

# GPT-ology, Computational Models, Silicon Sampling: How should we think about LLMs in Cognitive Science?

Desmond C. Ong (desmond.ong@utexas.edu)

Department of Psychology, The University of Texas at Austin

## Abstract

Large Language Models have taken the cognitive science world by storm. It is perhaps timely now to take stock of the various research paradigms that have been used to make scientific inferences about “cognition” in these models or about human cognition. We review several emerging research paradigms—GPT-ology, LLMs-as-computational-models, and “silicon sampling”—and review recent papers that have used LLMs under these paradigms. In doing so, we discuss their claims as well as challenges to scientific inference under these various paradigms. We highlight several outstanding issues about LLMs that have to be addressed to push our science forward: closed-source vs open-sourced models; (the lack of visibility of) training data; and reproducibility in LLM research, including forming conventions on new task “hyperparameters” like instructions and prompts.

**Keywords:** Large Language Models; Cognitive Science

## Introduction

Recent scientific discourse in cognitive science in 2023 and 2024 seems to be all about Large Language Models (LLMs) (Binz & Schulz, 2023; Trott et al., 2023), including entire workshops at CogSci2023<sup>1</sup> (Hardy et al., 2023) and similar conferences. Over this short time, we have seen a variety of different research paradigms emerge. Some research is focused on evaluating the cognitive capacities of these LLM models: what they can or cannot do. One might call this type of research “**GPT-ology**”, which involves studying how LLMs like GPT-4 process information<sup>2</sup>. These papers are usually characterized by inferences about the state of artificial cognition, and less on insights for human cognition. Other researchers suggest that we should instead abstract away from specific models, and use **LLMs as a computational model** of human cognition (e.g., Blank, 2023; Frank, 2023b). One example within this paradigm uses LLM performance as an “Existence Proof” about the sufficient conditions for certain cognitive capabilities to emerge (Contreras Kallens et al., 2023; Kauf et al., 2023; Piantadosi, 2023). And finally, some researchers have been using LLMs to simulate human behavior (Argyle et al., 2023; Dillion et

al., 2023; Grossmann et al., 2023; Park et al., 2023), and exploring how we can use these “**silicon samples**” to make inferences about people.

Given the interest in LLMs in cognitive science, we thought it timely to take stock of progress in these research paradigms. These paradigms are not exhaustive; In fact, we fully expect more creative approaches to appear. Neither are they mutually exclusive: the same empirical evidence—having an LLM respond to questions like in a psychological experiment—could be used to support research questions in more than one paradigm (see Fig. 1). The goal of this paper is to lay out a framework for thinking about these different approaches to using or studying LLMs. This is important because, in laying out these research questions at a “bird’s-eye” level, we can discuss the outstanding issues that affect most, if not all, of these research efforts, and that we feel have to be addressed as a field. We do not claim to have answers to all of these issues, although we provide some thoughts and suggestions for resolutions.

In this short review, we focus on examining the scientific logic and assumptions inherent in these approaches. Due to the “meta”-level of this discussion, we will not comment on specific psychological claims—for instance, claims about specific linguistic capabilities—vis-à-vis the literature and background knowledge required for adjudicating those claims. Instead, we will focus our discussion on more general principles: scientific reliability and validity, logical deductions, and epistemic support.

## A typology of research paradigms

Here, we define a typology of research paradigms that have been applied to study LLMs. These paradigms are not mutually exclusive; in fact, they all rely on the same base experiment of having LLMs provide responses to some input (e.g., see Fig. 1), and so the same data can be used to address multiple paradigms. We differentiate these paradigms by their specific research goals, and hence the inferences that are made from the results.

For each paradigm, we lay out some research questions, typical experiments, results and inferences that are made, and importantly, some of the challenges and concerns about interpretation that should be addressed.

<sup>1</sup><https://cogscillm.com/>

<sup>2</sup>This name is inspired by earlier work on studying how BERT, then the most successful NLP model, worked, which earned the moniker “BERT-ology” (Rogers et al., 2021).

Paradigm	(Example) Research Questions	Inference:	Example Experiment	Example Inference
GPT-ology	What do LLMs know?	about LLM capabilities		"GPT can <i>star-ify</i> !"
LLM-as-computational-model	What is learnable (by a model like GPT?)	about human cognition, via experimental comparisons		"Statistical learning is sufficient for <i>star-ifying</i> "
"Silicon Sampling" LLM simulating groups of participants	How would people respond?	about human cognition, via simulating humans		"Under <X> conditions, people may <i>star-ify</i> "

Figure 1: The same experiment, assessing LLM performance on a given task—in this cartoon, presenting an LLM with a choice, and the LLM output is *star*—leads to different inferences based on the initial research questions. Researchers may make inferences about the capabilities of specific LLMs (“GPT-ology”), such as: “GPT can *star-ify*”. Alternatively, we could use LLMs as a computational model of human learning. One example inference that could be made is that “statistical learning alone is sufficient for *star-ifying*”. And finally, we could treat samples from an LLM under some conditioning contexts as illustrative of how people might respond in that manner (“under <X> conditions, people may *star-ify*”). We note that these paradigms are not exhaustive (more creative ones could appear), nor are they mutually exclusive; the same paper or research program could make various claims.

### GPT-ology: Making inferences about LLMs

There have been much effort focused on evaluating the cognitive capacities of LLMs, in order to draw inferences about the capabilities of certain models, either in isolation or relative to other LLMs (Hagendorff et al., 2023).

**LLM “traits”.** Some researchers have LLMs respond to validated psychological scales, and make inferences about their “personality” or “psychometrics” (Pellert et al., 2024; Serapio-García et al., 2023). For instance, Schaaff et al. (2023) had chatGPT fill in empathy and autism-related scales, and concluded that “the empathic abilities of chatGPT are still below the average of healthy [neurotypical] humans”.

**Bias in LLMs.** Researchers have also studied potential biases that may be present in LLMs, and that may carry over to other contexts, such as when LLMs are generating answers. Various groups (Abdurahman et al., 2023; Fischer et al., 2023; Tao et al., 2023) had GPT respond to surveys (e.g., the World Values Survey), and compared the responses to different cultural groups and other subgroups (e.g., political orientation), as a way to identify cultural or other biases in LLMs (e.g., “*GPT exhibits cultural values resembling [more Western,] Protestant countries*”; Tao et al., 2023). In a similar vein, Hu et al. (2023) had LLMs complete sentences and showed that these models exhibit social identity biases.

**LLMs as lab participants.** In addition to survey questions, LLMs can respond to tasks that one might provide in a typical cognitive science experiment. For instance, Binz and Schulz (2023) assessed GPT-3’s performance on a variety of standard cognitive psychology tasks. Similar experiments have been done for tasks like analogical reasoning (Webb et al., 2023), logical reasoning (Lampinen et al., 2023), and inductive reasoning (Han et al., 2024). A specific task that has attracted much interest is whether LLMs have “Theory of Mind”,

operationalized as whether they can correctly answer questions that require representing others’ false beliefs and desires (Gandhi et al., 2023; Kosinski, 2023; Sap et al., 2022; Trott et al., 2023; Ullman, 2023).

**Other LLM capabilities.** There are also many examples beyond standard cognitive tasks. For instance, several groups have looked at whether LLMs can generate “empathic” responses (Lee et al., 2024; Li et al., 2024; Yin et al., 2024; Zhan et al., 2024).

**Challenges.** One obvious challenge, especially with regard to inferences about “traits”, is anthropomorphism. Many psychologists would intuitively reject the notion of an AI having “personality” or other traits in the same way a human does<sup>3</sup>, and it is unclear even if these properties reliably affect behavior (e.g., generated output text) in any systematic fashion.

Another challenge is reliability (see also Outstanding Issues 3 and 5 below). LLMs are not static in the sense that their knowledge depends upon their training data, and commercial models like GPT are updated regularly, and without the ability to use past checkpoints. Retesting the model after an update (or even the next major “version”, like GPT5) might already change the results, making them obsolete.

But reliability may be difficult to achieve even in the short-term. LLM performance is susceptible to seemingly innocuous changes in the stimuli prompts, such as reversing the order of answers (in a multiple choice prompt) or information (Binz & Schulz, 2023), or trivial alterations to the stimuli (Ullman, 2023). It is not quite clear why LLM performance is so brittle to the specific prompt. One possible (and scientifically uninteresting) answer is that the specific task (or tasks like

<sup>3</sup>Although there is a longstanding and sizable research community in AI working on creating social virtual agents with distinct or tuneable “personalities”

it) were present in the training data, and the LLM has simply memorized its answer, and thus is unable to handle superficial changes. This will have to be answered for proper interpretation.

### LLMs as computational models

For the previous research paradigm(s), the inference is about the LLM itself, such as whether a particular LLM possesses some reasoning capability or bias. Another set of research questions aims to make inferences about human cognition, by assuming that, at some level, the LLM may be a computational model of human cognition. This is not an unfamiliar paradigm; cognitive scientists have long been making inferences about human learning and cognition via comparison with much simpler models (e.g., feed-forward neural networks; rule-based systems, agent simulations). The main difference with current LLMs is the sheer scale and capabilities of LLMs compared to previous generations of models, which opens up new possibilities for experimentation, and consequently, new inferences about learning and reasoning. There is still no consensus about how to appropriately compare LLMs to human cognition, although many researchers are writing about it (Blank, 2023; Frank, 2023b). For example, borrowing Marr’s levels of analysis, are we comparing LLMs and human performance at the computational level (Blank, 2023)? Or can we make inferences about processing streams in LLMs versus humans, even at a neural level (Hosseini et al., 2024)?

One approach is to compare the performance and pattern of errors of LLMs with human performance (including “biases” and errors), which might yield interesting insights into how these errors might be learnt from text data (Aher et al., 2023). For instance, GPT-3 makes Kahneman and Tversky (1972)-esque errors, such as the conjunction fallacy, rating the probability that “Linda is a feminist bank teller” larger than “Linda is a bank teller” (Binz & Schulz, 2023). Other work has also looked at whether LLMs make similar moral acceptability judgments as humans (Dillion et al., 2023; Jin et al., 2022). At the moment it is difficult to understand “why” these LLMs make these patterns of human-like decisions and errors, although perhaps there will be progress made here, such as by, for example, using techniques from developmental psychology (Frank, 2023a).

**Existence Proofs.** For some research questions, the mere demonstration of an LLM’s capabilities serves as an Existence Proof about the learnability of some aspects cognition. For instance, how much of language can be learnt versus has to be innate (Contreras Kallens et al., 2023; Piantadosi, 2023), or the gap between language and thought (Mahowald et al., 2024). The general idea is that because we know, at a coarse level, what LLMs are exposed to during training, this might allow answering questions around the sufficient conditions for learning. For instance, even though LLMs are only trained on

language data, could they learn event knowledge (about whether one event is more likely than another), solely based on the statistics of word co-occurrences in its training data (Kauf et al., 2023)? Or how does nonsymbolic learning from natural language give rise to symbolic reasoning (Geiger et al., 2023)? In a recent example that includes language and visual input, Vong et al. (2024) trained a neural network with unlabeled audio and visual input from a single infant-worn head-camera taken over several months, and found evidence for grounded language acquisition from statistical learning alone.

**Internal representations.** And finally, we could theoretically peer into the inner workings of these models to see how they ‘think’. There are lots of “probing” and other introspection methods developed in NLP to study the internal representations of such models as they learn (Belinkov, 2022). One could also peer into individual ‘neuronal’ activation, or patterns of activation, and perhaps compare that with human brain activity (Hosseini et al., 2024; Kumar et al., 2022, see also Yamins et al., 2014 for the visual cortex and Computer Vision models), to make inferences. One could also run causal interventions or other mechanistic interpretability analyses on the model itself to test how information is processed in the model (Wu et al., 2024; Yamakoshi et al., 2023).

**Challenges.** When comparing LLM outputs to human behavior, there are many issues to consider. First, how much of the behaviors (e.g., errors like the conjunction fallacy) are due to them being present in the training data (Binz & Schulz, 2023), or due to specific styles of prompting (see Outstanding Issues 4 and 5 below)?

The existence proof logic is asymmetric, especially with respect to failures (null results): if a model can “do X”, we could make a claim about the sufficient (but not necessary) conditions for a capability to emerge. But if an LLM model “cannot do X”, that cannot be used as an argument for the necessary conditions (that the LLM lacks) for that capability. For instance, LLM failure on Winograd Schema tasks might lead to an inference that world knowledge or commonsense knowledge is necessary, or that statistical learning from language co-occurrences by itself is insufficient. But these inferences do not logically follow, and could easily be falsified with a newer and more capable model. Indeed, AI development over the past decade seems to be accelerating faster than many expect. Many researchers have catalogued the current inadequacies of LLMs—for example, failing to do certain types of reasoning or logic (e.g., Borji, 2023)—and it is important to do so. But these null results do not yet lend themselves well to lasting scientific inferences, if all it takes is an engineering counterproof. Arguing from a lack of ability is less scientifically sound than an argument from the presence of one. (See also Outstanding Issue 2, below)

Studying internal representations of LLMs may not

be possible, especially with proprietary models like the GPT-series models, which are only accessible via a limited API (see Outstanding Issue 3).

### Silicon Sampling: LLMs simulating humans

Another approach that has gotten some attention is using LLMs to simulate populations or subgroups of humans (Dillion et al., 2023; Grossmann et al., 2023), which has sometimes been referred to as “silicon sampling” (Argyle et al., 2023). Researchers have used LLMs to simulate human behavior, for example in economic experiments (Aher et al., 2023; Horton, 2023) or consumer preferences (Sarstedt et al., 2024). Researchers have also looked at specific subgroups, by conditioning the model with backstories of different subpopulations, and showed that LLMs could predict behaviors like voting (Argyle et al., 2023). To the extent that LLMs accurately “encode” or “compress” human knowledge, and that conditioning the model to reproduce the behavior of certain types of humans yields behavior of sufficient fidelity (two very strong assumptions), this approach might provide a scalable way to study human or “human-like” cognition and behavior (Park et al., 2023).

**Challenges.** This approach rests on several assumptions of fidelity, which is broadly whether LLM responses can accurately reflect human responses (Argyle et al., 2023; Grossmann et al., 2023). These have to be properly tested, and are also related to issues with LLM reliability (See Outstanding Issues 3, 5 below).

One appeal of this approach is that we could use LLMs to simulate subpopulations that might be more difficult to recruit in traditional studies, such as minority groups. But these minority groups are also underrepresented in the training data (See Outstanding Issue 4). Relatedly, there are concerns about whether such simulated behavior might reflect (biased) *stereotypes* about how certain groups of people behave, rather than actual behavior.

And lastly, this approach assumes that LLMs can simulate individual humans—or rather, LLM outputs are samples from an underlying distribution that might be in some way a good approximation to real human distributions. But other approaches view LLMs output as more of a population average, or “aggregate” summary of human knowledge (e.g., a cultural technology like a library; Yiu et al., 2023). These two conceptualizations are distinct and will affect experimental design and inferences drawn—an analogy in Bayesian cognitive science is whether people are sampling from a posterior distribution (probability matching), or whether people are reasoning using the *maximum a posteriori* estimate. If LLMs are actually representing some kind of population average, but are treated as mimicking individual humans, this might lead to biases in the inferences drawn from these results.

### Other uses of LLMs

We end this section with a brief mention of several other approaches that could become more developed in the future. First, LLMs can serve as building blocks in more complex models of cognition. For instance, LLMs can be used to extract features from unstructured text, as part of a larger neurosymbolic model (Kwon et al., 2023; Zhan et al., 2023). Second, LLMs also have broad applicability in other aspects of psychological research (Demszky et al., 2023), such as to generate, classify, or annotate stimuli (Rathje et al., 2023; Ziemis et al., 2023). Many of the challenges and outstanding issues (about reliability or validity) may also apply to these use-cases.

## Outstanding issues

### 1. Which LLMs should we use?

There exists a veritable zoo of language models (e.g., LLaMA, Alpaca, Vicuna), and most variants also come with different “version numbers” (GPT-3, chatGPT, GPT-4) and sizes (LLaMA-7B, -13B, etc.). Some are open-sourced, while others are proprietary. Which should we be using? For cognitive science research, should we just focus on one model, perhaps, the biggest—or more realistically, the best that one can have access to and can afford? This may introduce another layer of inequality as well-resourced labs may have greater access to unreleased or more expensive models. Should we instead be focused on experimenting with a range of models, as is done in machine learning research? How do these choices affect reproducibility?

### 2. What inferences should we make if one LLM, but not others, can “do X”?

Scientifically, the breadth of LLM choice poses an interesting conundrum. What should we make of contexts when “smaller models” fail at a certain task, while “bigger/better models” succeed (Gandhi et al., 2023; Haggendorff et al., 2023; Kosinski, 2023), For instance, if a particular “ability” was demonstrated by GPT-4 but not GPT-3. Is there some cognitively interesting answer about the model, learning algorithm, or data that leads to those changes? How does that affect arguments about the sufficiency of statistical learning or other conditions? Unfortunately, those questions seem like they would yield engineering answers, rather than cognitive insights. Relatedly, what if a particular cognitive “ability” was restricted to one particular model (say, LLaMa-3), but not shared by other models of similar specifications? We think there are deep meta-scientific conversations that we could have as a field, rather than only in peer-review.

### 3. LLMs are proprietary commercial products updated by companies.

This issue contributes to many challenges already described earlier. Many of the papers we reviewed used closed-source proprietary models, notably GPT-3 or GPT-4. Closed-source means we do not have access to the model and data that can help guide inferences (Frank, 2023b), via introspecting model activations or understanding trends in training data. Moreover, the fact that companies regularly update their models (and perhaps even learn from previous input) might render the idea of reproducibility meaningless. Researchers might lose access to these models for a variety of sudden, unforeseeable reasons, such as economic, political, or legal (there are pending lawsuits and legislation in several jurisdictions).

This brings up a deeper question that we should be asking as a field: Should we really be yoking the success of our science to such commercial products? Of course, commercial products are important to science, providing services (e.g., Qualtrics and other software, compute) and equipment necessary for scientific research. But these conditions are different, where the actual research artefact, the object of study, is a commercial product that is not regulated and that researchers have little influence over.

Another concern about propriety models is the lack of transparency around engineering changes that are built into the model. To minimize liability concerns, many commercial LLMs have what are called “**guardrails**” built into their system. For instance, GPT will refuse to discuss dangerous (e.g., making weapons), illegal (e.g., abuse), or offensive (e.g., racist jokes) information. But some of these guardrails might also affect research, for example, constraining a model’s moral judgments to conform to a particular view. Some of these guardrails (or perhaps due to human feedback in training) may also result in idiosyncratic behaviors: for example, when asked to generate first-person negative stories, GPT tends to offer a happy ending (or a “moral of the story”). Without more transparency, it is unclear which of these behaviors are learnt from data and which are engineered into the model.

One solution is to move to open-source models, but we as a field would still have to standardize many conventions (e.g., which model, Issue 1, or reproducibility, Issue 5).

### 4. Training Data

Training data is important for making claims about learnability, and also for “simulating humans”. However, the large amount of training data presents serious issues for scientific inference. First, for some proprietary models, the source and types of training data are not public information (e.g., GPT-4). Second, even if they were

public, the sheer scale of data makes it difficult to assess (let alone control) what went into a model. Third, the data the model is trained on might contain biases that subsequently will affect its output.

One particular concern is **data leakage**. If an experimental task happens to be in the training data (e.g., the “Sally-Anne” False-Belief task, or Kahneman-and-Tversky-style fallacy items), then the model might succeed on the task simply from having seen and memorized it in its training data. Memorization is a much less interesting scientific explanation for cognitive performance on a task, but it is often a concern, given that for example, trivial alterations to the stimuli like word order or changing names can break LLM performance on a task (Binz & Schulz, 2023; Ullman, 2023). If we cannot guarantee that our tasks were not part of the training data, then a large portion of our experimental approach will be rendered invalid.

It is also worrying that LLM-produced output may form the training data for future generations of LLMs. Is human-written language on the internet like the Ship of Theseus, gradually being replaced by LLM-written approximations of human text? At what point might such language be no longer “human”?

Lastly, LLMs are trained on data that predominantly comes from Western, Educated, Industrialized, Rich, and Democratic (WEIRD; Henrich et al., 2010, 2023) countries, and even within WEIRD societies, specific subcultures (mostly young, internet-savvy users) (Abdurahman et al., 2023; Henrich et al., 2023; Tao et al., 2023). This is not representative and is worrying if LLM-based cognitive science becomes mainstream.

### 5. Reproducibility in LLM research

Issue 3 makes it impossible to guarantee continued access to a stable, dated version of a proprietary model. But even for open-sourced models, there are other issues.

**Stochasticity** is a feature of language (modeling). Instead of always returning the same sequence of words, LLMs sample words probabilistically from a distribution, and the randomness is controlled by a parameter called the temperature. A common misconception is that the stochasticity of LLMs is a flaw for reproducibility, and some studies are run with temperature set to 0 (Binz & Schulz, 2023). This is an incorrect perception: stochasticity is a *feature*, not a bug, and is true also of human cognition more generally—humans do not always give the same answer either, but psychologists have learnt to sample from people. Ideally we would be measuring (and making inferences over) the *probability distribution of tokens* in LLMs. This is directly observable with some open-source LLMs, but not always true with proprietary LLMs. This property suggests that, in many experimental tasks, scientists should be thinking about **collecting samples from LLMs** (just like how we sample responses from many people, or even the same per-

son multiple times), and **doing statistics over those samples**. This is currently not common practice.

**Prompting** or prompt engineering is the process of iterating and deciding the best natural language input to an LLM to increase the performance of the output. This is more an “art” than a science: for example, some “best practices” include explicitly giving a role or persona to the LLM (e.g., “you are an expert in  $X$ ”), with the idea that these instructions will condition the model to respond according to those instructions. Another example is “chain of thought reasoning” (Wei et al., 2022), or asking the model to think “step-by-step”, which has been surprisingly effective, and some have even likened this to human “System 1/System 2” (intuitive vs. deliberative) thinking (Hagendorff et al., 2023; Yao et al., 2023). But this also means that many prompts may, for unknown reasons, produce lower-quality output, which may lead to false negative inferences: one might argue that some paper’s failure to get GPT-4 to perform  $X$ , is because they did not find the “right” prompt.

On the one hand, humans also respond differently based on how questions are phrased, and that is no surprise to cognitive scientists, especially those with a social psychology background. On the other hand, this clashes with our mental model of how LLMs (and “computer programs” more generally) work. The sensitivity (or less charitably, brittleness) of LLM performance to prompts suggests that as a first step, we need to **report prompts and procedures in full** (including hyperparameters: temperature; date on which the model was accessed for proprietary models, etc.). We might even need to do additional steps like **permuting answer choices on multiple-choice surveys**, or **permuting the order of information in presented stimuli**. As a field, we have lots of experience with such controlled experiments with humans (e.g., counterbalancing a blocked presentation design or counterbalancing the order of presentation of stimuli). There needs to be a similar paradigm shift—and field-wide discussions—when dealing with LLMs. We need to continually revise our scientific conventions.

But at a broader level, the brittleness of LLMs to seemingly irrelevant changes in prompts, especially those that would not meaningfully affect humans, is concerning. Mathematically, it suggests that the model is overfitting. Cognitively, it suggests a learnt stimulus-response (memorization), rather than true conceptual understanding. Should we be inferring such complex cognition when we still lack an understanding of these boundary conditions (which differ from humans)?

## 6. Generalizability and longevity of Results

In science we often want to produce knowledge that is generalizable and “true”—at least until future experiments falsify our theories. But, as the current review shows, many papers are making inferences based on the capabilities of currently-released LLM models. These

models, in all likelihood and given the recent history of face-paced development, will be updated and perhaps made obsolete in a matter of months. Would then the corresponding results and inferences made on these models, also be made obsolete? (At a practical level, this also matters given the long review time in publishing). It is perhaps worth thinking about whether we as a field should consider prioritizing research paradigms and agendas that produce more generalizable, lasting knowledge that will last more than a few months.

## Conclusion

It has been an exciting 2023–2024 for cognitive science, with many papers and preprints jumping on the opportunity to study these amazingly capable models, and the scientific inferences that we can glean from them. Indeed, these models are pushing the boundaries of our understanding of cognition, allowing more creative experiments with larger data, and increasing the external validity of our science.

However, as this review points out, there are still many challenges and issues that underlie these scientific endeavors. We have tried to briefly outline what types of inferences can be licensed with such evidence, and what are concerns that might undermine such inferences. We also note that most of the work has been on making inferences about LLM abilities—these inferences might be transient anyway, as models keep improving. We hope that with more time and as these research paradigms mature, we can draw more insights about human cognition. This paper is not meant to provide a definitive framing of the field, but rather to start conversations about the outstanding issues in these new research endeavors, and we hope that it will succeed in doing so.

## Acknowledgments

We would like to thank Robert Hawkins and Judith Fan for conversations that led to this paper, and four anonymous reviewers for their helpful feedback.

## References

- Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., Golazizian, P., Omrani, A., & Dehghani, M. (2023). Perils and opportunities in using large language models in psychological research. *psyArxiv*.
- Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *International Conference on Machine Learning*, 337–371.
- Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337–351.

- Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1), 207–219.
- Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6), e2218523120.
- Blank, I. A. (2023). What are large language models supposed to model? *Trends in Cognitive Sciences*.
- Borji, A. (2023). A categorical archive of chatgpt failures. *arXiv preprint arXiv:2302.03494*.
- Contreras Kallens, P., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, 47(3), e13256.
- Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., et al. (2023). Using large language models in psychology. *Nature Reviews Psychology*, 1–14.
- Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*.
- Fischer, R., Luczak-Roesch, M., & Karl, J. A. (2023). What does chatgpt return about human values? exploring value bias in chatgpt using a descriptive value theory. *arXiv preprint arXiv:2304.03612*.
- Frank, M. C. (2023a). Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, 2(8), 451–452.
- Frank, M. C. (2023b). Openly accessible LLMs can help us to understand human cognition. *Nature Human Behaviour*, 1–3.
- Gandhi, K., Fränken, J.-P., Gerstenberg, T., & Goodman, N. D. (2023). Understanding social reasoning in language models with language models. *arXiv preprint arXiv:2306.15448*.
- Geiger, A., Carstensen, A., Frank, M. C., & Potts, C. (2023). Relational reasoning and generalization using nonsymbolic neural networks. *Psychological Review*, 130(2), 308.
- Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. *Science*, 380(6650), 1108–1109.
- Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, 3(10), 833–838.
- Han, S. J., Ransom, K. J., Perfors, A., & Kemp, C. (2024). Inductive reasoning in humans and large language models. *Cognitive Systems Research*, 83, 101155.
- Hardy, M., Sucholutsky, I., Thompson, B., & Griffiths, T. (2023). Large language models meet cognitive science: LLMs as tools, models, and participants. *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*, 45(45).
- Henrich, J., Blasi, D. E., Curtin, C. M., Davis, H. E., Hong, Z., Kelly, D., & Kroupin, I. (2023). A cultural species and its cognitive phenotypes: Implications for philosophy. *Review of Philosophy and Psychology*, 14(2), 349–386.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3), 61–83.
- Horton, J. J. (2023). *Large language models as simulated economic agents: What can we learn from homo silicus?* (Tech. rep.). National Bureau of Economic Research.
- Hosseini, E. A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., & Fedorenko, E. (2024). Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiology of Language*, 1–21.
- Hu, T., Kyrychenko, Y., Rathje, S., Collier, N., van der Linden, S., & Roozenbeek, J. (2023). Generative language models exhibit social identity biases. *arXiv preprint arXiv:2310.15819*.
- Jin, Z., Levine, S., Gonzalez Adauto, F., Kamal, O., Sap, M., Sachan, M., Mihalcea, R., Tenenbaum, J., & Schölkopf, B. (2022). When to make exceptions: Exploring language models as accounts of human moral judgment. *Advances in Neural Information Processing Systems*, 35, 28458–28473.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3(3), 430–454.
- Kauf, C., Ivanova, A. A., Rambelli, G., Chersoni, E., She, J. S., Chowdhury, Z., Fedorenko, E., & Lenci, A. (2023). Event knowledge in large language models: The gap between the impossible and the unlikely. *Cognitive Science*, 47(11), e13386.
- Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.
- Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., Griffiths, T. L., Hawkins, R. D., & Nastase, S. A. (2022). Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. *BioRxiv*, 2022–06.
- Kwon, J., Levine, S., & Tenenbaum, J. B. (2023). Neurosymbolic models of human moral judgment: LLMs as automatic feature extractors. *Social Intelligence in Humans and Robots Workshop*.
- Lampinen, A. K., Dasgupta, I., Chan, S. C., Sheahan, H. R., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2023). Language models show human-

- like content effects on reasoning tasks. *arXiv preprint arXiv:2207.07051*.
- Lee, Y. K., Suh, J., Zhan, H., Li, J. J., & Ong, D. C. (2024). Large language models produce responses perceived to be empathic. *arXiv preprint arXiv:2403.18148*.
- Li, J. Z., Herderich, A., & Goldenberg, A. (2024). Skill but not effort drive gpt overperformance over humans in cognitive reframing of negative scenarios.
- Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models: A cognitive perspective. *Trends in Cognitive Sciences*.
- Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22.
- Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2024). Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*.
- Piantadosi, S. (2023). Modern language models refute Chomsky's approach to language. *Lingbuzz Preprint, lingbuzz, 7180*.
- Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjeh, R., Robertson, C., & Van Bavel, J. J. (2023). GPT is an effective tool for multilingual psychological text analysis.
- Rogers, A., Kovaleva, O., & Rumshisky, A. (2021). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8, 842–866.
- Sap, M., LeBras, R., Fried, D., & Choi, Y. (2022). Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*.
- Sarstedt, M., Adler, S. J., Rau, L., & Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology & Marketing*.
- Schaaff, K., Reinig, C., & Schlippe, T. (2023). Exploring ChatGPT's empathic abilities. *Proceedings of the 11th International Conference on Affective Computing and Intelligent Interaction*.
- Serapio-García, G., Safdari, M., Crepy, C., Fitz, S., Romero, P., Sun, L., Abdulhai, M., Faust, A., & Matarić, M. (2023). Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.
- Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2023). Auditing and mitigating cultural bias in LLMs. *arXiv preprint arXiv:2311.14096*.
- Trott, S., Jones, C., Chang, T., Michaelov, J., & Bergen, B. (2023). Do large language models know what humans know? *Cognitive Science*, 47(7), e13309.
- Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.
- Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, 383.
- Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9), 1526–1541.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.
- Wu, Z., Geiger, A., Icard, T., Potts, C., & Goodman, N. (2024). Interpretability at scale: Identifying causal mechanisms in alpaca. *Advances in Neural Information Processing Systems*, 36.
- Yamakoshi, T., McClelland, J. L., Goldberg, A. E., & Hawkins, R. D. (2023). Causal interventions expose implicit situation models for commonsense language understanding. *Findings of ACL*.
- Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23), 8619–8624.
- Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *NeurIPS 2023*.
- Yin, Y., Jia, N., & Waksalak, C. J. (2024). Ai can help people feel heard, but an ai label diminishes this impact. *Proceedings of the National Academy of Sciences*, 121(14), e2319112121.
- Yiu, E., Kosoy, E., & Gopnik, A. (2023). Transmission versus truth, imitation versus innovation: What children can do that large language and language-and-vision models cannot (yet). *Perspectives on Psychological Science*, 17456916231201401.
- Zhan, H., Ong, D. C., & Li, J. J. (2023). Evaluating subjective cognitive appraisals of emotions from large language models. *Findings of the Association for Computational Linguistics: EMNLP 2023*.
- Zhan, H., Zheng, A., Lee, Y. K., Suh, J., Li, J. J., & Ong, D. C. (2024). Large language models are capable of offering cognitive reappraisal, if guided. *arXiv preprint arXiv:2404.01288*.
- Ziems, C., Shaikh, O., Zhang, Z., Held, W., Chen, J., & Yang, D. (2023). Can large language models transform computational social science? *Computational Linguistics*, 1–53.