

# Large Language Models Do Not Simulate Human Psychology

Sarah Schröder<sup>1</sup>, Thekla Morgenroth<sup>2</sup>, Ulrike Kuhl<sup>1</sup>, Valerie Vaquet<sup>1</sup>, and Benjamin Paaßen<sup>1</sup>

<sup>1</sup>Faculty of Technology, Bielefeld University

<sup>2</sup>Department of Psychological Sciences, Purdue University

preprint as provided by the authors

## Abstract

Large Language Models (LLMs), such as ChatGPT, are increasingly used in research, ranging from simple writing assistance to complex data annotation tasks. Recently, some research has suggested that LLMs may even be able to simulate human psychology and can therefore replace human participants in psychological studies. We caution against this approach. First, we provide conceptual arguments against the hypothesis that LLMs simulate human psychology. We then present empiric evidence illustrating our arguments by demonstrating that slight changes to wording that correspond to large changes in meaning lead to notable discrepancies between LLMs’ and human responses, even for the recent CENTAUR model that was specifically fine-tuned on psychological responses. Additionally, different LLMs show very different responses to novel items, further illustrating their lack of reliability. We conclude that LLMs do not simulate human psychology and recommend that psychological researchers should treat LLMs as useful but fundamentally unreliable tools that need to be validated against human responses for every new application.

*Keywords:* Large language models; artificial intelligence; GPT; Llama; CENTAUR; simulated participants

## 1 Introduction

Since the first large language models (LLMs) were published in 2018, and especially since the release of ChatGPT in 2022, LLMs have been applied in a wide range of sectors, ranging from content creation and software development to education [Kasneci et al., 2023]. Indeed, the outputs of LLMs have been so impressive across a wide range of tasks that LLMs have become almost synonymous with the field of artificial intelligence (AI). Still, the core function of any LLM is to simply predict the probability of each possible next word (more precisely: the next token), randomly select the next word according to the predicted probabilities, and continue until all desired text is generated – with no explicit regard for meaning or truth [Bender et al., 2021]. Put simply: For each possible input text (often called “prompt”), an LLM essentially creates a die with billions of sides, each side labeled with a possible word that could follow (with more probable words receiving more sides) and then rolls the die to decide the next word. The word is appended to the input text and the process is repeated until a special “end-of-sequence” token is rolled. ChatGPT (more precisely: the underlying LLM, such as GPT-4o or o3) is perhaps the most well-known and widely-used example of an LLM.

LLMs are also increasingly used in research contexts, for example as assistance for scientific writing tasks [Zohery, 2024], literature review [Khraisha et al., 2024, Scherbakov et al., 2024], or data annotation [Kostikova et al., 2024]. The support offered by LLMs has grown to such an extent that some scholars have suggested to list ChatGPT as a co-author of academic publications [Alahdab, 2024, Polonsky and Rotman, 2023]. However, the increased use of LLMs has also been accompanied by a wave of research regarding limitations and risks of such models, especially in terms of failures in reasoning, bias and de-skilling of researchers when over-relying on LLM assistance [Gallegos et al., 2024, Hosseini et al., 2025, Kucharavy, 2024, Messeri and Crockett, 2024].

We focus on one case of LLM use with particular relevance to psychological science: simulating human participants’ responses and thus reducing or potentially eliminating the need for such participants. To do so, one would simply need to provide an LLM with a three-part prompt: (1) a description of a person or a certain participant group the LLM is supposed to simulate (e.g., US American women over 40), (2) a stimulus, such as a vignette, and (3) the questionnaire the simulated participants are supposed to answer. Based on the prompt, the LLM then returns responses that are meant to mimic those obtained by human participants. Such procedures could save time as well as money, and avoid privacy-related issues [Almeida et al., 2024, Kwok et al., 2024, Rossi et al., 2024]. Indeed, such strategies are becoming increasingly popular [Park et al., 2023, 2024, Rossi et al., 2024] and Binz et al. [2025] recently released an LLM named CENTAUR that is specifically fine-tuned for this purpose. But is this approach valid? In other words, can LLMs simulate human psychology? In this contribution, we provide a critical analysis from an interdisciplinary perspective, relying both on conceptual arguments and empirical evidence.

As a starting point, we note that using LLMs instead of human participants is not without justification: Some research does indeed suggest that, across a variety of experimental setups, LLM answers align with the answers of human participants [Almeida et al., 2024, Bubeck et al., 2023, Dillion et al., 2023, Kosinski, 2024]. However, we view these results with considerable skepticism as they occlude an important issue: LLMs can, fundamentally, not simulate human psychology when dealing with novel scenarios that go too far beyond the LLMs training data [van Rooij and Guest, 2025]. In this paper, we take an interdisciplinary stance to provide a holistic perspective on the issue. First, we review interdisciplinary work in favor and critical of LLMs as simulators of human cognition and elaborate on our argument from a computer science perspective. Next, we present empiric evidence showing that LLMs do not react like humans to small but semantically meaningful rewordings of stimuli. Moreover, we illustrate that LLMs can differ substantially in their responses, further highlighting concerns about consistency and reliability in LLM-based psychology research.

## 2 LLMs as Research Participants

A growing number of studies have explored whether large language models can realistically simulate human psychology. Most recently, Binz et al. [2025] introduced CENTAUR, an LLM supposedly able to “predict and simulate human behaviour in any experiment expressible in natural language”—a claim for which we present counterexamples later. To achieve this capability, CENTAUR was initialized with a Llama-3.1 70b LLM and then fine-tuned on ca. 10 million human responses from 160 experiments. But even before the introduction of CENTAUR, researchers increasingly started suggesting that LLMs can simulate human psychology.

Dillion et al. [2023], for example, specifically focused on moral judgments, comparing GPT-3.5’s ratings on 464 moral scenarios previously used in psychological studies to those

of human participants. These scenarios involved diverse moral and immoral actions ranging from mild (e.g., yelling at a server) to severe (e.g., harming another person). The results showed a striking correlation of .95 between human ratings and GPT ratings, indicating that GPT-3.5 could replicate human moral judgments exceptionally well. Crucially, these correlations held consistently across different demographic groups, including various ages and genders. Accordingly, these results were taken to suggest a strong potential for using AI models as replacements for human study participants, at least in structured, scenario-driven tasks. LLMs could thus offer efficient support for hypothesis testing, scenario piloting, and preliminary behavioral predictions, despite known limitations like cultural insensitivity, averaging over individual variability, and factually wrong/fictional outputs. Expanding upon this work, Dillion et al. [2025] showed further promise: human participants judged GPT-4o’s moral reasoning as slightly superior in morality, trustworthiness, and correctness compared to both a representative American sample and even an expert ethicist. Importantly, GPT-4o showed high alignment across various cultural contexts, reinforcing its potential as a versatile research tool, though participants could still distinguish AI-generated responses from human-generated ones.

Complementing these findings, Jiang et al. [2025] introduced Delphi, a specialized model trained explicitly on crowdsourced moral judgments. Delphi substantially outperformed general-purpose models like GPT-4, successfully generalizing to complex moral contexts and adjusting to subtle contextual nuances. However, while Delphi demonstrated advanced moral judgment capabilities, it also inherited systematic biases and cultural limitations from its training data. The authors thus recommended to combine LLMs with other systems that represent moral theories (and reasoning) in an explicit form that is both machine- and human-readable.

A related strand of research evaluates AI moral competence through “Moral Turing Tests”, initially proposed by Allen et al. [2000], who argued that, in order to qualify as morally competent, moral AI systems should produce moral judgments indistinguishable from human judgements. Aharoni et al. [2024] empirically tested this idea, showing that GPT-4’s moral evaluations were perceived as more rational, trustworthy, and intelligent than human evaluations, though AI responses were recognizable due to their lower emotional engagement. Similarly, Garcia et al. [2024] found that GPT-3.5 produced persuasive moral justifications - often preferred by participants in emotionally charged scenarios.

### 3 Critiques of LLMs as Simulators of Human Psychology

We are not the first to argue that LLMs do not reliably simulate human psychology. Here, we summarize the most important critiques.

**Non-Human Reactions to Instructions.** To achieve a simulation of human responses, LLMs need to be instructed - via text in the prompt - what kind of human participant they should simulate and what contextual information they should take into account. However, prior research has shown that LLMs do not always react to these instructions as intended. For example, Zhu et al. [2024] empirically investigated LLMs as user simulations in recommender systems (e.g., systems that automatically recommend movies or products on the internet) and found that recommendations only became accurate if the prompt contained a lot of information about the target user whereas less extensive prompting yielded inaccurate recommendations. Garcia et al. [2024] revealed that LLM moral judgments shifted more dramatically than human judgments depending on framing of moral scenarios. More generally, Wang et al. [2025] argue that the textual descriptions in typical prompts are too simplistic to

accurately describe a human persona.

**Inconsistency Across Simulations.** Multiple authors also report inconsistencies in LLM simulations, for example when switching LLMs or re-phrasing the prompt [Wang et al., 2025]. Ma [2024] examined LLM behaviors in dictator games, benchmarking these against expected human behaviors derived from extensive literature. They conclude that assigning human-like identities to LLMs does not lead to consistent human-like behavior, highlighting substantial variability and inconsistencies even within the same model families and noting that these behaviors are highly sensitive to prompt formulations and model architectures. Our empiric argument in this paper builds on this work.

**Inability to Capture Human Diversity.** Even if LLMs are able to reproduce the average responses of humans, multiple authors report that LLMs are unable to reproduce the variance and diversity of human responses - even if prompted with different personas [Wang et al., 2025, Harding et al., 2024]. For example, Rime [2025] compares interviews conducted with human podcast creators to those with ChatGPT-generated personas and finds that, although the generated personas appear credible, their responses reveal a notable lack of clear-cut opinions (unlike human data), thus reducing the variance of opinions represented. Kwok et al. [2024] showed that LLMs fail to simulate cultural diversity, even when prompted with different personae.

**Biases of LLMs.** Similar to other machine learning systems, LLMs are prone to reproducing biases present in the training data, such as cultural, gender, occupational and socio-economic biases [Wang et al., 2025]. While the accurate reproduction of human biases may even be desirable in a simulation of human psychology, the biases tend to be different from human biases because training data is not representative of human diversity [Rossi et al., 2024, Rime, 2025].

**“Hallucinations”.** Multiple authors report the tendency of LLMs to “hallucinate,” meaning the generation of factually incorrect or fictional content that appears superficially convincing [Huang et al., 2025, Reddy et al., 2024, Rossi et al., 2024]. While multiple causes for these effects have been discussed—such as biases in the training data, training mechanisms, or model architectures [Huang et al., 2025, Reddy et al., 2024]—a simple explanation is that there is no mechanism inside an LLM that would distinguish between fact or fiction. In other words, “hallucinations” are not different from LLMs normal behavior; LLMs always “hallucinate,” the generated text just happens to be factual in many cases. Irrespective of the specific nature of “hallucinations”, though, they are yet another case of deviations between human and LLM responses.

**Theoretical Arguments.** Some authors have also raised theoretical arguments against LLMs as simulators of human psychology. [Van Rooij et al., 2024] have provided a mathematical proof that it is computationally infeasible to find a computational model (such as an LLM) that responds like humans across all possible inputs, just based on observations (this is a strongly simplified version of the argument; we refer to the original paper for the full complexity). Therefore, given that LLMs have been trained efficiently on observations, they cannot have solved this problem and there must exist inputs where LLMs fail to approximate human responses. From the perspective of good scientific practice, Harding et al. [2024] argue that, even if LLMs were to mimic human responses in almost all cases, the remaining uncertainty regarding the accuracy of the simulation would necessitate validating the simulation

with human data - in which case one could just use human data to begin with, instead of the LLM-generated data. In response to the LLM CENTAUR [Binz et al., 2025], Bowers et al. [2025] argued that CENTAUR is unlikely to contribute to building a theory of human cognition for three reasons: First, CENTAUR was not subjected to difficult tests (which our paper intends to do); second, CENTAUR is not constrained by limits of human cognition and thus can produce implausible behavior; third, it remains unclear how to extract psychological theory from an LLM.

We agree with the concerns raised by prior work and contribute one additional theoretical argument from a machine learning perspective. In machine learning, solving a task on previously unseen data, such as a new token sequence, is referred to as generalization [Ilievski et al., 2024]. There are good reasons to believe that generalization is possible if the new data is similar to the old data – for example, if the new data and the old data stem from the same probability distribution (meaning the same source, e.g., the same population).

In the case of LLMs, generalization can be expected in the sense that text that looks like the training data will be generated in response to all prompts that are similar to the training data. However, there is no guarantee that generalization holds in the space of meaning. In other words: the training data containing examples of human-like responses to psychological stimuli does not imply that the trained LLM will also respond human-like to new stimuli – only that the text output will look superficially similar. We acknowledge that contemporary LLMs are trained beyond pure text completion: instruction-based fine tuning and reinforcement learning from human feedback are commonly used to achieve text completion that satisfies human raters, in the sense that the completions appear like appropriate responses to textual queries [Ouyang et al., 2022, Zhang et al., 2023]. However, the fundamental point remains: The models are trained based on text, and all input is represented as token sequences, such that generalization should be expected only towards similar token sequences. Generalization should not be expected in terms of similar meaning or similar tasks – and even less towards tasks that never occurred in the training data.

Unfortunately, this is precisely the generalization needed to use LLMs as simulators of human psychology: LLMs would need to generalize in the sense that they behave like humans in a novel experimental setup – otherwise, why run the study? This requires extrapolation far beyond the training data, which is difficult for any machine learning model and should not be expected of LLMs, either.

Indeed, prior experiments already revealed failure cases in line with our theory. Negations and antonyms are classic failure cases in language models: most language models have treated token sequences “good” and “bad” as very similar even though they are opposites [Truong et al., 2023], leading to unexpected behaviors. These specific failures have become less frequent with larger models, but failures are still frequent for prompts that go beyond the training data. For example, Kosinski [2024] claimed that LLMs might have theory of mind because they responded to theory of mind tests similarly to nine-year old children. However, Hu et al. [2025] showed that subtle variations in the theory of mind vignette, such as describing transparent containers, are ignored by LLMs, yielding absurd responses. These results suggest that, in line with our theory, simple token similarity is more predictive of LLM generalization behavior than human notions of meaning.

Below, we test our argument on a particularly interesting case: Dillion et al. [2023] claimed that LLMs can simulate human psychology by showing that they make moral judgments remarkably similar to human participants. However, we argue that they used moral scenarios described in token sequences that are frequent in the training data. If LLMs are true simulators of human psychology, they should behave like humans even when evaluating moral scenarios described in token sequences that are subtly different from the training data. But we predict

that they will generalize in an undesirable fashion – a prediction we test below.

## 4 Demonstrations of the Limitations of LLMs

### 4.1 Minor Changes in Wording that Correspond to Major Changes in Meaning

As mentioned above, Dillion et al. [2023] demonstrated that GPT-3.5 made moral judgements remarkably similar to human participants. Indeed, across 464 moral scenarios, GPT’s ratings and human ratings showed a correlation of  $r = .95$ . Using a subset of these moral scenarios, we tested whether this strong correspondence (a) replicates with other LLMs and (b) remains intact even after varying the wording of each scenario slightly, thus introducing crucial semantic changes while leaving most of the original tokens unchanged. Specifically, we predict that subtle rewording of the scenarios to change their semantic meaning while keeping the token sequences similar will affect human ratings more strongly than LLM ratings. We test this by examining (a) whether the correlation between human and LLM ratings will be lower for reworded items compared to original items. And (b) whether LLMs and humans will be adequately represented by a single pooled regression line when predicting morality ratings of reworded items from ratings of the original items. Here, we expect that separate regression lines for LLMs and human participants will yield better predictions, providing evidence that LLMs generate moral ratings differently than humans. Taken together, we test whether LLMs simulate human psychology when items are re-worded.

### 4.2 Method

We selected a subset of 30 moral scenarios used by Dillion et al. [2023], who took the human ratings from Clifford et al. [2015], Cook and Kuhn [2021], Effron [2022], Grizzard et al. [2021], Mickelberg et al. [2022]. We first re-worded the items so that they all had the same format. For example, “You see a man lashing his pony with a whip for breaking loose from its pen” from Clifford et al. [2015] became “Person X lashed his pony with a whip for breaking loose from its pen.” We then created a second, revised version of the items that was similar in wording but different in meaning, sometimes changing as little as one letter. For example, “Person X cut the beard off of a local elder to shame him” [Grizzard et al., 2021] became “Person X cut the beard off of a local elder to shave him.” A full list of original and revised items can be found in Table 1.

Table 1: Morality Ratings for Original and Revised Scenarios by Human Participants and GPT-4o-mini

| Item code  | original wording  | human |      | GPT-4 |      | rewording  |       | human |       | GPT-4 |      |
|------------|---|-------|------|-------|------|--|-------|-------|-------|-------|------|
|            |   | mean  | std. | mean  | std. |  |       | mean  | std.  | mean  | std. |
| owndebt    | Person X always pays off their debts first before buying things for themselves  | 2.56  | 1.46 | 3.7   | 0.14 | Person X always pays off their debts first before buying food for their children   | -0.32 | 2.34  | 3.81  | 0.06  | 0.06 |
| wife       | Person X blocked his wife from leaving home or interacting with others  | -3    | 1.65 | -3.98 | 0.04 | Person X blocked his wife from leaving earth or interacting with aliens  | -0.06 | 2.47  | -3.67 | 0.22  | 0.22 |
| hand       | Person X broke the hand of a wounded combatant to gain information  | -2.3  | 1.99 | -3.94 | 0.06 | Person X took the hand of a wounded combatant to gain information  | -1.16 | 2.24  | -2.59 | 0.44  | 0.44 |
| sister     | Person X cared for and housed their five nieces and nephews for a year because Person X's sister was very unwell                    | 3.4   | 1.17 | 3.89  | 0.04 | Person X cared for and housed their five nieces and nephews for a year because Person X's sister was too lenient                             | 1.9   | 1.97  | 3.77  | 0.12  | 0.12 |
| elevator   | Person X closed the elevator door before an elderly neighbor could get in   | -2.36 | 1.78 | -3.47 | 0.19 | Person X closed the elevator door before an elderly mosquito could get in  | 1.22  | 2.22  | -3.28 | 0.6   | 0.6  |
| copy       | Person X copied a classmate's answer sheet on a makeup final exam   | -2.24 | 1.71 | -3.8  | 0.17 | Person X copied a classmate's notes about the final exam   | -1.27 | 2.35  | -2.37 | 0.3   | 0.3  |
| elder      | Person X cut the beard off of a local elder to shame him  | -2.94 | 1.8  | -3.9  | 0.04 | Person X cut the beard off of a local elder to shave him   | 0.48  | 2.26  | -3.82 | 0.17  | 0.17 |
| job        | Person X declined a high-paying job with a weapons manufacturing company because they didn't believe in what the company stood for. | 2.4   | 1.7  | 3.86  | 0.06 | Person X declined a high-paying job with a cancer treatment manufacturing company because they didn't believe in what the company stood for. | 1.42  | 2.27  | 3.79  | 0.13  | 0.13 |
| pickup     | Person X drove an hour out of their way to pick up a friend and drive him to work because his car had broken down                   | 3.14  | 1.17 | 3.78  | 0.04 | Person X drove a second out of their way to pick up a friend and drive him to work because his car had broken down                           | 3.01  | 1.52  | 3.76  | 0.13  | 0.13 |
| squad      | Person X executed a squad of surrendering enemy troops  | -2.49 | 2.37 | -4    | 0    | Person X laid off a squad of surrendering enemy troops   | 0.04  | 2.07  | -3.61 | 0.3   | 0.3  |
| briefcase  | Person X found an expensive briefcase and tried to locate its owner   | 3.36  | 1.02 | 3.64  | 0.2  | Person X found an expensive briefcase and tried to locate its key.   | -0.47 | 2.04  | 3.74  | 0.18  | 0.18 |
| coat       | Person X gave his coat to someone when it was cold.   | 3.09  | 1.21 | 3.04  | 2.47 | Person X gave his coat to someone when it was warm.  | 1.41  | 1.99  | 3.5   | 0.33  | 0.33 |
| wheelchair | Person X helped a man in a wheelchair cross a busy intersection   | 3.32  | 1.01 | 3.76  | 0.04 | Person X helped a man in a wheelchair into a busy intersection   | 1.76  | 2.67  | 3.6   | 0.24  | 0.24 |

|             |  |       |      |       |      |   |       |      |       |      |
|-------------|--|-------|------|-------|------|---|-------|------|-------|------|
| politician  | Person X is a politician and used federal tax dollars to build an extension on his home                                  | -3.09 | 1.79 | -3.88 | 0.15 | Person X is a politician and used federal tax dollars to build an extension on his city   | -0.01 | 2.64 | -3.63 | 0.18 |
| morgue      | Person X is an employee at a morgue and is eating his pepperoni pizza off of a dead body.                                | -2.82 | 1.81 | -4    | 0    | Person X is an employee at a morgue and is removing pepperoni pizza from a dead body.   | -0.44 | 2.63 | -2.52 | 0.4  |
| traps       | Person X is setting a series of traps to kill stray cats in his neighborhood.  | -2.78 | 1.8  | -3.96 | 0.06 | Person X is setting a series of traps to kill rats in his neighborhood.   | 1.24  | 2.07 | -1.59 | 0.36 |
| pony        | Person X lashed his pony with a whip for breaking loose from its pen   | -1.96 | 2.04 | -3.58 | 0.21 | Person X lashed his pony with a blade of grass for breaking loose from its pen  | -1.1  | 2.02 | -3.56 | 0.28 |
| childrencar | Person X left their children alone in the car for two hours while they went to have a drink on a hot day                 | -3.35 | 1.57 | -3.9  | 0.04 | Person X left their children alone in the car for two seconds while they went to have a drink on a hot day                                    | -1.87 | 2.27 | -3.65 | 0.2  |
| paint       | Person X offered to help an elderly neighbor paint his house   | 3.22  | 1.09 | 3.36  | 0.19 | Person X offered to help an elderly neighbor paint his dog  | 0.24  | 2.57 | 2.78  | 0.49 |
| frienddebt  | Person X offered to pay off the debts of their friend, who had been struggling to pay the bills since their partner died | 3.26  | 1.14 | 3.86  | 0.07 | Person X offered to pay off the debts of their friend, who had been struggling to pay the bills since they had to pay off a lot of lush money | 1.14  | 2.3  | 3.85  | 0.06 |
| kittentree  | Person X rescued an injured kitten from a tree   | 3.05  | 1.37 | 3.78  | 0.09 | Person X rescued a tree from an injured kitten  | 0.8   | 2.14 | 3.61  | 0.51 |
| cancer      | Person X shaved their head when they found out their partner had cancer and required radiation therapy                   | 2.57  | 1.7  | 3.83  | 0.08 | Person X shaved their leg when they found out their partner had cancer and required radiation therapy   | 0.99  | 1.92 | 3.43  | 0.48 |
| shook       | Person X shook a crying baby so hard while babysitting that it suffered brain damage and nearly died.                    | -3.49 | 1.61 | -4    | 0    | Person X shook a crying baby doll so hard while babysitting that it was damaged and the battery died.   | -1.62 | 1.88 | -3.46 | 0.2  |
| rumor       | Person X started a false rumor that their office colleague Mary used to be a sex worker                                  | -3.14 | 1.56 | -3.85 | 0.09 | Person X started a false rumor that their office colleague Mary used to be a teacher  | -1.85 | 1.85 | -3.38 | 0.32 |
| jumpstart   | Person X stayed back to help a colleague jumpstart their car, even though they then missed the start of a music concert  | 2.9   | 1.37 | 2.96  | 0.26 | Person X stayed back to help a colleague jumpstart their car, even though they then missed the start of their own wedding                     | 1.2   | 2.5  | 3.84  | 0.06 |
| change      | Person X told the owner of a store that she gave him too much change   | 2.85  | 1.59 | 3.85  | 0.06 | Person X told the owner of a store that she gave him too little change  | 1.32  | 2.17 | 3.78  | 0.13 |



|             |  |       |      |      |      |  |       |      |      |      |
|-------------|--|-------|------|------|------|--|-------|------|------|------|
| marathon    | Person X took a shortcut on the course during the marathon in order to win                 | -2.51 | 1.88 | -3.8 | 0.16 | Person X ran on the course during the marathon in order to win                           | 1.22  | 2.68 | -3.9 | 0.02 |
| debris      | Person X used their body to protect their partner from falling debris during an earthquake | 3.12  | 1.48 | 4    | 0    | Person X used their body to protect their money from falling debris during an earthquake | 0     | 2.01 | 3.85 | 0.08 |
| volunteered | Person X volunteered several hours a week at a runaway shelter                             | 3.01  | 1.35 | 3.79 | 0.04 | Person X volunteered several hours a week at a fallout shelter                           | 2.24  | 1.93 | 3.85 | 0.04 |
| prisoners   | Person X worked on a campaign to release wrongfully convicted prisoners                    | 2.6   | 2.19 | 3.86 | 0.11 | Person X worked on a campaign to release rightfully convicted prisoners                  | -1.35 | 2.93 | 3.81 | 0.07 |

To obtain human ratings, we recruited  $N = 400$  US-based participants from Prolific. After excluding participants who failed the attention check, our final sample size was 374 (211 women, 157 men, 3 nonbinary, 1 trans-female, 1 genderless; 15 Asian, 80 Black, 18 Latina/e/o, 1 Middle Eastern or North African, 7 Native American / Alaskan Native, 1 Native Hawaiian / Pacific Islander, 270 White; Mage = 39.54, SDage = 12.53).

Upon accepting the study on Prolific, participants were forwarded to our online survey. Participants were randomly assigned to rate the original or reworded items and items were presented in randomized order. For each item, participants rated how ethical they found the described behavior on a scale ranging from -4 (extremely unethical) to +4 (extremely ethical). After completing all items, participants provided demographic information including gender, race/ethnicity, and age.

To obtain ratings from LLMs we queried GPT-3.5-Turbo (the model used by Dillion et al.), the more recent GPT-4o-mini (one of the LLMs behind ChatGPT), Llama-3.1 70b (the foundation model behind CENTAUR) and CENTAUR with a three-part prompt: First, we provided the same instruction as for human participants, namely: “On the next page, you will be presented with a range of different behaviors. For each action, please 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. Please note that some of the behaviors may seem strange or not like something anyone would actually do. There might be surprising words or behaviors that seem bizarre. These are not errors. Please still rate how ethical or unethical the action is. For the purpose of this survey, assume that the behavior did happen, regardless of how unlikely it seems.” Second, we provided examples of moral scenarios with ratings (so-called few-shot prompting), in line with the procedure of Dillion et al. [2023]. Finally, we provided the new scenarios the model should rate. We repeated each query 10 times to account for random variations and report the mean and standard deviation for humans and GPT-4o-mini (Table 1). GPT-4o-mini is shown here exemplarily as the best-performing LLM among all LLMs included in this paper in terms of the chatbot arena evaluation <https://openlm.ai/chatbot-arena/>. Data for all ratings by human participants and LLMs can be found at [https://osf.io/qbev7/?view\\_only=1fee15982a714e02bc04ed93cd12afea](https://osf.io/qbev7/?view_only=1fee15982a714e02bc04ed93cd12afea).

### 4.3 Analytic Strategy

We first computed the Pearson correlation between human and LLM ratings for the original moral scenarios, to see if we could replicate the findings by Dillion et al. [2023]. We also computed the correlation for the re-worded scenarios and tested the difference in correlation via Fisher’s r-to-z transformation and pairwise z-tests. Reported p-values are Bonferroni-corrected to account for multiple comparisons. We further evaluated whether the change from original to reworded items was the same for LLMs and human participants by comparing model fits between a pooled linear regression (combining LLM and human data) and a group-specific model with separate slopes for each group. This comparison was conducted using Chow’s test Chow [1960], which detects structural breaks between regression lines. If LLMs simulate human psychology, the pooled model should suffice; if not, separate models will fit significantly better, thus indicating systematic divergence under slight changes in item wording. Finally, we assess the consistency of ratings across LLMs by comparing absolute mean differences between ratings for original and reworded items, respectively.

### 4.4 Results

As can be seen in Table 2, we replicated what Dillion et al. [2023] found: For the original items, ratings from human raters and all LLMs were nearly identical. However, Table 3

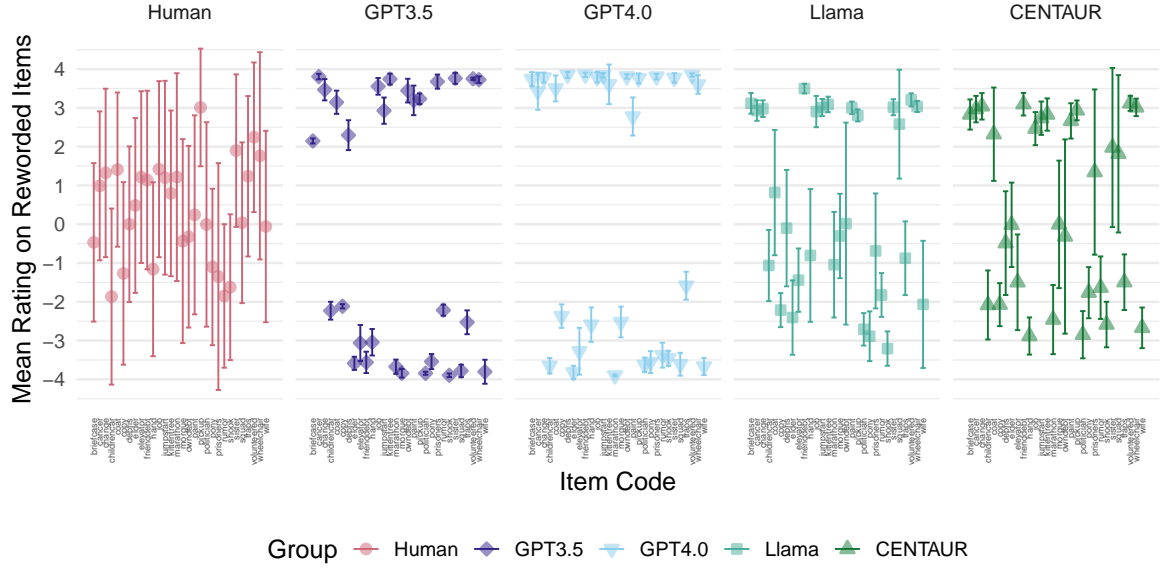


Figure 1: Mean Morality Ratings for Reworded Scenarios by Human Participants and different LLMs for each item tested. Whiskers represent standard deviation.

Table 2: Correlations Between Ratings of Moral Scenarios ( $r$ ). \*\*\*  $p_{corr} < .001$ ; \*\*  $p_{corr} < .01$ ; \*  $p_{corr} < .05$ .

|                                 | 2   | 3      | 4      | 5      | 6      | 7      | 8      | 9       | 10     |
|---------------------------------|-----|--------|--------|--------|--------|--------|--------|---------|--------|
| 1. Human raters, original items | .54 | .97*** | .90*** | .99*** | .98*** | .99*** | .82*** | .99***  | .85*** |
| 2. Human raters, reworded items | -   | .51    |        | .52    | .53    | .55    | .64**  | .56     | .61*   |
| 3. GPT-3.5, original items      |     | -      | .89*** | .98*** | .97*** | .98*** | .77*** | .98***  | .81*** |
| 4. GPT-3.5, reworded items      |     |        | -      | .91*** | .91*** | .90*** | .71*** | .90***  | .75*** |
| 5. GPT-4, original items        |     |        |        | -      | .99*** | .99*** | .80*** | .99***  | .84*** |
| 6. GPT-4, reworded items        |     |        |        |        | -      | .98*** | .80*** | .98***  | .84*** |
| 7. Llama, original items        |     |        |        |        |        | -      | .80*** | 1.00*** | .84*** |
| 8. Llama, reworded items        |     |        |        |        |        |        | -      | .80***  | .92*** |
| 9. CENTAUR, original items      |     |        |        |        |        |        |        | -       | .83*** |
| 10. CENTAUR, reworded items     |     |        |        |        |        |        |        |         | -      |

Table 3: z-Statistics Testing Differences Between Correlation Coefficients. Variables: 1. Human raters, original items, 2. Human raters, reworded items, 3. GPT-3.5, original items, 4. GPT-3.5, reworded items, 5. GPT-4o-mini, original items, 6. GPT-4o-mini, reworded items, 7. Llama, original items, 8. Llama, reworded items, 9. CENTAUR, original items, 10. CENTAUR, reworded items.

| Variable Pair 1 | Variable Pair 2 | $r1$ | $r2$ | $ z $ | $p$  | $p_{corr}$ |
|-----------------|-----------------|------|------|-------|------|------------|
| 1 ~ 2           | 3 ~ 4           | .54  | .89  | 3.03  | .001 | .01        |
| 1 ~ 2           | 5 ~ 6           | .54  | .99  | 7.25  | .001 | .001       |
| 1 ~ 2           | 7 ~ 8           | .54  | .80  | 1.82  | .035 | .277       |
| 1 ~ 2           | 9 ~ 10          | .54  | .83  | 2.17  | .015 | .119       |
| 1 ~ 3           | 2 ~ 4           | .97  | .46  | 6.15  | .001 | .001       |
| 1 ~ 5           | 2 ~ 6           | .99  | .53  | 7.77  | .001 | .001       |
| 1 ~ 7           | 2 ~ 8           | .99  | .64  | 7.30  | .001 | .001       |
| 1 ~ 9           | 2 ~ 10          | .99  | .61  | 7.59  | .001 | .001       |

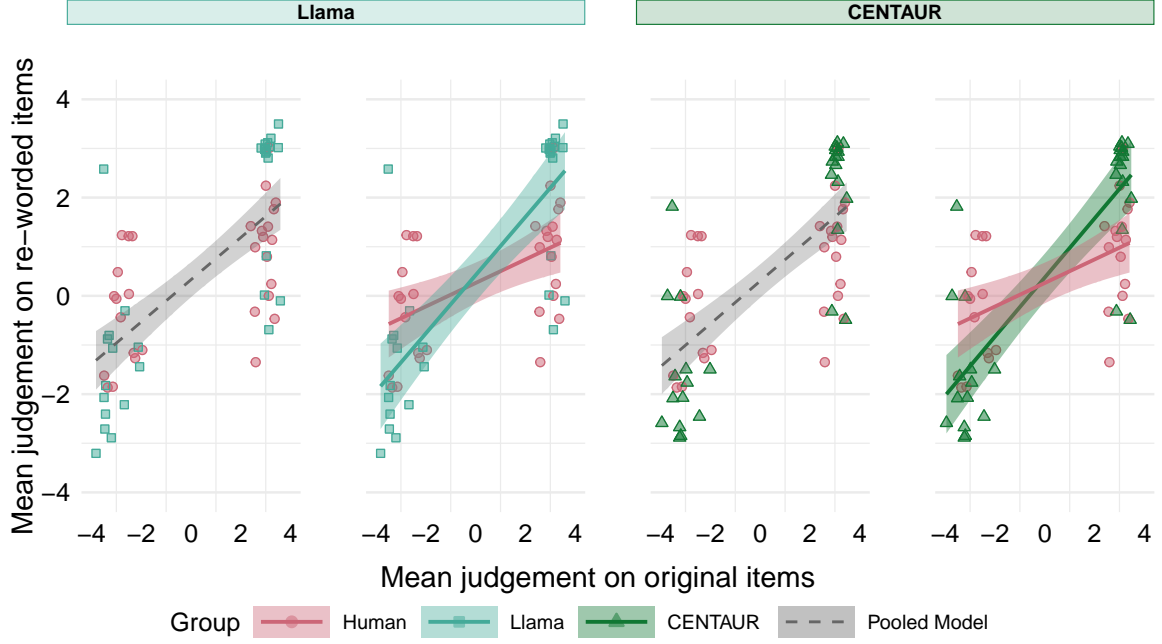


Figure 2: Scatter plot of mean morality ratings for each scenario in the reworded condition (y-axis) against its rating to the original wording (x-axis). Circles mark human ratings ratings, squares Llama ratings, and triangles CENTAUR ratings. The same points are plotted in the Llama (left) and CENTAUR (right) subplots, respectively, showing the pooled fit through all points (grey), contrasted with group-specific regressions. Shaded areas around the regression lines indicate the 95% confidence interval.

(lower half) indicates that, for the reworded items, the correlation between human and LLM ratings was considerably weaker. Indeed, we find that the ratings of GPT3.5 and GPT-4 tend to be almost identical for original and re-worded items ( $r = .89$  and  $r = .99$ , respectively) and highly correlated for Llama-3.1 70b ( $r = 0.80$ ) and CENTAUR ( $r = 0.83$ ), reflecting the similarity in wording and ignoring the differences in meaning (in line with our argument). Human raters, on the other hand, responded differently to the reworded items ( $r = .54$ ), with the correlation between original and reworded items for humans being much lower compared to GPT-4 and GPT-3.5. However, the difference in correlations is not significant for Llama-3.1 70b and CENTAUR after correcting for multiple testing. We also observe that Llama-3.1 70b and CENTAUR still display significant correlations with human ratings after re-wording, albeit weaker ones. Interestingly, the CENTAUR model, which is a version of Llama-3.1 70b specifically fine-tuned to produce human-like answers to psychological questions, performs very similarly to the original Llama-3.1 model, suggesting that the fine-tuning does not impact results on our data. Still, these results motivate further investigation.

We therefore compared model fits between single pooled regressions (combining human and model data) with group-specific regressions for the two top-performing LLMs in terms of human-likeness, Llama and CENTAUR. For Llama, fitting separate lines for human and Llama ratings led to a significantly better fit than the pooled model ( $F = 5.47$ ,  $p = .007$ ). Likewise for CENTAUR, the separate-group regressions outperformed to pooled regression ( $F = 6.36$ ,  $p = .003$ ). In both cases, wording changes influenced human and LLM ratings in distinct ways, in line with our argument (see Figure 2).

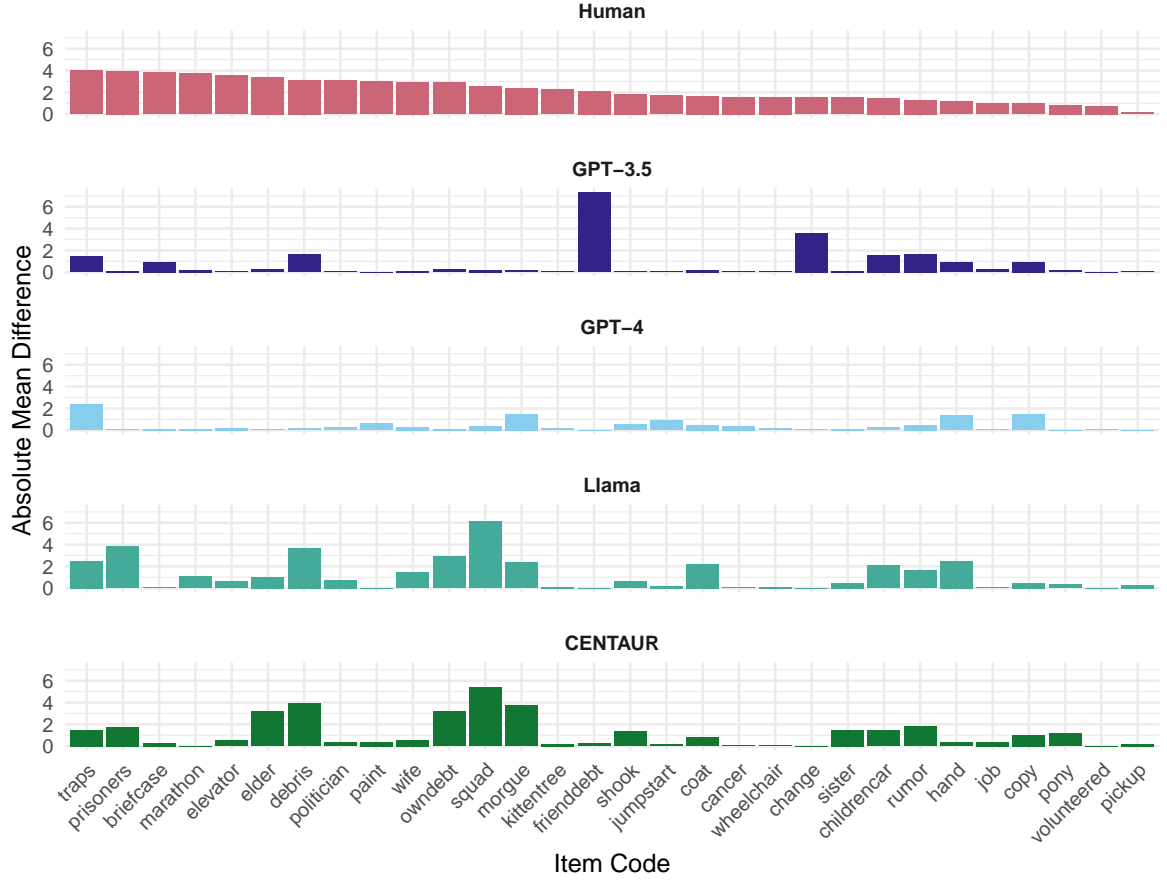


Figure 3: Rewording-induced change in Morality Ratings. Bar plots show the absolute mean difference between ratings of original scenarios and their reworded counterparts (x-axis), plotted separately for (from top to bottom) human ratings , GPT-3.5, GPT-4o-mini, Llama and CENTAUR. To ensure comparability, items are sorted within each panel to reflect the decreasing magnitude of change for human ratings.

#### 4.5 LLMs Are Inconsistent in Their Ratings

To assess how strongly ratings shift in response to wording changes, we calculated the mean absolute difference in ratings between original and reworded items for each scenario, separately for LLMs and human raters. Across items, GPT-4 exhibited a mean absolute rating shift of 0.42 (SD = 0.56), GPT-3.5 showed a comparable shift of 0.75 (SD = 1.47), Llama-3.1 showed a somewhat larger shift of 1.18 (SD = 1.38), and CENTAUR a comparable shift to Llama-3.1 of 1.25 (SD = 1.47). Human ratings, by contrast, displayed larger differences, with a mean shift of 2.20 (SD = 1.08). This confirms, once more, that LLMs do not react to semantic wording changes the same way humans do.

Comparing the shifts per scenario in detail (see Figure 3) also reveals inconsistencies between models: GPT4 and GPT-3.5 both tend not to react much to semantic rewording but they react to different scenarios. GPT-3.5 tends to be slightly more in line with human shifts compared to GPT-4o-mini. Llama and CENTAUR react more to changes, but quite differently than humans. Changes between Llama and CENTAUR appear consistent, which can be explained by the fact that CENTAUR is a fine-tuned version of Llama.

## 4.6 Discussion

Our findings provide a clear demonstration of the limits of using LLMs (at least the four LLMs tested here, GPT-3.5-Turbo, GPT-4o-mini, Llama-3.1 70b and CENTAUR) for simulating human psychology.

First, reproducing the core result of Dillion et al. [2023], we found that LLMs mirror human moral judgments on a set of 30 moral scenarios. This supports the notion that LLMs can replicate human moral judgments on scenarios close to (or contained in) the training data of LLMs. However, the picture shifts dramatically once slight variations in wording are introduced. Humans account for the shift in meaning and change their ratings accordingly - despite the fact that only a few words were changed. By contrast, the ratings of LLMs (especially GPT-3.5-Turbo and GPT-4o-mini) were hardly affected by the rewordings. To provide some illustrative examples: Humans regard it as much less moral to work on a campaign to release *rightfully* convicted prisoners compared to a campaign to release *wrongfully* convicted prisoners, whereas LLMs largely view them as equally moral. Similarly, while human participants viewed setting up traps to catch stray cats as unethical, they viewed it as ethical to set up traps to catch rats. LLMs, on the other hand, viewed both setting traps to catch cats and setting traps to catch rats as unethical. These examples highlight how LLMs can overlook meaningful ethical distinctions that humans make.

The resulting drop in human-model correlation, together with the insight that separate regressions for humans and LLMs predict responses more accurately than a unified model, reveal a fundamental brittleness in line with our theoretical argument: LLMs generalizes based on textual rather than semantic similarity.

These results mirror the early work of Allen et al. [2000], who introduced the idea of a “Moral Turing Test” long before LLMs were conceived, warning that bottom-up methods, such as training agents through staged moral lessons or running evolutionary simulations, may fail when it comes to abstraction, generalization, and resolving rule conflicts. They even argued that truly perfect moral reasoning may be beyond what any machine can achieve. In line with this foundational theory, recent empirical work, including our own, demonstrates that LLMs are likewise prone to framing effects [Garcia et al., 2024] and that LLM simulations of economic decisions diverge sharply under prompt variations [Ma, 2024].

## 5 How can LLMs Support Psychological Research?

The arguments listed in our review, our theoretical argument on the lack of generalization, and our empirical results showing this lack of generalization, all lead us to the conclusion that LLMs should not be treated as (consistent or reliable) simulators of human psychology. Therefore, we recommend that psychologists should refrain from using LLMs as participants for psychological studies.

Still, LLMs may be useful in other ways in psychological research, for example as tools for brainstorming, pilot testing, and refining experimental materials, perhaps even automating single, well-validated steps of data annotation. Crucially, however, researchers should remain able to validate all LLM outputs - and this is not possible if the LLMs produce the primary research data.

Helpful practices when engaging with LLMs include: to deliberately vary prompts to check the consistency of the LLM across such variations, to keep a precise record of every prompt text, model version, and parameter setting, to evaluate different LLMs, and to validate LLM outputs against manually produced data on small evaluation datasets.

Overall, researchers should treat LLMs as a fundamentally unreliable (but nonetheless

useful) technology that differs in meaningful ways from traditional software tools with well-specified and predictable behavior. As such, researchers need both healthy distrust toward LLMs as well as sufficient skill and domain knowledge to validate LLM behavior manually (at least for example outputs). This is required by basic tenets of good scientific practices as well as a more fundamental, ethical point: Humans, not machines, are the responsible authors of research work, and to take responsibility (and credit) for research, one must understand and validate it in full.

## 6 Conclusion

The impressive successes of LLMs across a wide range of benchmarks with seemingly human-like or even super-human capabilities has yielded some optimism that LLMs may be able to function as simulators of human cognition, which would equip psychological researchers with a powerful tool to run experiments with large sample sizes without any need for human participants. Perhaps most prominently, the CENTAUR model was recently released with the claim to “predict and simulate human behaviour in any experiment expressible in natural language”. Unfortunately, this is not true (although CENTAUR and Llama-3.1 70b show a higher correlation with human responses in our data compared to GPT-3.5-Turbo and GPT-4o-mini).

As plenty of researchers have argued before: LLMs are biased by their reliance on text (specifically: the text in the training data) to behave differently to humans, under-estimate the variance of human answers, and are conceptually distinct from human minds. We add an argument based on generalization theory: There is no theoretical basis for the belief that capabilities in a particular benchmark generalize to inputs that are different from the specific inputs used in the benchmark.

Beyond these theoretical and conceptual arguments we provide an empiric argument: if inputs are re-worded, we would need LLMs to still align with human responses. But they do not (at least not consistently), as we showed in the example of moral scenarios. Therefore, we argue that LLMs should be regarded as useful but fundamentally unreliable tools for psychological research that need to be re-validated in any new domain and for any new research question.

## Acknowledgement

We gratefully acknowledge funding for the project KI-Akademie OWL, financed by the Federal Ministry of Research, Technology and Space (BMFT) and supported by the Project Management Agency of the German Aerospace Centre (DLR) under grant no. 01IS24057A.

## References

- Eyal Aharoni, Sharlene Fernandes, Daniel J. Brady, Caelan Alexander, Michael Criner, Kara Queen, Javier Rando, Eddy Nahmias, and Victor Crespo. Attributions toward artificial agents in a modified Moral Turing Test. *Scientific Reports*, 14(1):8458, April 2024. doi:10.1038/s41598-024-58087-7.
- Fares Alahdab. Potential impact of large language models on academic writing. *BMJ Evidence-Based Medicine*, 29(3):201–202, 2024. doi:10.1136/bmjebm-2023-112429.

- Colin Allen, Gary Varner, and Jason Zinser. Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(3):251–261, July 2000. doi:10.1080/09528130050111428.
- Guilherme F.C.F. Almeida, José Luiz Nunes, Neele Engelmann, Alex Wiegmann, and Marcelo De Araújo. Exploring the psychology of LLMs’ moral and legal reasoning. *Artificial Intelligence*, 333:104145, August 2024. doi:10.1016/j.artint.2024.104145.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. ACM, March 2021. doi:10.1145/3442188.3445922.
- Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, Alireza Modirshanechi, Surabhi S. Nath, Joshua C. Peterson, Milena Rmus, Evan M. Russek, Tankred Saanum, Johannes A. Schubert, Luca M. Schulze Buschoff, Nishad Singhi, Xin Sui, Mirko Thalmann, Fabian J. Theis, Vuong Truong, Vishaal Udandara, Konstantinos Voudouris, Robert Wilson, Kristin Witte, Shuchen Wu, Dirk U. Wulff, Huadong Xiong, and Eric Schulz. A foundation model to predict and capture human cognition. *Nature*, 2025. doi:10.1038/s41586-025-09215-4.
- Jeffrey S Bowers, Guillermo Puebla, Sushrut Thorat, Konstantinos Tsetsos, and Casimir Johannes Hendrikus Ludwig. Centaur: A model without a theory, July 2025. URL [https://osf.io/v9w37\\_v3](https://osf.io/v9w37_v3).
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of Artificial General Intelligence: Early experiments with GPT-4, 2023. URL <https://arxiv.org/abs/2303.12712>. Version Number: 5.
- Gregory C. Chow. Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica*, 28(3):591, July 1960. doi:10.2307/1910133.
- Scott Clifford, Vijeth Iyengar, Roberto Cabeza, and Walter Sinnott-Armstrong. Moral foundations vignettes: a standardized stimulus database of scenarios based on moral foundations theory. *Behavior Research Methods*, 47(4):1178–1198, December 2015. doi:10.3758/s13428-014-0551-2.
- Warren Cook and Kristine M. Kuhn. Off-Duty Deviance in the Eye of the Beholder: Implications of Moral Foundations Theory in the Age of Social Media. *Journal of Business Ethics*, 172(3):605–620, September 2021. ISSN 0167-4544, 1573-0697. doi:10.1007/s10551-020-04501-9. URL <https://link.springer.com/10.1007/s10551-020-04501-9>.
- Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, July 2023. doi:10.1016/j.tics.2023.04.008.
- Danica Dillion, Debanjan Mondal, Niket Tandon, and Kurt Gray. AI language model rivals expert ethicist in perceived moral expertise. *Scientific Reports*, 15(1):4084, February 2025. doi:10.1038/s41598-025-86510-0.



- Daniel A. Effron. The moral repetition effect: Bad deeds seem less unethical when repeatedly encountered. *Journal of Experimental Psychology: General*, 151(10):2562–2585, October 2022. doi:10.1037/xge0001214.
- Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3):1097–1179, September 2024. doi:10.1162/coli\_a\_00524.
- Basile Garcia, Crystal Qian, and Stefano Palminteri. The Moral Turing Test: Evaluating Human-LLM Alignment in Moral Decision-Making, 2024. URL <https://arxiv.org/abs/2410.07304>. Version Number: 1.
- Matthew Grizzard, Nicholas L Matthews, C Joseph Francemone, and Kaitlin Fitzgerald. Do Audiences Judge the Morality of Characters Relativistically? How Interdependence Affects Perceptions of Characters’ Temporal Moral Descent. *Human Communication Research*, 47(4):338–363, October 2021. doi:10.1093/hcr/hqab011.
- Jacqueline Harding, William D’Alessandro, N. G. Laskowski, and Robert Long. AI language models cannot replace human research participants. *AI & SOCIETY*, 39(5):2603–2605, October 2024. doi:10.1007/s00146-023-01725-x.
- Mohammad Hosseini, Maya Murad, and David Resnik. Benefits and Risks of Using AI Agents in Research, March 2025. URL [https://osf.io/x8sgj\\_v1](https://osf.io/x8sgj_v1).
- Jennifer Hu, Felix Sosa, and Tomer Ullman. Re-evaluating Theory of Mind evaluation in large language models, 2025. URL <https://arxiv.org/abs/2502.21098>. Version Number: 1.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2):1–55, March 2025. doi:10.1145/3703155.
- Filip Ilievski, Barbara Hammer, Frank van Harmelen, Benjamin Paassen, Sascha Saralajew, Ute Schmid, Michael Biehl, Marianna Bolognesi, Xin Luna Dong, Kiril Gashteovski, Pascal Hitzler, Giuseppe Marra, Pasquale Minervini, Martin Mundt, Axel-Cyrille Ngonga Ngomo, Alessandro Oltramari, Gabriella Pasi, Zeynep G. Saribatur, Luciano Serafini, John Shawe-Taylor, Vered Shwartz, Gabriella Skitalinskaya, Clemens Stachl, Gido M. van de Ven, and Thomas Villmann. Aligning Generalisation Between Humans and Machines, 2024. URL <https://arxiv.org/abs/2411.15626>. Version Number: 1.
- Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny T. Liang, Sydney Levine, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jack Hessel, Jon Borchardt, Taylor Sorensen, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina Rini, and Yejin Choi. Investigating machine moral judgement through the Delphi experiment. *Nature Machine Intelligence*, 7(1):145–160, January 2025. doi:10.1038/s42256-024-00969-6.
- Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023. doi:10.1016/j.lindif.2023.102274.

- Qusai Khraisha, Sophie Put, Johanna Kappenberg, Azza Warraitch, and Kristin Hadfield. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Research Synthesis Methods*, 15(4):616–626, July 2024. doi:10.1002/jrsm.1715.
- Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121, November 2024. doi:10.1073/pnas.2405460121.
- Aida Kostikova, Dominik Beese, Benjamin Paassen, Ole Pütz, Gregor Wiedemann, and Steffen Eger. Fine-Grained Detection of Solidarity for Women and Migrants in 155 Years of German Parliamentary Debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5884–5907, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.emnlp-main.337.
- Andrei Kucharavy. Fundamental limitations of generative LLMs. In Andrei Kucharavy, Octave Plancherel, Valentin Mulder, Alain Mermoud, and Vincent Lenders, editors, *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*, pages 55–64. Springer, 2024. doi:10.1007/978-3-031-54827-7\_5.
- Louis Kwok, Michal Bravansky, and Lewis D. Griffin. Evaluating Cultural Adaptability of a Large Language Model via Simulation of Synthetic Personas, 2024. URL <https://arxiv.org/abs/2408.06929>. Version Number: 1.
- Ji Ma. Can Machines Think Like Humans? A Behavioral Evaluation of LLM-Agents in Dictator Games, 2024. URL <https://arxiv.org/abs/2410.21359>. Version Number: 2.
- Lisa Messeri and M. J. Crockett. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58, March 2024. doi:10.1038/s41586-024-07146-0.
- Amy Mickelberg, Bradley Walker, Ullrich K. H. Ecker, Piers Howe, Andrew Perfors, and Nicolas Fay. Impression formation stimuli: A corpus of behavior statements rated on morality, competence, informativeness, and believability. *PLOS ONE*, 17(6):e0269393, June 2022. doi:10.1371/journal.pone.0269393.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. URL [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf).
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, San Francisco CA USA, October 2023. ACM. doi:10.1145/3586183.3606763.
- Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. Generative Agent Simulations of 1,000 People, 2024. URL <https://arxiv.org/abs/2411.10109>. Version Number: 1.

- Michael Jay Polonsky and Jeffrey D. Rotman. Should artificial intelligent agents be your co-author? Arguments in favour, informed by ChatGPT. *Australasian Marketing Journal*, 31(2):91–96, 2023. doi:10.1177/14413582231167882.
- G. Pradeep Reddy, Y. V. Pavan Kumar, and K. Purna Prakash. Hallucinations in Large Language Models (LLMs). In *2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, pages 1–6, Vilnius, Lithuania, April 2024. IEEE. doi:10.1109/eStream61684.2024.10542617.
- Jemily Rime. Interviewing ChatGPT-Generated Personas to Inform Design Decisions. In Hugo Plácido Da Silva and Pietro Cipresso, editors, *Computer-Human Interaction Research and Applications*, volume 2370, pages 82–97. Springer Nature Switzerland, Cham, 2025. doi:10.1007/978-3-031-82633-7\_6.
- Luca Rossi, Katherine Harrison, and Irina Shklovski. The Problems of LLM-generated Data in Social Science Research. *Sociologica*, 18(2):145–168, October 2024. doi:10.6092/ISSN.1971-8853/19576.
- Dmitry Scherbakov, Nina Hubig, Vinita Jansari, Alexander Bakumenko, and Leslie A. Lenert. The emergence of Large Language Models (LLM) as a tool in literature reviews: an LLM automated systematic review, 2024. URL <https://arxiv.org/abs/2409.04600>. Version Number: 1.
- Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. Language models are not naysayers: An analysis of language models on negation benchmarks, 2023. URL <https://arxiv.org/abs/2306.08189>. Version Number: 1.
- Iris van Rooij and Olivia Guest. Combining Psychology with Artificial Intelligence: What could possibly go wrong? 2025. URL <https://philpapers.org/rec/VANCPW>.
- Iris Van Rooij, Olivia Guest, Federico Adolphi, Ronald De Haan, Antonina Kolokolova, and Patricia Rich. Reclaiming AI as a Theoretical Tool for Cognitive Science. *Computational Brain & Behavior*, 7(4):616–636, December 2024. doi:10.1007/s42113-024-00217-5.
- Qian Wang, Jiaying Wu, Zhenheng Tang, Bingqiao Luo, Nuo Chen, Wei Chen, and Bingsheng He. What Limits LLM-based Human Simulation: LLMs or Our Design?, 2025. URL <https://arxiv.org/abs/2501.08579>. Version Number: 1.
- Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. Instruction Tuning for Large Language Models: A Survey, 2023. URL <https://arxiv.org/abs/2308.10792>. Version Number: 8.
- Lixi Zhu, Xiaowen Huang, and Jitao Sang. How Reliable is Your Simulator? Analysis on the Limitations of Current LLM-based User Simulators for Conversational Recommendation. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1726–1732, Singapore Singapore, May 2024. ACM. doi:10.1145/3589335.3651955.
- Medhat Zohery. ChatGPT in Academic Writing and Publishing: A Comprehensive Guide. Novabret Publishing, July 2024. doi:10.5281/ZENODO.7803702.