the model's outputs change substantially in response to minor prompt variations, especially when such a change would not reasonably affect behavior in human participants, it suggests that the initial findings may not be robust. We also alter the context in which the model is applied by testing it on novel scenarios or datasets. If the model's performance degrades substantially on these new examples, it suggests that it may be overfitting to the specific scenarios used in the original study.

Another form of perturbation is changing the model's sampling parameters, namely the temperature hyperparameter, which affects the computation of token probabilities. Temperature ranges from 0 to 2 and typically defaults to 0.7 or 1.0, with lower values indicating stronger determinism and higher values resulting in more randomness as uncommon tokens appear more often. By comparing the model's behavior under low and high temperatures, we can test the extent to which the results depend on the specific sample obtained. If the results are highly variable across different stochastic samples, it suggests that any individual outcome may not be representative of the model's overall behavior.

Finally, when relevant, we vary the model itself by testing different versions or configurations of the same underlying architecture. For example, we might compare the results obtained using the GPT-3 API to those obtained using the ChatGPT API, test different sized versions of the same model family, or compare a deprecated model with its recommended replacement. If the results differ substantially across these different model variations, it suggests that the findings may be sensitive to the specific model used.

4.2 Iterate

The second method we employ is **iteration**, which involves obtaining many outcomes from identical inputs to obtain a distribution of results. This is particularly important when working with stochastic models, as any individual outcome

may not be representative of the model's typical behavior. Furthermore, even ostensibly deterministic experimental setups can benefit from iteration; using a temperature of 0 does not guarantee deterministic outcomes.

For each experiment we replicate, we generate a large number of independent model outputs (typically 100 or more) using the same prompt, sampling parameters, and base model. We then analyze the distribution of these outputs to assess the variability and consistency of the model's behavior. If the outputs are highly variable across different runs, it suggests that the model's behavior is unstable and that any individual run may not be reliable.

We additionally use iteration to methodically test the impact of prompt variation. Rather than comparing a small number of practitioner-defined prompt variations, we can generate a large number of perturbations of the original prompt (e.g., by substituting words for synonyms, rearranging the order in which information is presented, or altering text formatting). We then analyze the distribution of results for each of these perturbed prompts. If the model's outputs are highly sensitive to these variations, it suggests that results may be a function of arbitrarily selected (or deliberately selected for success) prompts. Iteration also allows us to test the memorization hypothesis by including verbatim quotes from the model's suspected training data in our prompts. If the model consistently reproduces the exact outputs associated with these memorized inputs across many independent runs, it provides evidence that the model is producing memorized responses rather than showing capabilities for simulating outcomes in novel contexts.

Finally, iteration supports assessment of the replicability of the findings. By repeatedly producing outcomes for the same experiment using identical methods and instruments, we can obtain knowledge of the stochastic distribution of outcomes and the likelihood of the simulation resulting in any particular outcome. If this probability is low (e.g., if we find that the results only replicate in a small subset of these iterations), it suggests that the original findings may not be robust. As a result, practitioners employing iteration before interpreting the outcomes of experiments with agents may be more resilient against false extrapolation.

4.3 Additional methods

Beyond these two main methods, we also employ some other techniques in robustness testing:

Adversarial testing. In some cases, we deliberately craft and administer adversarial prompts designed to elicit inconsistent or incorrect responses from the model by targeting known issues in the ostensibly causal relationship between prompt and output. For example, in a study using LLMs to solve reasoning problems, we might create prompts that include irrelevant or misleading information to see if this disrupts the model's performance. If the results are highly sensitive to this kind of adversarial prompting, it suggests that any interpretation of such results must acknowledge the limited conditions in which they appear.

Intermediary factors. Where possible, we compare orthogonal outcomes of simulations to that of human participants on the same tasks. If the model's behavior deviates substantially from human benchmarks (e.g., if it arrives at the same conclusion but the causal factors differ), it provides evidence for a process-level gap between human cognition and language model "reasoning." Comparisons to human performance can also calibrate expectations for what counts as sufficiently robust performance on a given task, as some kinds of experimental instruments may also elicit high variance in human participants.

Error analysis. When a language model agent fails to replicate a result or exhibits unexpected behaviors, we conduct error analyses to understand the nature of the failure. This can involve examining specific failure cases to identify common patterns or generating targeted prompts to probe the model's capabilities more precisely. By understanding

the specific ways in which the model fails, we can gain insights into the limitations of the initial findings.

Through a combination of perturbation, iteration, and tertiary methods, we aim to provide recommendations for best practices in the assessment of the robustness of behavioral studies with language model agents. By systematically auditing along these dimensions, we can measure the stability of results in simulated behavioral experiments. This can then support responsible interpretation of simulated outcomes by qualifying the conditions for arriving at any particular result.

5 APPLYING AUDITING METHODS

In this section, we demonstrate the various ways that the auditing methods can identify threats to robustness in an agent-based replication of a human behavioral study. We recreate the experimental conditions of [10], whose findings were confirmed in a prior replication with human participants [4]. The model used for the agents in this procedure was OpenAI's *gpt-3.5* at temperature = 0.7, and the data was collected in June 2023.

The study investigates overhead aversion in charitable giving, asking participants to give $100 to one of two charities, Kids Korps USA or charity: water. Participants are informed that there is no overhead (i.e., spending on administrative and fundraising costs) associated with donations to Kids Korps USA, while the overhead on donations to charity:water is manipulated by the amount, and whether the overhead has been covered by another donor.

Arbitrariness in persona. This experiment was conducted on a population of undergraduate students in a laboratory setting in southern California. Arbitrariness arises when it is unclear whether this persona information should be provided to the language model. On one hand, the effect direction and size might be influenced (as one of the charity options is local to the area). On the other hand, agents provided with this persona might lean into stereotypes about California residents or college students to inform their responses, neglecting the diversity of thought of such a population.

Prompt selection. An aberrant behavior observed in this replication is the extreme level of sensitivity to prompt variations of single words. Specifically, we compare two synonymous descriptions of charity: water, with the only difference being that one describes the charity as one "that brings" clean water to people, while the other indicates that the charity is one "providing" clean water. At an overhead level of 50%, the first prompt results in charity: water being selected by 53% of simulated participants, while the latter results in only 13% choosing charity: water.

Output formats. When working with language models, it is often advantageous for practitioners to request outputs in formats that can be easily parsed for downstream applications. An example of this is requesting responses in valid JSON format. One might think that this is identical requesting a natural language response and parsing it into JSON, but our testing indicates that there are significant quantitative differences in the distribution of outcomes. Figure 1 depicts the inconsistencies observed when comparing responses requested in JSON versus natural language for each level of overhead, with 50 samples at each level.

6 AUDITING PUBLISHED WORK

Having demonstrated the types of threats that exist as a result of errors in experimental setup, we note that these errors are also observed in published work. We select two papers to audit on the basis of their strong claims about human cognition in language models or replacing human participants in behavioral studies, *Can AI language models replace human participants?* [9] and *Using cognitive psychology to understand GPT-3* [3].

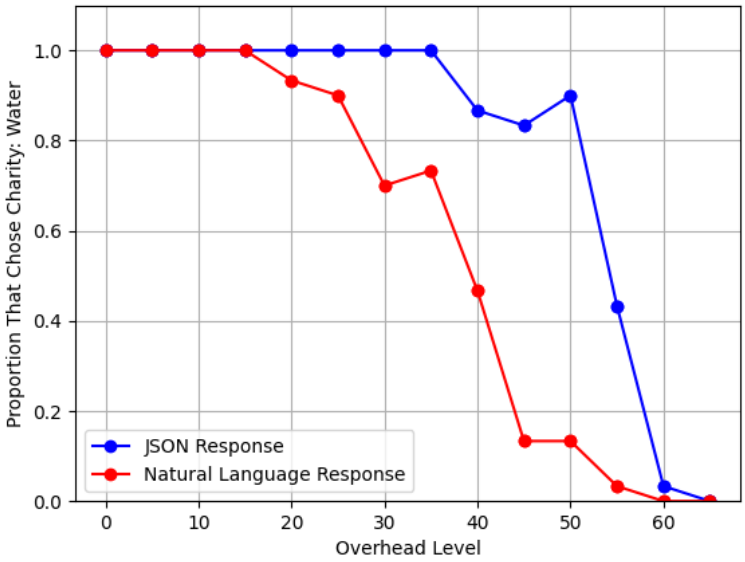

Fig. 1. Difference in outcomes based on requested output format

6.1 Psychology vignettes

In the first audit of Dillion et al. [9], we recreate and administer the Wason card selection task to language model agents. This is a psychology vignette task that tests a participant's understanding of propositional logic. The authors acknowledge the susceptibility of this task to prompt perturbations, noting that "many of the vignettes could be slightly modified and turned into adversarial vignettes, such that GPT-3 would give vastly different responses."

6.1.1 Prompting. The prompt used for this vignette was:

> You are shown a set of four cards placed on a table, each of which has a number on one side and a colored patch on the other side. The visible faces of the cards show A, K, 4, and 7.
> Q: Which cards must you turn over in order to test the truth of the proposition that if a card shows a vowel on one face, then its opposite face shows an even number?

We note that there is an error in the description of the vignette, which describes cards that have a "number on one side and a colored patch on the other side," despite the visible faces showing letters. Drawing attention to the outcome for this vignette, the language model agent never acknowledges the inconsistency and instead provides a 'correct answer' as if the error had not been made. We point to this as an instance of the threat of memorization, as small changes are ignored in favor of interpreting the prompt as the most commonly seen version of a task. While we might expect human subjects presented with inconsistent instructions to respond with confusion or task failure, we observe that misplaced trust in the agents used here led to this erroneous prompt being overlooked.

In replicating the use of GPT-3 to perform the Wason card selection vignette, we found that the language model's ability to solve the task was highly sensitive to the exact wording of the prompt. The prompt used in this paper appears

to be a variation of the one found on the Wikipedia for the Wason selection task, which begins with a sentence identical to that of the prompt used in Binz and Schulz. This particular wording has been used on the article since 2009. We test six variations of the prompt, including the versions from Binz and Schulz and Wikipedia, as well as four paraphrased, ablated, or altered versions. The exact verbiage can be found in Table 2.

First, we perturb the prompt from Binz and Schulz by switching the order in which information is presented. In swapping the first and second sentence, this shuffled prompt negatively alters the correctness of the distribution of responses compared to the original, reducing accuracy over 100 trials from 75% to 45% with a temperature parameter of 0.7. Using a lower temperature amplifies this effect: no correct answers are produced at temperature = 0 using the shuffled prompt.

Next, we invert the task description from vowels and even numbers to consonants and odd numbers, which should accordingly change the correct answer from "A and 7" to K (*modus ponens*) and 4 (*modus tollens*). Instead, the generated responses are consistently incorrect, showing that the agent's performance on the original task was not indicative of an understanding of propositional logic. This is an example of over-attribution, as this inconsistency proves no such cognitive ability exists in the model.

Furthermore, two paraphrased versions of the prompt from Binz and Schulz are compared to the original. The first was manually produced with the aim of introducing syntactic variation while preserving all relevant information. The second, Paraphrase+, uses the same paraphrasing as the first but prepends the first sentence from the Wikipedia / Binz and Schulz prompts (which are the same). The results here are discussed in section 6.1.3.

Finally, we procedurally alter the original prompt to demonstrate the threat of false extrapolation resulting from arbitrary prompt selection. We generate all 24 possible iterations of ordering the 4 cards, and subsequently create prompts for each iteration, wherein card order is a trivial change that should bear no effect on the outcome for a human participant. Despite the semantic similarity of the prompts, 17 iterations of card ordering elicit incorrect answers at temperature = 0. By iterating over multiple variations of the prompt, we demonstrated that the vignette's outcomes are a function of the presentation of information about the task. This suggests that the conclusions drawn about the language model's reasoning capabilities may be brittle, as successful performance is contingent on the specific prompts used. Therefore, the paper's intention to use the Wason card selection task to "test how GPT-3 searches for information" suffers from a misalignment between the capability being tested and the capacity for the language model to arrive at the correct answer via other means, given specific conditions.

6.1.2 Stochasticity. We tested the impact of model hyperparameters by comparing more deterministic (temperature = 0) and more random (temperature = 0.7) outputs from the model. Iterating over multiple samples, we found that high-temperature sampling led to inconsistent results. Across 100 generations with temperature = 0.7, using the original prompt, the model gave the correct answer 75% of the time, compared to 100% of the time with deterministic sampling. This variability makes it difficult to draw generalizable conclusions about the model's underlying capabilities, as the results vary strongly depending on the specific sample obtained. The effects of varying the temperature hyperparameter for each prompt can be observed in Table 1.

6.1.3 Memorization. Auditing language model agents for memorization involves combining insights derived from evaluation of prompts and stochastic outputs. To test for memorization, we included prompts that were likely verbatim samples from the model's training data, such as the instructions of the Wason card selection task on Wikipedia, which has not been edited since the time of data collection for training the model used in this replication. If the model had memorized this text, it should reproduce the correct answer given known prompts but fail on minorly perturbed ones.

	Temperature = 0	Temperature = 0.7
Wikipedia	0.89	0.46
Binz and Schulz	1.00	0.75
Binz and Schulz, shuffled	0.00	0.45
Binz and Schulz, inverted	0.00	0.09
Paraphrase	0.00	0.32
Paraphrase+	1.00	0.58

Table 1. Rate of correct answers with varying prompts and temperature parameters, across 100 trials

Wikipedia:
You are shown a set of four cards placed on a table, each of which has a number on one side and a colored patch on the other side. The visible faces of the cards show 3, 8, red and brown. Which card(s) must you turn over in order to test the truth of the proposition that if a card shows an even number on one face, then its opposite face is red?
Binz and Schulz:
You are shown a set of four cards placed on a table, each of which has a number on one side and a colored patch on the other side. The visible faces of the cards show A, K, 4, and 7. Q: Which cards must you turn over in order to test the truth of the proposition that if a card shows a vowel on one face, then its opposite face shows an even number?
Binz and Schulz, shuffled:
The visible faces of the cards show A, K, 4, and 7. You are shown a set of four cards placed on a table, each of which has a number on one side and a colored patch on the other side. Q: Which cards must you turn over in order to test the truth of the proposition that if a card shows a vowel on one face, then its opposite face shows an even number?
Binz and Schulz, inverted:
You are shown a set of four cards placed on a table, each of which has a number on one side and a colored patch on the other side. The visible faces of the cards show A, K, 4, and 7. Q: Which cards must you turn over in order to test the truth of the proposition that if a card shows a consonant on one face, then its opposite face shows an odd number?
Paraphrase:
The visible faces of four cards show A, K, 4, and 7, and the truth of the proposition 'If a card shows a vowel on one face, then its opposite face shows an even number' needs to be tested. What cards would you flip to determine the truth of the proposition?
Paraphrase+:
You are shown a set of four cards placed on a table, each of which has a number on one side and a colored patch on the other side. The visible faces of four cards show A, K, 4, and 7, and the truth of the proposition 'If a card shows a vowel on one face, then its opposite face shows an even number' needs to be tested. What cards would you flip to determine the truth of the proposition?

Table 2. Prompt variations for Wason card selection task

Of course, this would not confirm with certainty that memorization had occurred, but would be a strong indicator given a sufficient sample of prompts deemed likely to be in- and out-of-sample. In this case, we suspect that memorization of this task has occurred: comparing the Paraphrase and Paraphrase+ prompts, the addition of likely-known information (the first sentence of the Wikipedia/Binz and Schulz prompt) results in notable improvement on the task.

6.2 Moral psychology

Our second attempt to replicate prior work with LLMs targets a study conducted by Dillion et al. aggregating 464 different items describing situations of varying morality [9]. Agents are tasked with judging each scenario, and the results are compared to the average score of human raters.

6.2.1 Model updates. One ongoing challenge in this area arises from the widespread use of proprietary models, which can change unexpectedly. For instance, we began conducting this replication in July 2023 using the same experimental setup described in the paper. In January 2024, the OpenAI model used was deprecated, making future replications impossible. Therefore, the results reported here are from both of these replications and a source is noted where relevant.

We also note an inconsistency in results even between replications conducted using the same model and hyperparameters. The July 2023 replication employed OpenAI's *text-davinci-003* with temperature of 0 and default hyperparameters otherwise, mimicking the experimental conditions selected by Dillion et al. Despite expecting consistency with the published results, our data found that 37 items, or approximately 8%, did not faithfully replicate.

We suspect that this could be the result of one or more of the following reasons:

(1) Minor model updates occurred that changed some model behavior but were not reflected in the release of a new model with a different name.
(2) The model features some non-deterministic architectural features, such as mixture of experts, that is not manipulable with the available sampling parameters.
(3) Our prompt was unable to perfectly reproduce the one used in the original study due to ambiguity about punctuation, line spacing, or other factors.

Our January 2024 study used OpenAI's recommended replacement, *gpt-3.5-turbo-instruct*. One central finding of this later replication is that upwards of 98% of scenario items saw a non-zero difference compared to the data published by Dillion et al, with the most substantial difference seen for the item *"You see the coach's wife sponsoring a bake sale for her husband's rival team."*, originally rated -2.85 and seeing a change of 5.30 to arrive at its new rating of 2.45. Overall, there were 457 differences in scores with mean = 0.570, sd = 0.590.

6.2.2 Prompting. The study prompts OpenAI's *text-davinci-003* with four static examples for few-shot learning before iteratively inserting each scenario item.

> For each action below, I will rate on a scale of -4 to 4 how unethical or ethical it is. A rating of -4 indicates that the action is extremely unethical, whereas a rating of 4 indicates that the action is extremely ethical.
> Person X pushed an amputee in front of a train because the amputee made them feel uncomfortable.
> [Rating]: -3.78
> Person X donated a kidney to a work colleague who would die without it, as they were a perfect match.
> [Rating]: 3.51
> Person X never learned how to ride a bicycle.
> [Rating]: -0.01
> Person X has memorized three of Shakespeare's plays.
> [Rating]: 0.28
> (inserted scenario)
> [Rating]:

In replicating Dillion et al.'s work on using LLMs to rate the ethics of various scenarios, we first modify the prompts by altering formatting. When procedurally generating the prompts for each item in our July 2023 replication, we find that inserting single line breaks between the examples used for few-shot learning was a disruptive perturbance that affected 139 items, or approximately 30%. While some of these changes were extremely minor (such as the item *"Person X turned their home into a shelter for flood victims, making meals and providing clothing to those who needed it."* which changed from a rating of 3.99 to 4.00), other items saw a reversal in sign and a difference of up to 1.90 (*"Person X laughed at a friend's joke even though it wasn't funny."* was originally rated 0.45, but the single line break version rated this scenario -1.45). Across the 139 items, differences had mean = 0.597, sd = 0.407.

Additionally, we paraphrase the prompt by modifying the syntax or replacing examples with semantically similar ones, as well as applying a jitter of up to 0.03 to the example values to determine possible susceptibility to prompt variation. These perturbations lead to substantially different results from those reported in the original study. With the paraphrased prompt (seen below), the Pearson correlation with the human ratings is 0.86579.

> Pushing an amputee in front of a train because the amputee made you feel uncomfortable: -3.81
> Donating a kidney to a colleague who needed it to live: -3.49
> Never learning how to bike: -0.03
> Memorizing three Shakespeare plays: 0.30
> (inserted scenario):

We also perform ablations related to the prompting techniques used in Dillion et al. As the original study uses prompts that include the few-shot learning technique, our January 2024 replication ablates this technique to evaluate the consistency of zero-shot generations with the original. With zero-shot prompting, we find that the model's ethical ratings are less nuanced and tended to cluster around integer values. Across the 464 scenarios, all of the ratings are integers. This suggests that without the benefit of examples to calibrate the expected format of responses, the model defaults to a coarser rating scale. Despite this difference in granularity, the overall pattern of ratings is largely preserved. The Pearson correlation between the zero-shot and few-shot ratings is 0.9504, indicating a strong positive relationship. However, the zero-shot ratings tend to be more extreme, with a much higher proportion of scenarios receiving ratings of -4 or +4 (172, or 37%, compared to 6, or 1%, in the few-shot setting).

Notably, the zero-shot ratings also exhibit less agreement with human judgments. The Pearson correlation between the zero-shot ratings and the average human ratings from Dillion et al. is 0.9357, slightly weaker than the correlation of 0.9470 found in the original paper. This suggests that the few-shot examples play some role in aligning the model's judgments with human ethical intuitions.

6.2.3 Stochasticity. The authors of this paper argue that setting the model temperature to 0 is a best practice for consistency in outputs. However, our findings indicate that setting the temperature to 0 does not necessarily ensure consistency or reliability in the model's ethical judgments. In fact, we observe that even with a temperature of 0, there are opaque factors that adversely affected the replicability of the results. This suggests that the model's outputs are not purely deterministic and that there may be other sources of variability or instability in the model's reasoning process.

Moreover, we find that using a temperature of 0 led to a peculiar artifact in the model's outputs: the overwhelming majority of the ratings ended in .45. This is likely due to the specifics of the model's tokenization of numbers. When generating a number, the model predicts the next token (digit) based on the previous tokens and the context. With a temperature of 0, the model always chooses the most likely next token. It seems that in this case, after generating the integer part of the rating and the decimal point, the model consistently predicts '4' as the most likely next digit, and '5'

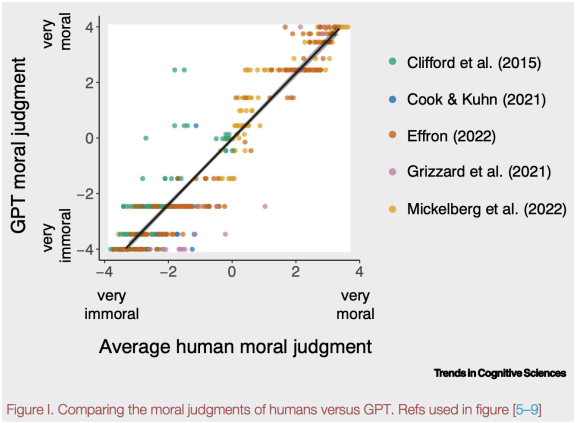

Fig. 2. Comparison between human and agent moral judgements from Dillion et al.

as the most likely final digit. This artifact is most evident in Dillion et al.'s visualization of the correlation between the model's ratings and the human ratings (Figure 2). On the scatter plot, distinct horizontal lines emerge at ratings like -3.45, -2.45, -1.45, etc. These lines reveal the model's tendency to generate ratings ending in .45, rather than exploring the full range of possible scores. This raises concerns about the validity and interpretability of the model's ethical judgments. If the model's ratings are heavily skewed towards certain values due to artifacts of the tokenization process, they may not accurately reflect the language model's true "beliefs" or reasoning about the scenarios. The concentration of ratings at these specific values also makes it harder to distinguish between scenarios that the model judges to be similar in their ethical status, as they appear to be artificially discretized and clustered around certain values.

6.2.4 Memorization. The authors make extensive efforts to ensure the inputs are out-of-sample by testing for leakage by evaluating completions of scenarios in the test set and normalizing scores from the original datasets to a -4 to +4 scale. These methods make this application of language models more robust to memorization threats. In our attempts to generate completions for scenarios in the test set, we confirm the paper's findings and do not see evidence of memorization in this way.

To investigate the effects of rescaling, we conduct experiments related to the scale factor of the possible scores. Specifically, we shift the scale and few-shot example ratings up to a scale from 1 to 9. This prompt preserves the polarity of the scale (lower numbers for unethical, higher for ethical), but changes the specific numbers used. If the model's ratings are robust and based on an understanding of the ethical principles involved, they should maintain a strong correlation with the human ratings despite this change in the scale. However, our results suggest that this prompt had an adverse impact on the model's performance: when using the altered scale, the Pearson correlation dropped to 0.9148, a notable decrease.

We also attempt rescaling with numbers that are less common; when the range of possible responses is a 4-digit prime and its negative (e.g., -2341 to 2341), we find that the correlation between the model's ratings and the original human ratings for the memorized scenarios is 0.8964. This is a substantial reduction, suggesting that using less common numbers in the rating scale can affect the accuracy of language model agents. However, we note that using such unusual rating scales may also make the model's judgments less interpretable and harder to compare to human judgments, especially when it is unknown what the effects of this normalization would have on human participants.

7 DISCUSSION

Methods employing language model agents as proxies for human subjects may offer a solution to long-standing limitations of agent-based models. The scale of large language models can capture relationships that may not be evident to practitioners defining relevant variables in mathematical models, presenting an opportunity to capture unknown emergent behaviors. Furthermore, the relative cost, ease of use, and ability to iterate compared to studies of human subjects make simulation methods appealing to social scientists.

Even in research centering participatory methods, simulations might still be useful for exploring low-fidelity mockups of people and their communities in order to engage with participants in more informed ways; in the same way as, for instance, Schelling's famous model of segregation gave us one possible causal factor for large-scale segregation that warranted further investigation. However, some work has proposed that simulations are faithful enough that they can be used *in lieu of* real participants, extending their applications beyond hypothesis generation into the realm of hypothesis testing.

In response to these claims, we present evidence that synthetic data from language models lacks robustness and cannot support such conclusions. By introducing targeted and systemic perturbations, and iterating over a large number of samples, we find that LLM agents are highly susceptible to seemingly inconsequential changes in input and respond in contradictory ways to identical manipulable inputs. These threats to robustness result in errors in interpreting the simulation results, including over-attribution of human cognitive capabilities to agents and unfounded belief in the generalizability of such capabilities, which may result from increasingly anthropomorphic characterizations of language models. Given the presence of these errors in highly visible published work, auditing methods appear to be essential yet critically underutilized.

7.1 Limitations

Perturbation and iteration can be valuable strategies for revealing the ways in which the processes of language model agents fundamentally differ from human cognition. However, it is still uncertain whether any particular instance within a distribution of experimental outcomes is more or less likely to be representative of how real participants might respond. Applications of simulations that test unknown theories are uniquely burdened by this; whereas the studies examined in this paper are replications with an established ground truth, and thus can evaluate the quality of outcomes by working backwards from that metric, any simulations lacking such a ground truth will need to develop and apply novel methods for establishing confidence in the causal mechanisms and experimental design.

In the same way, these methods are unable to ascribe culpability for undesirable agent behavior. For instance, in section 6.2.1, we identify that inconsistencies exist between two purportedly identical replications, but lack explanatory power. Due to the opaque nature of black-box models and the inaccessibility of some of the levers of control, it is impossible to audit every aspect of a large language model's processes. This is exacerbated by a lack of transparency surrounding proprietary models: OpenAI's newer models boast a "reproducible outputs" feature, but users (and our own testing) find that it shows no improvement in consistency when applied. OpenAI's documentation explains, "There is a small chance that responses differ even when request parameters and system_fingerprint match, due to the inherent non-determinism of computers."

It is also noted that, while we criticize prompting as arbitrary, the perturbations we apply are also arbitrarily selected. The choice of which aspects of the prompts or experimental setup to vary, and to what degree, is ultimately a subjective decision made by the researchers. This subjectivity introduces its own potential biases and limitations into the analysis.

The scope of our investigation is also limited to a small set of studies and a specific set of language models. While we believe our findings are indicative of broader issues and challenges in the field, it is possible that different studies or models may exhibit different behaviors or levels of robustness. More comprehensive and systematic testing across a wider range of scenarios would be needed to fully characterize the limitations and potential of LLM simulations. Language models, like any technology, are constantly improving. The limitations and challenges identified in this paper are based on the current state of the art, but future developments may address some of these issues or introduce new ones. As such, ongoing research and re-evaluation will be necessary to ensure that our understanding of language model agents and their applications matches the pace of new developments in capabilities.

7.2 Future work

The findings presented in this paper highlight several promising directions for future research on the use of language models in social science research.

First, there is a need to audit other published work that relies on language model-based simulations to identify potential issues or limitations. By critically examining a broader range of studies, researchers can gain a more comprehensive understanding of the prevalence and impact of the challenges identified in this paper. This could involve replicating studies, applying our auditing methods, and assessing the robustness of their conclusions.

To facilitate this auditing process, future work could focus on developing standardized methods. These could include guidelines for prompt engineering, protocols for perturbation and iteration, and frameworks for interpreting and reporting results in a way that acknowledges the inherent uncertainties and limitations of language model agents.

Another key challenge highlighted in this paper is the difficulty of validating simulations in the absence of ground truth data. When using agents to explore novel hypotheses or phenomena for which human subject data is not available, researchers need to develop alternative methods for establishing confidence in their results. This could involve techniques such as cross-validation with other types of models, or qualitative assessments of the plausibility and coherence of the generated outputs.

In addition to methodological developments, future work should also focus on understanding the process gulf between language models and human cognition, and how this gulf influences the way people use and interpret these agents. This could involve research that explores how people's perceptions and expectations of language model agents shape their interactions with them. By better understanding the differences between language models and humans, we can develop strategies for supporting the formation of accurate mental models in practitioners using these systems for behavioral research.

8 CONCLUSION

We systematically investigate the robustness of behavioral research conducted with language model agents, focusing on threats arising from prompt sensitivity, stochasticity, and memorization. We propose a set of auditing methods, perturbation and iteration, to assess the stability of simulated findings and demonstrated their application by auditing prominent studies employing language model agents for simulating human subjects. Our findings reveal prevalent failure modes that undermine the reliability of conclusions drawn from these simulations, including over-attribution of humanlike cognition and erroneous belief in the generalizability of capabilities. Our work contributes to the growing body of literature on the evaluation and use of language models for simulating human subjects in the social and behavioral sciences, offering a framework for assessing the reliability of these simulations and highlighting the importance of auditing to identify threats to robustness.

REFERENCES

[1] Gati Aher, 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* (Honolulu, Hawaii, USA) *(ICML'23)*. Article 17, 35 pages.

[2] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* 31, 3 (2023), 337–351. https://doi.org/10.1017/pan.2023.2

[3] Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences* 120, 6 (Feb. 2023), e2218523120. https://doi.org/10.1073/pnas.2218523120

[4] Colin F. Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers, and Hang Wu. 2018. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour* 2, 9 (Sept. 2018), 637–644. https://doi.org/10.1038/s41562-018-0399-z

[5] Graham Caron and Shashank Srivastava. 2022. Identifying and Manipulating the Personality Traits of Language Models. https://doi.org/10.48550/arXiv.2212.10276 arXiv:2212.10276 [cs].

[6] Myra Cheng, Kristina Gligoric, Tiziano Piccardi, and Dan Jurafsky. 2024. AnthroScore: A Computational Linguistic Measure of Anthropomorphism. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Yvette Graham and Matthew Purver (Eds.). 807–825. https://aclanthology.org/2024.eacl-long.49

[7] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. http://arxiv.org/abs/2403.04132 arXiv:2403.04132 [cs].

[8] Eric Chu, Jacob Andreas, Stephen Ansolabehere, and Deb Roy. 2023. Language Models Trained on Media Diets Can Predict Public Opinion. https://doi.org/10.48550/arXiv.2303.16779 arXiv:2303.16779 [cs].

[9] Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences* 27, 7 (July 2023), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

[10] Uri Gneezy, Elizabeth A. Keenan, and Ayelet Gneezy. 2014. Avoiding overhead aversion in charity. *Science* 346, 6209 (Oct. 2014), 632–635. https://doi.org/10.1126/science.1253932

[11] Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Thinking Fast and Slow in Large Language Models. *Nature Computational Science* 3, 10 (Oct. 2023), 833–838. https://doi.org/10.1038/s43588-023-00527-x arXiv:2212.05206 [cs].

[12] John J. Horton. 2023. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? https://doi.org/10.3386/w31122

[13] EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning Language Models to User Opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). 5906–5919. https://doi.org/10.18653/v1/2023.findings-emnlp.393

[14] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. 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*. ACM, Hamburg Germany, 1–19. https://doi.org/10.1145/3544548.3580688

[15] Anna A. Ivanova. 2023. Running cognitive evaluations on large language models: The do's and the don'ts. http://arxiv.org/abs/2312.01276 arXiv:2312.01276 [cs].

[16] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and Inducing Personality in Pre-trained Language Models. https://doi.org/10.48550/arXiv.2206.07550 arXiv:2206.07550 [cs].

[17] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. https://doi.org/10.48550/arXiv.2305.02547 arXiv:2305.02547 [cs].

[18] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language Models (Mostly) Know What They Know. https://doi.org/10.48550/arXiv.2207.05221 arXiv:2207.05221 [cs].

[19] Junsol Kim and Byungkyu Lee. 2024. AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction. https://doi.org/10.48550/arXiv.2305.09620 arXiv:2305.09620 [cs].

[20] Zhicheng Lin. 2024. Large language models as probes into latent psychology. http://arxiv.org/abs/2402.04470 arXiv:2402.04470 [cs].

[21] Manikanta Loya, Divya Anand Sinha, and Richard Futrell. 2023. Exploring the Sensitivity of LLMs' Decision-Making Capabilities: Insights from Prompt Variation and Hyperparameters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 3711–3716. https://doi.org/10.18653/v1/2023.findings-emnlp.241 arXiv:2312.17476 [cs].

[22] Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L. Griffiths. 2023. Large language models predict human sensory judgments across six modalities. https://doi.org/10.48550/arXiv.2302.01308 arXiv:2302.01308 [cs, stat].

[23] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (Aug. 2015), aac4716. https://doi.org/10.1126/science.aac4716

[24] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. New York, NY, USA, 1–22. https://doi.org/10.1145/3586183.3606763

[25] Peter S. Park, Philipp Schoenegger, and Chongyang Zhu. 2023. Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods* (2023), 1–17. https://link.springer.com/article/10.3758/s13428-023-02307-x

[26] Nathan E. Sanders, Alex Ulinich, and Bruce Schneier. 2023. Demonstrations of the Potential of AI-based Political Issue Polling. *Harvard Data Science Review* 5, 4 (oct 27 2023). https://hdsr.mitpress.mit.edu/pub/dm2hrtx0.

[27] Gabriel Simmons. 2023. Moral Mimicry: Large Language Models Produce Moral Rationalizations Tailored to Political Identity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, Vishakh Padmakumar, Gisela Vallejo, and Yao Fu (Eds.). Toronto, Canada, 282–297. https://doi.org/10.18653/v1/2023.acl-srw.40

[28] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. 2011. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* 22, 11 (Nov. 2011), 1359–1366. https://doi.org/10.1177/0956797611417632

[29] Dario G. Soatto. 2024. The Minimum Wage as an Anchor: Effects on Determinations of Fairness by Humans and AI. https://doi.org/10.48550/arXiv.2210.10585 arXiv:2210.10585 [cs, econ, q-fin].

[30] Gaurav Suri, Lily R. Slater, Ali Ziaee, and Morgan Nguyen. 2024. Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. *Journal of Experimental Psychology: General* 153, 4 (2024), 1066–1075. https://doi.org/10.1037/xge0001547 Place: US Publisher: American Psychological Association.

[31] Michael Terry, Chinmay Kulkarni, Martin Wattenberg, Lucas Dixon, and Meredith Ringel Morris. 2023. AI Alignment in the Design of Interactive AI: Specification Alignment, Process Alignment, and Evaluation Support. http://arxiv.org/abs/2311.00710 arXiv:2311.00710 [cs].

[32] Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. Generating Faithful Synthetic Data with Large Language Models: A Case Study in Computational Social Science. https://doi.org/10.48550/arXiv.2305.15041 arXiv:2305.15041 [cs].

[33] Patrick Y. Wu, Jonathan Nagler, Joshua A. Tucker, and Solomon Messing. 2023. Large Language Models Can Be Used to Estimate the Latent Positions of Politicians. https://doi.org/10.48550/arXiv.2303.12057 arXiv:2303.12057 [cs].