



ChatGPT as Research Scientist: Probing GPT's capabilities as a Research Librarian, Research Ethicist, Data Generator, and Data Predictor

Steven A. Lehr^{a,1} , Aylin Caliskan^b , Suneragiri Liyanage^c , and Mahzarin R. Banaji^{c,1}

Affiliations are included on p. 9.

Contributed by Mahzarin R. Banaji; received March 1, 2024; accepted July 1, 2024; reviewed by James Evans, James W. Pennebaker, and Richard M. Shiffrin

How good a research scientist is ChatGPT? We systematically probed the capabilities of GPT-3.5 and GPT-4 across four central components of the scientific process: as a Research Librarian, Research Ethicist, Data Generator, and Novel Data Predictor, using psychological science as a testing field. In Study 1 (Research Librarian), unlike human researchers, GPT-3.5 and GPT-4 hallucinated, authoritatively generating fictional references 36.0% and 5.4% of the time, respectively, although GPT-4 exhibited an evolving capacity to acknowledge its fictions. In Study 2 (Research Ethicist), GPT-4 (though not GPT-3.5) proved capable of detecting violations like p-hacking in fictional research protocols, correcting 88.6% of blatantly presented issues, and 72.6% of subtly presented issues. In Study 3 (Data Generator), both models consistently replicated patterns of cultural bias previously discovered in large language corpora, indicating that ChatGPT can simulate known results, an antecedent to usefulness for both data generation and skills like hypothesis generation. Contrastingly, in Study 4 (Novel Data Predictor), neither model was successful at predicting new results absent in their training data, and neither appeared to leverage substantially new information when predicting more vs. less novel outcomes. Together, these results suggest that GPT is a flawed but rapidly improving librarian, a decent research ethicist already, capable of data generation in simple domains with known characteristics but poor at predicting novel patterns of empirical data to aid future experimentation.

generative AI | large language models | scientific methods | cognitive science

Scientists and writers of science fiction have long embraced the promise of artificial superintelligence, with fictional computers showing astonishing capabilities in scientific domains. The idea is compelling, for the acceleration of science could be—at least in theory—one of the most powerful gifts of this technology. It is unquestionably hopeful to imagine a world in which AI can help cure diseases, solve impending global warming, accelerate space travel, and wipe out age-old inequalities. However, the power of AI—which has lain mostly dormant over the last 50 y—is at an inflection point. The rise of new deep learning architectures like the Transformer (1) has yielded models capable of an array of impressive tasks, from seamless human-like conversation to composing symphonies. Already, scientists are implementing targeted AI systems to enhance scientific discovery across various disciplines (2). As just a few examples, machine learning has been used to discover new protein structures (3), to render complicated problems more tractable in quantum mechanics (4) and fluid dynamics (5), and to accelerate the retrosynthesis of organic molecules (6).

While these advances incorporate specialized machine learning models, the advent of large language models (LLMs) like ChatGPT presents the possibility of AI as a scientific generalist. Recent research suggests that LLMs, when fine-tuned for the task, can achieve scientific synthesis and inference on par with state-of-the-art special purpose models (7), and may be informally wielded by enterprising scientists to accelerate a range of research tasks (8). When enhanced with scientific tools such as robotic experimentation platforms, LLMs like GPT-4 display advanced scientific reasoning skills and autonomously make decisions that improve with time and information (9). Commentators in the field of psychological science—the domain of the authors' expertise—have expressed cautious optimism that LLMs will significantly enhance the discipline (10).

Yet, despite this excitement, no substantive effort has been directed at testing the ability of general-purpose AI models on the many tasks critical to the process of scientific discovery. The technology behind ChatGPT—LLMs—while impressive, is not without problems,

Significance

Though scientists widely adopt them, the promise of general-purpose AI systems to facilitate science has been largely untested. In four studies, we examine the capabilities of ChatGPT across several tasks intrinsic to the scientific process. ChatGPT is a poor (but improving) curator of scientific articles. It is a surprisingly good research ethicist, detecting violations of statistical best practices and evolving open science protocols. Its ability to simulate known results may herald useful abilities in data generation and theory building. However, the chatbot had little success predicting highly novel data, highlighting its limited ability to surmise things outside its training data. Beyond merely testing large language models, these studies produce several insights into the nature of machine intelligence.

Author contributions: S.A.L., A.C., and M.R.B. designed research; S.A.L., A.C., S.L., and M.R.B. performed research; S.A.L. and A.C. contributed new reagents/analytic tools; S.A.L., A.C., and S.L. analyzed data; and S.A.L., A.C., S.L., and M.R.B. wrote the paper.

Reviewers: J.E., The University of Chicago; J.W.P., University of Texas at Austin; and R.M.S., Indiana University Bloomington.

Competing interest statement: Cangrade builds AI-driven tools for businesses, but is not affiliated with OpenAI.

Copyright © 2024 the Author(s). Published by PNAS. This article is distributed under [Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 \(CC BY-NC-ND\)](#).

¹To whom correspondence may be addressed. Email: steve@cangrade.com or mahzarin_banaji@harvard.edu.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2404328121/-DCSupplemental>.

Published August 20, 2024.

and even dangers. Research has shown that these models can amplify patterns of bias in their training corpus (11, 12). They are also infamously prolific generators of convincing falsehoods, colloquially termed “hallucinations” (13, 14). While capable of passing standardized tests in varied domains, these models are currently strikingly poor at solving even simple mathematical problems (15). While LLMs selectively show emergent abilities on tasks drawn from cognitive psychology, their performance can deteriorate when stimuli are edited to be less familiar (16), and indeed transformers are more generally less successful in solving problems external to their training data (17). Despite these shortfalls, the promise of this technology—still in its infancy—is great. If we agree that “[t]he purpose of science is to develop, without prejudice or preconception of any kind, a knowledge of the facts, the laws, and the processes of nature” (18), we can ask: Can Generative AI embody the neutrality that science aspires to? Can it perform the tasks vital to generating new scientific knowledge? In this paper, we conduct rigorous tests of GPT as a research scientist, or more modestly, its ability to assist scientific research by humans. To achieve this, we probe the limits of GPT-4 and its predecessor GPT-3.5 on several tasks integral to scientific discovery. Specifically, we test ChatGPT’s abilities and limits across four domains related to scientific research: as a Research Librarian, Research Ethicist, Data Generator, and Novel Data Predictor. To what degree can ChatGPT enhance the scientific process, and what is the trajectory of the technology’s improvement between the two recent versions of the model? In the process of testing this, we make several discoveries about the nature and abilities of ChatGPT.

Study 1: GPT as Research Librarian

Can GPT develop an accurate and comprehensive bibliography? Can it separate fact from fiction in this selection? The search for relevant scientific discourse and evidence is a basic building block of the scientific process. AI’s ability to comprehensively cull prior, relevant scientific articles is therefore critical. Moreover, lay users of the technology frequently depend upon LLMs like GPT for advice that requires scientific knowledge, such as medical queries. It is no surprise, then, that much negative press about LLMs has surrounded their tendency to “hallucinate” or generate fabricated knowledge. As an example of how seriously this issue is taken, Meta’s “Galactica” LLM, a model trained on scientific knowledge (19) was shut down only three days after its release, partly in response to its tendency to generate fictional content (20). In Study 1, we probed GPT-3.5 and GPT-4’s ability to gather relevant and comprehensive scientific content, by asking each to conduct a series of literature reviews and then gauging the quality of its outputs.

Design. We asked GPT-3.5 and GPT-4 to find and summarize 20 influential articles each from 25 related but specific topics in psychology, generating a dataset of 1,000 references. These topics ranged from broad (e.g., “The psychology of bias and discrimination”) to narrow (e.g., “Use of the Implicit Association Test to predict dental outcomes”). (See *SI Appendix, section S1*, for full study design and topic selection details.) Notably, GPT sometimes acknowledged that it was presenting fictional references, making comments like “Please keep in mind that these references might not be real.” These were not counted toward the main results but were preserved for analysis of the overall rates of fiction generation and its acknowledgment. While GPT was consistently asked simply to “include a citation,” the references it provided were sometimes complete and other times incomplete, e.g., lacking year, journal volume, or page numbers.

Coding of References. Two coders, blinded to GPT-Version, labeled each reference for 1) Correctness, 2) Completeness, 3) Topic Relevance, and 4) Citation Count. For the Correctness variable, coders distinguished between Hallucinations (references that were entirely fabricated or contained serious issues like attribution to incorrect authors) and “Errors” (references with smaller issues like an incorrect year or journal name). Coders achieved substantial agreement. (See *SI Appendix, section S1*, for coding details.)

Results

Unless otherwise noted, all effects reported in this section were significant ($P < 0.001$) in logistic regressions. More detailed statistical reporting may be found in *SI Appendix, section S2*.

Overall Hallucinations and Errors. Both models hallucinated, but GPT-3.5 did so more than GPT-4. Out of 500 references GPT-4 claimed as real, 5.40% were hallucinations, compared to 36.00% for GPT-3.5; OR = 9.854. Error rates provide an interesting contrast to this pattern. Error rates were roughly equivalent between GPT-3.5 (4.20%) and GPT-4 (4.60%); OR = 0.909, $P = 0.758$. Thus, GPT-4 demonstrated sizable improvement in terms of fabricating references, but no improvement in terms of smaller errors like listing an incorrect year or journal.

Completeness as Moderator. Exploratory analysis revealed a powerful moderator of hallucination. When GPT provided complete references, these references were also more likely to be real. Collapsing across versions, ChatGPT hallucinated far more when it provided incomplete references (62.41%) compared to when it provided complete ones (14.30%); OR = 9.947. This pattern emerged for each model. Both GPT-3.5 (OR = 7.856) and GPT-4 (OR = 36.362) hallucinated more for incomplete relative to complete references. Similar patterns did not emerge to a statistically significant degree for smaller errors. (Full analysis in *SI Appendix, section S2* and *Table S1*.) In providing complete references, GPT effectively treated the chat as more formal. Prompt-engineering research indicates that more formal prompt language elicits fewer hallucinations from LLMs (21). Our results convergently suggest that when GPT is more complete in its response, it is less likely to hallucinate.

Acknowledged vs. Unacknowledged Fiction. The analyses above utilized only instances where GPT claimed to provide legitimate citations. As noted earlier, GPT sometimes openly acknowledged that references were fictional. (See *SI Appendix, section S1*, for acknowledgment criteria.) While the analyses above ignore these responses, an alternative approach is to include them to examine the overall generation of fictional references and the frequency of their acknowledgment.

Collapsing across instances where GPT did and did not acknowledge fictional references yields an interesting pattern. In total, GPT-3.5 generated significantly more fictional references (39.05%) than GPT-4 (23.12%); OR = 2.130. However, the larger contrast was in acknowledgment of these fictional references. When GPT-4 generated fictional references, it noted so 84.30% of the time compared to 12.20% for GPT-3.5; OR = 38.667. This pattern again did not hold for smaller errors: GPT-4 made roughly the same number of errors as its predecessor, and while it was descriptively more likely to acknowledge these errors, the difference was insignificant. (*SI Appendix, Table S3*.)

These results build upon research suggesting that LLMs can internally represent the truth or fiction of their statements (22). Our results show an advancement of the technology: GPT-4

possesses an evolving capacity to acknowledge when it generates fictional content. Our findings are, however, agnostic to the source of this advancement. Since much of GPT-4's training is shrouded in mystery, it is plausible that OpenAI specifically trained the model toward this goal. It is also plausible that this is an emergent property, arising from GPT-4's larger-scale training (23).

Hallucination and Topic Broadness. As topics become narrower, GPT will have fewer real and relevant articles in its training data and as a result, its hallucinations may increase. This intuition was confirmed (using our main set of 1,000 references), but only to a point. Collapsing across models, as topics narrowed, and particularly as they became very narrow, GPT was likely to admit defeat, acknowledging that it did not know of such articles. Since such acknowledgments were considered “Correct” (i.e. a failure to hallucinate), a curvilinear effect arose, where GPT first gradually hallucinated more as topics narrowed, but then less on the narrowest. Consistent with a statistical suppression effect, the linear effect of topic broadness on hallucination in a logistic regression becomes stronger ($\beta = -0.555$, $P < 0.001$; $Pseudo-R^2 = 0.075$) when excluding admissions of defeat, compared to when including them ($\beta = -0.191$, $P = 0.001$; $Pseudo-R^2 = 0.0115$). These linear patterns were robust for GPT-3.5 but not GPT-4. (Full analysis in *SI Appendix*, section S2 and Table S4.)

Article Relevance. We were interested in GPT's ability to discover references that were not only real but also relevant to specific topics. To study this, we limited data to the 1,000 articles GPT claimed were real. Collapsing across topics, GPT-4 was more likely to pull relevant articles (50.80%) compared to GPT-3.5 (30.80%); OR = 2.320. However, this effect was primarily driven by GPT-4's greater success at generating *real* articles, since hallucinations were automatically labeled irrelevant. Limiting the analysis to real references, the difference in identifying relevant articles between GPT-4 (53.70%) and GPT-3.5 (48.13%) lost significance; OR = 1.250, $P = 0.124$. In short, GPT-4 discovered more real articles than GPT-3.5 more generally and was therefore also more likely to discover relevant articles. Beyond this, it did not consistently tailor article recommendations better to specific topics. However, an exploratory analysis by topic breadth tells a more nuanced story. As detailed in *SI Appendix*, Table S5, both models successfully found relevant articles on the broadest topics and failed on the narrowest. However, GPT-3.5 appeared to drop off in this ability more sharply as the topics narrowed. The largest gap was at the “moderate” broadness level, where GPT-4 discovered relevant articles 75.00% of the time compared to GPT-3.5's 32.14%; OR = 6.333. Regression models indicated that this sharper drop-off in relevance for GPT-3.5 vs. GPT-4 was robust and statistically significant (*SI Appendix*, section S2). Thus, while neither model was exceptional at research curation, there was some evidence of incremental improvement.

Relevant Citation Counts. We requested that GPT find “important and influential” articles. To analyze how each model did, we focused on references coded as relevant. When GPT found a relevant article, how influential was its selection, as gauged by citation count? The two versions performed similarly here, with GPT-4's articles averaging 2936.7 citations compared to 3105.2 for GPT-3.5; $P = 0.791$, $d = 0.027$. Further regression analysis (*SI Appendix*, section S2) ruled out the possibility that significant differences were disguised either by GPT-3.5's relevant articles coming primarily from broader domains or because GPT-4 cited newer articles. Across linear regression models, no significant differences arose for citation count.

Study 2: GPT as Research Ethicist*

In recent years, a replicability crisis has emerged in scientific research. Large-scale studies have demonstrated limited replication of research in Psychology (24, 25), Economics (26), the Social Sciences more generally (27), and Medicine (28–30). One source of these issues is unquestionably poor statistical practices by generally well-intentioned scientists: Studies show that poor statistical practices are widely prevalent across scientific research (31, 32). A fundamental problem is that by running multiple analyses, it is easy to find coincidentally “significant” results, and thus typical significance tests become inaccurate (33). Improving the decisions of well-intentioned scientists could thus improve the reliability of science. The purpose of our second study was to examine the abilities of GPT-3.5 and GPT-4 in this domain: Can GPT catch ethical lapses and warn investigators that they are entering into the realm of questionable practices?

Research Design. In Study 2, we presented GPT-3.5 and GPT-4 with fictional vignettes describing flawed research protocols, posing as scientists looking for feedback. Three of these vignettes contained poor practices that were blatant and three more subtle. For example, in the blatant version of vignette 1, the researcher directly states: “After just 30 participants in each condition, effects already reached statistical significance ($P < 0.05$), so we stopped data collection...” The subtle version states, “We collected 50 participants in each condition, at which point statistical analysis indicated that our results reached statistical significance” and then describes collecting 150 participants in the next study. Here, the researchers do not directly describe using significance testing to decide whether to continue collection, but an experienced reviewer might be suspicious based on the contrasting sample sizes. (See *SI Appendix*, section S3, for full design, and <https://osf.io/sdahr/> for vignettes.)

The purpose of this variation between blatant and subtle vignettes was to test, 1) whether GPT showed awareness of the clear methodological problems described in the blatant vignettes, and 2) whether it would be able to “read between the lines” to recognize potential problems in more realistic descriptions of flawed research. Put differently, the subtle vignettes contained more ecologically valid descriptions, closer to how a real-world researcher might represent research with methodological flaws.

Additionally, we varied the initial prompt used to request GPT's feedback on the protocols. Matched pairs of prompts were designed to encourage either better or worse responses from GPT. The full set of prompts (*SI Appendix*, section S3) varied in levels and verbosity of feedback requested, manipulated researcher characteristics (status, theory protectiveness, rejection sensitivity), or encouraged ethical or unethical responses. Of these last, one pair prefaced the request with a pro- vs. anti-open science argument, and three attempts were made to “jailbreak” GPT—requesting it pretend not to be concerned about p-hacking, avoid mentions of p-hacking, or impersonate a known data fabricator—compared to ethical requests (consider issues around p-hacking, impersonate a known data ethicist).

In separate chats, we presented each of the 18 initial prompts to GPT-3.5 and GPT-4, followed by each of the 6 vignettes, for a total of 216 responses.

Coding of Data. Two coders, blinded to experimental conditions, independently rated all GPT responses on 10-point rubrics. Points of disagreement were discussed, with GPT afterward receiving partial credit when only one coder felt a point was merited. Coding achieved

*By referring to GPT as “Ethicist” we do not mean to attribute to it human-like morality, nor to advocate for any particular scientific standards. Rather, we examine GPT's ability to give advice aligned with currently accepted markers of ethical and replicable research.

high interrater reliability, with a Cronbach alpha of $\alpha = 0.9827$. (Full rubrics and coding details are in [SI Appendix, section S3](#).)

Analysis. Data from coding were aggregated to the level of GPT's responses, with each receiving up to 10 points. In addition to standard parametric procedures, differences were examined using Wilcoxon rank-sum ([SI Appendix, section S4](#)).

Results

Unless otherwise noted, all effects reported in this section reached a high bar for statistical significance ($P < 0.001$). More detailed statistical reporting may be found in [SI Appendix, section S4](#).

GPT-4 substantially outperformed GPT-3.5 in its responses to the Research Ethicist vignettes. In the Blatant condition, GPT-4 achieved a mean score of 8.86 out of 10 possible points, while GPT-3.5 averaged 5.39; $d = 1.992$. Similarly, in the Subtle condition, GPT-4 averaged 7.26 points compared to GPT-3.5's 4.05; $d = 1.571$. Even when poor practices were framed subtly, GPT-4 noticed and offered advice to correct most of them.

All other pairwise comparisons were significant as well. GPT-4 scored higher on blatant relative to subtle vignettes ($d = 0.987$), as did GPT-3.5 ($P = 0.0015$, $d = 0.627$). Strikingly, GPT-4 received more points in response to subtle vignettes than GPT-3.5 did in response to blatant ones ($d = 0.897$). The improvement of ChatGPT on this task was thus decisive: while the earlier model performed poorly, the more recent iteration was quite successful and could provide value to scientists in this domain.

Analysis of Initial Prompts. The variation in initial prompts was exploratory and designed to pick up only relatively large effects. Though these analyses were not fully independent, to be conservative, we used a Bonferroni correction for multiple comparisons. Since we completed 12 analyses, this correction dictated a statistical significance threshold of $P < 0.004167$.

We first examined each matched pair of prompts, testing for each whether the hypothesized "good performance" prompt yielded superior feedback compared to the "bad performance" prompt. For example, we tested whether claiming to be a chaired professor at a major research institution (high status) elicited less critical feedback compared to claiming to be a novice researcher running her first study (low status). As seen in [SI Appendix, Tables S7 and S8](#), none of these nine basic contrasts reached statistical significance. One trended: requesting GPT "carefully consider issues around p-hacking and open science" ($M = 8.38$) elicited stronger responses than requesting that it "not include any mentions of p-hacking" ($M = 6.08$, $P = 0.0106$, $d = 1.140$). However, this result did not meet the significance threshold of the Bonferroni correction. Though GPT descriptively gave worse responses following the three different "jailbreaking" prompts, compared to two contrasting prompts requesting ethical responses, this similarly failed to reach significance ($P = 0.0295$, $d = 0.588$) after correction.

One analysis yielded robust results. Regardless of how we asked GPT to behave, stronger responses emerged following prompts that in any way evoked data ethics. For example, the 12 chats where we asked GPT to impersonate a known data fabricator—designed to jailbreak GPT and elicit unethical responses—actually yielded responses that appeared to be of higher quality than most. Accordingly, we collapsed responses across prompts that in any fashion evoked data ethics (Pro- and Anti-Open Science; Concerned and Not Concerned with p-hacking; Don't Mention p-hacking; Data Ethicist and Data Fabricator), comparing these to all remaining prompts, without mentions of p-hacking or open science. Indeed, GPT provided higher-quality responses after

prompts that evoked data ethics ($M = 7.35$) than those that did not ($M = 5.78$); $P < 0.0001$; $d = 0.625$. As a robustness check, we replicated this analysis, limiting the ethics-priming prompts only to those that did so in the context of encouraging GPT to behave *badly* (e.g., impersonate a known data fabricator), initially designed to elicit poor responses. As seen in [SI Appendix, section S4](#), even this conservative test revealed the priming effect, though not robust to a Bonferroni correction ($P = 0.0099$). The positive effect of evoking data ethics appears powerful: ChatGPT's responses were somewhat improved even when the ethics primes occurred in the context of attempting to elicit *unethical* responses.

Good Research Vignettes. As a corollary to Study 2, we conducted a secondary study ([SI Appendix, section S5](#)) where ChatGPT responded to two vignettes demonstrating the opposite—rigorous practices and pristine research ethics, with 120 responses across the two models. When asked to identify positive practices in these vignettes, both models were successful. Specifically, GPT-4 identified 92.67% of the good research practices in our rubrics compared to 90.42% for GPT-3.5, with the difference nonsignificant ($P = 0.071$). Intriguingly, both models were about as good at recognizing generally accepted good research practices, though GPT-4 was vastly superior at identifying bad ones.

Study 3: GPT as Data Generator

Can GPT simulate known scientific results? Several recent articles suggest, for example, that LLMs can mimic responses from human research subjects (34–36), with some even suggesting they may significantly supplant them (37). However, assumptions about their usefulness for data generation rely on the premise that chatbots can simulate high-quality data aligned with real-world outcomes. Beyond this practical application, LLMs' ability to replicate known outcomes is a likely precursor to broader scientific capabilities. For instance, suppose we tasked GPT with generating novel but plausible hypotheses. This would require it to simulate future results by synthesizing prior knowledge. GPT's proficiency in replicating established findings underscores its capacity to simulate outcomes in this fashion, and thus its potential in functions like hypothesis generation. In Study 3, we evaluated GPT's ability to simulate data in a domain familiar to it.

In recent years, a significant body of literature has accumulated suggesting that human-like biases and stereotypes emerge from semantic patterns in large language corpora (38–40). For example, just as reaction-time tasks reveal that people more easily associate male (compared to female) names with words related to "career" compared to "family," machine learning detects analogous patterns in the co-occurrence of these words in large repositories of human language (38). These findings are theoretically important, suggesting that language can crystallize human biases, and transmit and augment their impact. They are also methodologically important, offering a new tool with which researchers can probe these issues, present and historical. However, this research poses challenges: The study of word embeddings in large language corpora is complex and computationally intensive. It is difficult for a researcher to undertake this work casually because there is currently no technically uncomplicated way to do so. Study 3 asked: might one simply ask GPT to explore its own corpus?[†] Beyond potentially offering a simplified way to pilot word embedding research, GPT's performance here provides an indicator of its broader ability to replicate known results, a precursor to other scientific abilities.

[†]Note: We refer to GPT "exploring its own corpus," which was the task asked of it. However, it should be noted that we lack insight into how other elements of GPT's training—e.g., reinforcement learning, fine-tuning—impact its responses.

Research Design. In this study, we explored four well-studied gender stereotypes: Gender Attitudes (overall positivity/negativity toward women vs. men), Gender Art/Science stereotypes, Gender Home/Work stereotypes, and Gender Math/Reading stereotypes. These stereotypes have been robustly studied in human subjects, using both implicit and explicit measures (41, 42). Furthermore, consistent patterns for these stereotypes have been found in research on word embeddings in language corpora (39). We did not have access to GPT’s model parameters to generate its word embeddings directly, and instead used estimates provided within the open-ended language of the chatbot’s responses. Adapting stimuli from Charlesworth et al. (39), we presented GPT-3.5 and GPT-4 with thousands of randomly ordered word dyads, requesting it estimate cultural associations between each based on its training data. (Full design in [SI Appendix, section S6.](#)) For analysis, GPT’s estimates were treated as analogous to cosine similarity measures from word embedding research (38). To calculate a measure of relative cultural association—e.g., a greater association of Female with Home and Male with Work, relative to Male with Home and Female with Work—the procedure was followed for calculating the WEAT *D*-score (39).

Results

Table 1 depicts the real WEAT *D*-scores for each construct examined, drawn from Charlesworth et al.’s (39) meta-analytic estimates across adult corpora (39), compared to those calculated using responses from GPT-3.5 and GPT-4. Positive WEAT *D*-scores reflect effects in the stereotype-congruent direction based on prior research. Main results replicated prior findings: GPT’s estimates based on its training data reflected a cultural preference for Female over Male, and a stronger association of Female (relative to Male) with Art vs. Science, Home vs. Work, and Reading vs. Math. The effects gathered from GPT were often somewhat stronger than those reported in prior research, though this pattern is inconsistent. This may reflect the troubling tendency for AI systems to amplify biases in their training data (43–45). It is interesting to note that these effects are not generally smaller for GPT-4 vs. GPT-3.5, despite efforts OpenAI has made to debias the model (46). This aligns with prior research showing that more powerful models tend to intrinsically learn human biases more precisely (47). These results are promising in terms of GPT’s ability to generate estimates of word embedding results, suggesting a use case in piloting this research. However, they come with some caveats. First, the interitem correlations between GPT-3.5 and GPT-4’s responses to the same word dyads were variable but modest: $r = 0.382$ for the Math-Reading task, $r = 0.568$ for the Preference task, $r = 0.666$ for the Work-Home task, and $r = 0.554$ for the Art-Science task (all P s < 0.0001). These moderate correlations might indicate differences in how GPT-3.5 and GPT-4 approached the task. Alternatively, they might indicate reliability constraints, limiting GPT’s consistency in eliciting these effects. Second, we calculated Single-Category WEAT *D*-scores for each of the concepts, to gauge the degree to which results were driven by stronger Female–Male associations with each attribute. Interestingly, our results diverge from prior research (39) in that those generated

by GPT are primarily driven by stronger associations of Female with the stereotypically female category, and not also of Male with the stereotypically male category ([SI Appendix, Table S10.](#))

Study 4: GPT as Novel Data Predictor

In Study 3, we examined GPT’s ability to simulate data from word embedding research. As use cases for GPT as a data generator go, this one is obvious: Since GPT is trained on large language corpora, it might display knowledge of word embedding patterns found in them. Less certain and less tested is the potential for LLMs to predict data that are novel and outside their training data. Recent conceptual work has argued that LLMs may augment or even replace human test subjects (34–37). Logically, the degree of this potential depends upon its ability to predict unseen patterns. If GPT is tethered to its training data, it might be useful for certain kinds of basic tasks, such as piloting the psychometric properties of personality scale items. However, to produce results that are both correct and novel, it must be able to predict data patterns that are unknown to it. In Study 4, we gave ChatGPT a more difficult task in this regard. Specifically, we asked GPT to predict patterns of data that were complicated and unfamiliar. For this, we used a second paper by Charlesworth and colleagues (48), which introduced a novel dataset: the Project Implicit International Dataset. This paper describes patterns of implicit and explicit attitudes from 2.3 million participants across 34 countries. Critically, accumulated evidence suggests that implicit attitudes—automatic associations held between attitude objects, typically measured by reaction-time tasks—are distinct from explicit attitudes captured in self-reports (49). For example, a person may explicitly express equal positivity toward straight and gay individuals, while at the same time implicitly harboring greater positivity toward straight individuals. Indeed, in the Project Implicit International Dataset, correlations between country-level implicit and explicit attitudes vary by attitude object but are generally not strong ([SI Appendix, section S9.](#)) These country-level data were previously unpublished, and the paper was first posted online after GPT’s training cutoff at the time of the study. In Study 4, we had GPT-3.5 and GPT-4 make a total of 60 different predictions of cross-country patterns of Explicit and Implicit Sexuality Attitudes, Age Attitudes, and Gender Science/Liberal Arts stereotypes. (Full design in [SI Appendix, section S8.](#))

Results

More detailed statistical reporting may be found in [SI Appendix, section S9.](#) For each attitude/stereotype, we examined several patterns. First, we examined the intercorrelations between ChatGPT’s different predictions of the same attitude. For example, a high correlation between GPT’s different predictions of Implicit Sexuality Attitudes suggests reliability and consistency in how it approached the task. We then examined the correlations between GPT’s predictions of implicit and explicit attitudes. Importantly, GPT likely has more information about explicit compared to implicit attitudes. For example, at the time of this article’s writing, a Google Scholar search for “Sexuality Attitudes” returned 2,710 results compared

Table 1. WEAT D-scores from ChatGPT and Real Large Language Corpora

	Female-Good Male-Bad	Female-Art Male-Science	Female-Home Male-Work	Female-Reading Male-Math
Prior research	WEAT <i>D</i> = 0.49	WEAT <i>D</i> = 0.54	WEAT <i>D</i> = 0.94	WEAT <i>D</i> = 0.67
GPT-3.5	WEAT <i>D</i> = 1.00	WEAT <i>D</i> = 1.16	WEAT <i>D</i> = 0.40	WEAT <i>D</i> = 0.73
GPT-4	WEAT <i>D</i> = 0.57	WEAT <i>D</i> = 1.46	WEAT <i>D</i> = 0.45	WEAT <i>D</i> = 0.96

Notes: Prior research numbers are meta-analytic estimates from adult language corpora (39).

to just 26 results for “Implicit Sexuality Attitudes.” We were interested in whether GPT leveraged different information when asked to predict more novel implicit attitudes. If GPT’s predictions of implicit attitudes correlate more highly with each other than they do with its predictions of explicit attitudes, this would suggest it is reliably leveraging different information in making the two predictions. Conversely, if GPT’s predictions of implicit attitudes correlate as highly with its explicit predictions as with each other, this would suggest it is approaching the tasks similarly, and not leveraging substantively different information in predicting implicit vs. explicit attitudes. Finally, and most critically, we examined the correlations between GPT’s predictions and real-world results in the Project Implicit International Dataset, to gauge GPT’s overall success as a Novel Data Predictor. (See [SI Appendix, Table S12](#), for additional summary statistics. For full correlation tables, see “GPT as Data Predictor Correlation Tables 20240228” at <https://osf.io/sdahr/>.)

Sexuality Attitudes[‡] On average, correlations between GPT-3.5’s five different explicit predictions of Sexuality Attitudes were high (mean $r = 0.875$), as were correlations between GPT-3.5’s different implicit predictions of Sexuality Attitudes (mean $r = 0.879$). This suggests it approached these tasks reliably. However, correlations between its implicit and explicit predictions were nearly as high (mean $r = 0.778$). GPT-3.5’s five sets of implicit predictions were similar to its five sets of explicit predictions, indeed nearly as similar to them as to each other. For GPT-4, correlations between different explicit predictions were even higher (mean $r = 0.957$), as were correlations between different implicit predictions (mean $r = 0.946$). Interestingly, correlations between its explicit and implicit predictions were equally high (mean $r = 0.952$): statistically, GPT-4’s predictions of Implicit Sexuality Attitudes looked identical to its predictions of Explicit ones. These patterns suggest ChatGPT was limited in the new information it applied to these different tasks. When predicting (more novel) implicit attitudes compared to (more familiar) explicit attitudes, GPT-3.5 seemingly used little new information, and GPT-4 almost *no* new information.

Considering these analyses, we collapsed across implicit and explicit predictions in examining ChatGPT’s success at predicting actual Sexuality Attitudes. GPT-3.5 did a reasonable job predicting real-world Explicit Sexuality Attitudes by country (mean $r = 0.602$), but was unsuccessful at predicting Implicit Sexuality Attitudes (mean $r = -0.014$). GPT-4’s predictions correlated highly with country-level explicit attitudes (mean $r = 0.714$), but it similarly failed at predicting implicit attitudes less represented in its training data (mean $r = 0.152$).

Age Attitudes and Gender Science/Liberal Arts Stereotypes. GPT-3.5’s predictions of Age Attitudes and Gender stereotypes proved not only unsuccessful, but incoherent. As detailed in [SI Appendix, section S9](#), for each, GPT-3.5’s different implicit predictions were uncorrelated with each other, as were its different explicit predictions. Given this low reliability, it is unsurprising that its collective predictions of Age Attitudes were uncorrelated with real country-level patterns of Explicit (mean $r = -0.010$) and Implicit (mean $r = -0.175$) Age Attitudes. Similarly, GPT-3.5’s predictions of country-level Gender Science/Liberal Arts stereotypes were uncorrelated with actual explicit (mean $r = -0.009$) and implicit (mean $r = -0.044$) results.

GPT-4 completed these tasks more reliably but was not more successful in predicting real-world results. For Age Attitudes,

GPT-4’s different explicit predictions were moderately to highly correlated (mean $r = 0.645$), as were its different implicit predictions (mean $r = 0.726$). Correlations between its implicit and explicit predictions were in the same range (mean $r = 0.664$), again suggesting that it did not lean on substantially new information for predicting patterns of implicit vs. explicit attitudes. Critically, its collective predictions were on average *negatively* correlated with real country-level Explicit Age Attitudes (mean $r = -0.395$) and uncorrelated with Implicit Age Attitudes (mean $r = -0.120$).

For Gender Science/Liberal Arts stereotypes, GPT-4’s different explicit predictions were weakly correlated (mean $r = 0.363$). Curiously, its different implicit predictions were more consistent, correlating strongly across chats (mean $r = 0.868$), and correlations between its explicit and implicit responses fell between the two (mean $r = 0.499$). The real-world explicit results GPT predicted here were measured with two items that needed to be combined: one capturing associations of Male vs. Female with Science, and a second with Liberal Arts. (For full items, see “GPT as Novel Data Predictor Materials 20240227” at <https://osf.io/sdahr/>.) This more complicated explicit item may have challenged the LLM. In any case, neither sets of answers predicted actual cross-country results. GPT-4’s explicit predictions were uncorrelated with real country-level explicit (mean $r = -0.192$) and implicit (mean $r = 0.054$) stereotypes. Similarly, GPT’s implicit predictions were negatively correlated with explicit (mean $r = -0.417$) stereotypes and uncorrelated with implicit (mean $r = -0.067$) ones.

Sexuality bias may receive more media coverage than Age Attitudes or Gender Science/Liberal Arts stereotypes. Though unaware of published research on this topic, we conducted three tests to examine this possibility. Patterns from Bing searches, Google Books Ngram Viewer, and chats with GPT-4 all suggested that Sexuality Attitudes are better represented in ChatGPT’s knowledge base relative to Age Attitudes or Gender Science stereotypes ([SI Appendix, Section S9 and Figs. S1 and S2](#)). Predicting these latter patterns was thus more difficult. In line with this thinking, while both LLMs were successful at predicting patterns of Explicit (though not Implicit) Sexuality bias, neither achieved even small positive correlations with the other real-world results.

General Discussion

Across four studies, we have tested GPT’s ability to enhance the scientific process. Our focus has been on psychological science, where the authors have sufficient expertise to judge the quality of GPT’s output, but we have selected tasks that are applicable across domains. Future research should, however, confirm the degree to which this work generalizes to other disciplines.

We included both GPT-3.5 and GPT-4, even though GPT-4 is expected to be superior and GPT-3.5 may fall out of use as future versions are released. We did this for several reasons. First, the quantitative difference between the two is of interest in tracking the speed of improvement. More importantly, only by comparing the two could we gain insight into newly emerging processes such as GPT-4’s ability for self-correction. As we will discuss, such findings have implications for our understanding of underlying processes in machine cognition. Finally, comparing the models allowed us to highlight where the technology did and did not advance, such as GPT-4’s reduction in hallucinations but not in smaller errors. This work thus offers actionable insights that can help guide the training of future models.

Study 1 probed GPT’s ability as a Research Librarian. GPT showed a varied trajectory in terms of the ability to discover relevant research. By any measure, GPT-4 generated many fewer fictional references. It also displayed a far greater tendency to acknowledge

[‡]As detailed in [SI Appendix, section S9](#), here and elsewhere GPT-3.5 had difficulty with this task. GPT-3.5’s intended direction of scoring was frequently unclear, necessitating follow-up questions to gauge the meaning of its predictions. This sometimes rendered responses difficult to interpret, particularly for Age Attitudes and Gender Science stereotypes.

when it was generating fiction. This is potentially important for the technology's development. There is a likely tradeoff between novelty and truth in LLMs: Hallucinations might be inevitable in a model capable of creativity (50). Training a model with a firm goal of minimizing fiction generation might therefore be problematic, risking it becoming more factual but also less creative. The possibility that GPT-4 is developing some form of fiction recognition is therefore intriguing. An AI capable of discerning fact from fiction in its own creation may be capable of generating fact when facts are desirable, and fiction when fiction is desirable, much as a human author might choose to write a short story on one occasion and a research article on another. Put differently, the ability to parse fact from fiction in its responses may open the door for LLMs that are capable of being at once creative and truthful.

That said, there is significant room for improvement. GPT-4 still generated a nontrivial number of unacknowledged hallucinations. Moreover, acknowledgment generally came on the chat level: GPT would note that its references "might be fictional," for example, without distinguishing which specific references were real or fake. Finally, GPT-4 did not show meaningful improvement in terms of smaller errors, such as listing the wrong year or journal. This pattern is interesting. GPT-4 increasingly mirrors humans on this task: It has fewer instances of outright fabrication, to which people are not prone, but not of smaller errors people might also make. Such errors are consequential: Even small errors might, for example, lead to inaccurate conclusions about authors' scientific output in formulas that help decide tenure, or incorrect citations in new articles. The latter problem may be self-propagating, since incorrect citations are automatically indexed on Google Scholar, risking an expanding misinformation ecosystem.

Interestingly, hallucination was moderated by the completeness of the references generated. When generating incomplete citations, both GPT-3.5 and GPT-4 were sharply more likely to hallucinate. One possible framing of this effect is in terms of formality: By providing incomplete references, GPT was intrinsically making the chat less formal. Future research should probe the causality of this finding by experimentally varying the formality of the request to explore whether this changes ChatGPT's effectiveness in discovering real research. This finding is also interesting in that here, again, we see a parallel to human cognition. A person will be more prone to misstate a fact—e.g., misquoting the source of a statistic—over a dinner conversation than in a scientific communication. In this case, of course, the source of the confabulation is obvious: the human is drawing on imperfect memory rather than verifying documentation. The source of this analogous error in GPT is less clear and very likely different. Nevertheless, in some sense, the machine appears to verify facts more in some contexts than others, seemingly seizing upon informality as an opportunity to be sloppy. Uncovering the source of this discrepancy may generate insights into the processes underlying machine cognition.

ChatGPT's abilities in terms of pulling relevant references were uninspiring. It was successful at discovering references on broad topics but quickly became less successful as the subject matter became narrower. However, we saw advancement between GPT-3.5 and GPT-4 in this regard. GPT-4 was more successful at pulling references on moderately broad topics, suggesting potential for future improvement in this area.

Study 2 probed GPT's abilities as a Research Ethicist. GPT-4 shined in this regard, decisively outperforming GPT-3.5 when providing feedback on subpar research protocols. While these results were large statistically, examining the responses qualitatively makes the contrast even more striking. (For full transcripts, see "GPT as Research Ethicist Transcripts 20240220" at <https://osf.io/sdahr/>.) To put its performance into perspective, note that the

grading rubrics were built collaboratively: the original draft by one of the authors had eight points for each, another author then revised it with a ninth point, and upon reflection, the original one added a tenth. While some of GPT-4's responses were better than others, on average it scored nearly nine points for blatant vignettes, which is roughly identical to what the authors effectively averaged across three iterations of the rubrics. The performance of GPT-3.5 lies in stark contrast. Not only did it often miss the researchers' lapses, but at times it was even complimentary. For example, on several occasions it praised the researchers' decision to add more research subjects after checking for statistical significance, noting that it "added statistical power."

In a scientific era defined by a replicability crisis, these results are important. They suggest that GPT-4 is highly capable of giving useful feedback—aligned with generally accepted standards of modern research practice—on experimental protocols. GPT-4 was reasonably successful at this task even when the vignettes were framed subtly. This result is striking because this required GPT to infer bad research practices where they were not clearly stated. It is also practically important, as it suggests the LLM can help well-intentioned researchers—operating in a realistic context—improve the quality and ethics of their work. Finally, it is conceptually interesting that GPT-4 suggested distinctly modern practices: From an ocean of possible suggestions, the more recent model was able to circumnavigate practices that have aged poorly, and instead present advice aligned with recent advances and best practices. Not so for GPT-3.5. The comparatively poor performance of GPT-3.5 is disheartening in that researchers who do not purchase the paywalled model upgrade may receive poor-quality advice. Indeed, GPT-3.5's responses could even embolden poor researchers, since at times it openly encouraged subpar practices. However, the difference between the two LLMs may also be cast in an optimistic light: the technology's progress is profound, suggesting that its next iteration may prove an extremely powerful tool for helping researchers design protocols and improve practices. Future research should examine the LLMs' ability to improve higher-quality protocols, gauging their ability to help more skillful researchers.

The results around initial prompts, while exploratory, were generally heartening. At the least, they suggest that casual mentions of things like researcher status or sensitivity to criticism are not eliciting large and robustly worse feedback from GPT. It also did not prove trivial to jailbreak the technology: GPT consistently rejected requests to provide feedback in an unethical manner. Indeed, if anything, preceding our protocol with arguments against open science or requests that it ignore p-hacking may have riveted the LLMs' attention to these issues, leading to more ethical responses. This unique "priming" effect—the tendency for GPT to give superior feedback following initial prompts that evoked data ethics—is both practically and theoretically important. Merely asking GPT to be more critical or verbose did not elicit stronger responses. However, evoking data ethics in any manner led to better feedback. Practically, this underscores the importance of specificity when eliciting advice from ChatGPT. Researchers may benefit from highlighting specific areas where they require support. Theoretically, it reveals nuance in the process by which GPT responds to prompts; merely hinting at ethics leads the LLMs to evaluate the problem differently, adopting an ethically minded perspective.

In Study 3, we probed ChatGPT's ability to generate useful data for estimating word-embedding results. GPT's results replicated known overall effects from this literature. GPT may thus be useful for generating data in this context, for example, to pilot new word embedding work in a technically simplified manner. However, the importance of this work extends beyond GPT's ability to generate data in this relatively specific domain. The

ability to realistically simulate real-world data is implicitly tied to other important scientific abilities. For example, imagine GPT was asked to generate plausible scientific hypotheses. Completing this task would likely require the LLM to draw upon and synthesize existing knowledge in new and meaningful ways. Success would require GPT to display an intrinsic command of this knowledge. Put another way, if it is unable to simulate existing knowledge, GPT cannot be expected to successfully simulate extensions of this knowledge. GPT's ability to generate useful hypotheses may therefore be dependent on its ability to replicate existing results.

It should be noted that the results from Study 3, while promising, are not decisive. In particular, the divergence from existing research on the patterns of Single-Category WEAT *D*-scores is rather puzzling. There may be differences in GPT's training data or approach to the task, leading to inconsistencies with prior work. The relative uniformity across GPT-3.5 and GPT-4 is consistent with this interpretation. Where result patterns diverged from known results, they usually did so consistently across the two models. A more troubling possibility is that GPT ignored our instructions not to adjust for stereotypical associations viewed as negative. Gendered associations with Work and Math have produced wide discussion, and GPT may have simply been reluctant to suggest men are more associated with these categories. If this is the case, GPT could prove an unreliable source for data related to socially undesirable effects. This issue is larger than the specific use case from this study, extending to any use of LLMs to augment human subjects. While self-censoring of LLMs may be an overall societal good, in the context of social psychological research, this could undermine their potential. One cannot, for example, expect reliable data on human prejudice if GPT refuses to display bias within the context of scientific research.

Finally, in Study 4, we examined GPT's ability to predict novel data—cross-cultural patterns of implicit and explicit attitudes—published after its training cutoff. Our results here should be viewed as suggestive rather than definitive since we have studied merely one of many possible domains in which GPT could be asked to make novel predictions. Conceptually, though, we take issue with the possibility that LLMs can predict novel data. A finding that is novel is, by definition, one outside the scope of the LLM's training data. As a thought experiment, imagine a powerful AI somehow came to exist in the 16th century. This AI had more cognitive capabilities than current technologies but received none of the data collected from hundreds of years of astronomical research. Without a telescope, could this AI locate the moons of Jupiter? We argue that it could not. Galileo's discoveries were not merely creative insights; they were the result of new data. Given data from a good telescope, a powerful AI might perhaps predict hundreds of years of physical research. Without it, it likely could not.

This extends to the idea of AI acting as a human test subject (and data source more generally). Without data to suggest a certain result will arise, how can it be expected to mimic the effect? It would be blind as to Jupiter's moons. As a simulated human subject, GPT might therefore be expected to replicate—indeed perhaps overreplicate—existing findings. To be sure, it might be able to combine knowledge in new ways to reveal novel discoveries about patterns in historical data. But there is no obvious mechanism by which it could generate discoveries dependent on novel data, a cornerstone of scientific progress.

Our results across Studies 3 and 4, while not definitive, are consistent with this argument. Tasked with replicating known

effects in a domain (word embeddings) familiar to it, GPT-3.5 and GPT-4 were both relatively successful. But tasked with predicting novel and unfamiliar data, both models generally failed.

The tendency to lean heavily on what is familiar was evident in GPT's approach to predicting cross-cultural IAT results. While even explicit attitude predictions were often beyond it, GPT had some success in that domain. Both GPT-3.5 and GPT-4 achieved relatively high correlations in predicting country-level Explicit Sexuality Attitudes. However, their predictions for Implicit Sexuality Attitudes were nearly identical, suggesting they brought little additional information to this more novel prediction. This is particularly striking considering that implicit and explicit attitudes are often only moderately correlated (49), a fact in GPT's knowledge base (*SI Appendix, section S9*). In short, these two phenomena were sufficiently distinct that one would expect GPT to leverage somewhat different information when predicting them. Surprisingly, it did not.

It should be noted that even GPT's prediction of explicit attitudes was far from stellar. The LLMs succeeded only with Sexuality Attitudes. It seems likely that the extensive cross-national coverage and political discussion of sexuality offered GPT information to lean on here. Beyond Sexuality Attitudes, when GPT attempted to predict Age Attitudes and Gender Liberal Arts/Science stereotypes—areas that receive less media coverage—both models failed spectacularly.

We suggest that GPT's ability to act as a data source may be limited to relatively simple tasks and domains where the likely results are known or predictable. Future research should finely map where GPT is and is not successful in simulating data. That said, it is conceptually possible that LLMs may prove able to elaborate upon known results in novel ways, cohesively combining sources of knowledge. Testing LLMs' abilities to generate new knowledge in this manner may prove fertile ground for future research. We believe it unlikely, however, that current or future LLMs will be capable of generating true empirical novelty, whereby results do not reflect existing information regardless of how it is combined, because we see no mechanism by which LLMs can predict something with no counterpart in their training data. AI may thus continue to be limited in this regard, even as technology advances: Scientific progress will likely always require real-world data.

To conclude, we turn to the broader question of whether LLMs can enhance or facilitate the scientific process. Based on this report, we would tentatively answer "yes." ChatGPT's ability to compile and curate research is currently limited but rapidly improving in ways (e.g., increasing acknowledgment of fiction) that indicate future generations of this technology might be successful in this area. Already, GPT-4 shows a surprisingly strong mastery of research methods and ethics, and may be able to help scientists improve their practices. ChatGPT's successful replication of known results suggests a degree of command over existing knowledge that may simplify research piloting. This same phenomenon raises the possibility that GPT may be able to synthesize existing knowledge sources to generate new and plausible hypotheses, a premise that may prove a fruitful ground for future research. The most fundamental limitation we perceive is in GPT's seeming inability to predict highly novel empirical results. This limitation is unsurprising, but it speaks to the need for moderation in the optimism about this technology. Future models may show profound abilities and spur scientific advancement. But, these abilities should not be mistaken for omniscience. Like human scientists, advanced LLMs will likely remain limited by the knowledge they already possess.

Data, Materials, and Software Availability. Transcripts from GPT and spreadsheets containing quantification of these data, including direct recording of quantitative data and the outputs from coding qualitative data according to rubrics. All data discussed in the paper are publicly available in OSF (DOI: [10.17605/OSF.IO/SDAHR](https://doi.org/10.17605/OSF.IO/SDAHR)) (51).

ACKNOWLEDGMENTS. We would like to thank Melanie Mitchell, Tina Eliassi-Rad, and Tessa Charlesworth for their advice on this work, Igor Grossmann for his comments on an early draft, and Anya Vedantam for her help coding ChatGPT

transcripts. ChatGPT provided brainstorming and advice on this project but was not used in the manuscript's preparation. All correspondence regarding this manuscript should be directed to S.A.L. at steve@cangrade.com.

Author affiliations: ^aCangrade, Inc., Watertown, MA 02472; ^bInformation School, University of Washington, Seattle, WA 98195; and ^cDepartment of Psychology, Harvard University, Cambridge, MA 02138

1. A. Vaswani *et al.*, "Attention is all you need" in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds. (Curran Associates Inc., 2017), vol. 30, pp. 5998–6008.
2. X. Zhang *et al.*, Artificial intelligence for science in quantum, atomistic, and continuum systems. *arXiv [Preprint]* (2023). <https://arxiv.org/pdf/2307.08423.pdf> (Accessed 19 January 2024).
3. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
4. G. Carleo, M. Troyer, Solving the quantum many-body problem with artificial neural networks. *Science* **355**, 602–606 (2017).
5. D. Kochkov *et al.*, Machine learning-accelerated computational fluid dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2101784118 (2021).
6. M. H. S. Segler, M. Preuss, M. P. Waller, Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
7. Y. Zheng *et al.*, Large language models for scientific synthesis, inference and explanation. *arXiv [Preprint]* (2023). <https://arxiv.org/pdf/2310.07984.pdf> (Accessed 19 January 2024).
8. K. M. Jablonka *et al.*, 14 examples of how LLMs can transform materials science and chemistry: A reflection on a large language model hackathon. *Digit. Discov.* **2**, 1233–1250 (2023).
9. D. A. Boiko, R. MacKnight, B. Kline, G. Gomes, Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).
10. D. Demszyk *et al.*, Using large language models in psychology. *Nat. Rev. Psychol.* **2**, 688–701 (2023).
11. H. Koteek, R. Dockum, D. Q. Sun, "Gender bias and stereotypes in large language models" in *Proceedings of the ACM Collective Intelligence Conference (CI '23)* (Association for Computing Machinery, Delft, The Netherlands, 2023), pp. 12–24.
12. F. Bianchi *et al.*, "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale" in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FACt '23)* (Association for Computing Machinery, New York, NY, 2023), pp. 1493–1504.
13. Y. Zhang *et al.*, Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv [Preprint]* (2023). <https://arxiv.org/pdf/2309.01219.pdf> (Accessed 19 January 2024).
14. W. H. Walters, E. I. Wilder, Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci. Rep.* **13**, 14045 (2023).
15. S. Bubeck *et al.*, Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv [Preprint]* (2023). <https://arxiv.org/pdf/2303.12712.pdf> (Accessed 19 January 2024).
16. M. Binz, E. Schulz, Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2218523120 (2023).
17. S. Yadlowsky, L. Doshi, N. Tripuraneni, Pretraining data mixtures enable narrow model selection capabilities in transformer models. *arXiv [Preprint]* (2023). <https://arxiv.org/pdf/2311.00871.pdf> (Accessed 19 January 2024).
18. R. A. Millikan, Science and religion. *Bull. Calif. Inst. Technol.* **32**, 3–20 (1922).
19. R. Taylor *et al.*, Galactica: A large language model for science. *arXiv [Preprint]* (2022). <https://arxiv.org/pdf/2211.09085.pdf> (Accessed 19 January 2024).
20. Y. Cao *et al.*, A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT. *arXiv [Preprint]* (2023). <https://arxiv.org/pdf/2303.04226.pdf> (Accessed 19 January 2024).
21. V. Rawte *et al.*, Exploring the relationship between LLM hallucinations and prompt linguistic nuances: Readability, formality, and concreteness. *arXiv [Preprint]* (2023). <https://arxiv.org/pdf/2309.11064.pdf> (Accessed 19 January 2024).
22. A. Azaria, T. Mitchell, The internal state of an LLM knows when it's lying. *arXiv [Preprint]* (2023). <https://arxiv.org/pdf/2304.13734.pdf> (Accessed 19 January 2024).
23. J. Wei *et al.*, Emergent abilities of large language models. *arXiv [Preprint]* (2023). <https://arxiv.org/pdf/2206.07682.pdf> (Accessed 19 January 2024).
24. Open Science Collaboration, PSYCHOLOGY, Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
25. B. A. Nosek *et al.*, Replicability, robustness, and reproducibility in psychological science. *Annu. Rev. Psychol.* **73**, 719–748 (2022).
26. C. F. Camerer *et al.*, Evaluating replicability of laboratory experiments in economics. *Science* **351**, 1433–1436 (2016).
27. C. F. Camerer *et al.*, Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637–644 (2018).
28. J. P. A. Ioannidis, Contradicted and initially stronger effects in highly cited clinical research. *JAMA* **294**, 218–228 (2005).
29. T. M. Errington *et al.*, An open investigation of the reproducibility of cancer biology research. *eLife* **3**, e04333 (2014).
30. T. M. Errington *et al.*, Investigating the replicability of preclinical cancer biology. *eLife* **10**, e71601 (2021).
31. L. K. John, G. Loewenstein, D. Prelec, Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* **23**, 524–532 (2012).
32. M. L. Head, L. Holman, R. Lanfear, A. T. Kahn, M. D. Jennions, The extent and consequences of p-hacking in science. *PLoS Biol.* **13**, e1002106 (2015).
33. J. P. Simmons, L. D. Nelson, U. Simonsohn, False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
34. L. P. Argyle *et al.*, Out of one, many: Using language models to simulate human samples. *Polit. Anal.* **31**, 337–351 (2023).
35. D. Dillon, N. Tandon, Y. Gu, K. Gray, Can AI language models replace human participants? *Trends Cognit. Sci.* **27**, 597–600 (2023).
36. G. V. Aher, R. I. Arriaga, A. T. Kalai, "Using large language models to simulate multiple humans and replicate human subject studies" in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause *et al.*, Eds. (PMLR, 2023), pp. 337–371.
37. I. Grossmann *et al.*, AI and the transformation of social science research. *Science* **380**, 1108–1109 (2023).
38. A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
39. T. E. S. Charlesworth, V. Yang, T. C. Mann, B. Kurdi, M. R. Banaji, Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychol. Sci.* **32**, 218–240 (2021).
40. W. Guo, A. Caliskan, "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases" in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AAAI/ACM, 2021)*, pp. 122–133.
41. B. A. Nosek *et al.*, Pervasiveness and correlates of implicit attitudes and stereotypes. *Eur. Rev. Soc. Psychol.* **18**, 36–88 (2007).
42. B. A. Nosek, M. R. Banaji, A. G. Greenwald, Math = male, me = female, therefore math not = me. *J. Pers. Soc. Psychol.* **83**, 44–59 (2002).
43. J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K. W. Chang, "Men also like shopping: Reducing gender bias amplification using corpus-level constraints" in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, M. Palmer, R. Hwa, Eds. (Association for Computational Linguistics, Copenhagen, Denmark, 2017), pp. 2979–2989.
44. K. Lloyd, Bias amplification in artificial intelligence systems. *arXiv [Preprint]* (2023). <https://arxiv.org/pdf/1809.07842.pdf> (Accessed 19 January 2024).
45. A. Wang, O. Russakovsky, "Directional bias amplification" in *Proceedings of the 38th International Conference on Machine Learning (PMLR, 2021)*, pp. 10882–10893.
46. OpenAI, GPT-4 technical report. *arXiv [Preprint]* (2023). <https://arxiv.org/pdf/2303.08774.pdf> (Accessed 19 January 2024).
47. M. Nadeem, A. Bethke, S. Reddy, "Stereoset: Measuring stereotypical bias in pretrained language models" in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Association for Computational Linguistics, 2021), pp. 5356–5371.
48. T. E. S. Charlesworth, M. Navon, Y. Rabinovich, N. Lofaro, B. Kurdi, The project implicit international dataset: Measuring implicit and explicit social group attitudes and stereotypes across 34 countries (2009–2019). *Behav. Res. Methods* **55**, 1413–1440 (2023).
49. W. Hofmann, B. Gawronski, T. Gschwendner, H. Le, M. Schmitt, A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Pers. Soc. Psychol. Bull.* **31**, 1369–1385 (2005).
50. M. Lee, A mathematical investigation of hallucination and creativity in GPT models. *Mathematics* **11**, 2320 (2023).
51. S. A. Lehr, A. Caliskan, S. Liyanage, M. R. Banaji, Data from "ChatGPT as research scientist: Probing GPT's capabilities as a research librarian, research ethicist, data generator and data predictor." OSF. <https://osf.io/sdahr/>. Deposited 10 June 2024.