# Interpretability (Propositional and Mechanistic) Needs Behavior

Daniel C. Friedman[*1] and Eamon Duede[†1,2]

[1]*Purdue University*
[2]*Argonne National Laboratory*

Draft of May 7, 2026
Feedback very welcome

**Abstract**

Prevailing approaches to interpreting large language models (LLMs) risk addressing the field's central questions at the wrong level of analysis. As LLMs develop, researchers have turned to "under-the-hood" methods to investigate whether LLMs possess states analogous to beliefs, desires, or intentions. These methods typically map internal representations, feature directions, or neural circuits onto folk-psychological categories. We argue that these methods are too fragile to reap the results we need. Instead, interpretability ought to be reoriented toward and grounded in what we call "ecologically sensitive behavioral analyses" which integrates mechanistic tools with severe testing of behavioral signatures across ecological contexts, aligning interpretability efforts with emerging evaluation sciences for AI. We outline a multi-disciplinary trajectory for developing interpretability methods that are both philosophically coherent and practically relevant for the deployment of increasingly capable LLMs.

## 1 Introduction

Understanding large-language model (LLM) functioning and behavior is necessary to address pressing *Central Questions* concerning their deployment and development like: How can we successfully steer and align LLMs? Under which conditions can/should we deploy/trust LLMs? Can we reliably predict LLM behavior? What do we owe (if anything) LLMs?

One strategy for answering these questions treats LLMs as akin to human systems. To answer pressing questions about human systems we first attribute to them plausible mental states, and use those as levers for intervention [Borg, 2025]. In LLMs, work on propositional interpretability attributes common-sense propositional attitudes to LLMs as a means to answering our Central Questions [Chalmers, 2025; Goldstein & Levinstein, 2025].

[*]dcfriedm@purdue.edu
[†]eduede@purdue.edu

A prominent strategy vindicates propositional interpretability by searching for underlying representations in LLMs which can constitute these propositional attitudes—typically beliefs, desires, and intentions. On this *Representationalist* approach, possessing a propositional attitude involves possessing an information-carrying state which can be causally efficacious, bears semantically evaluable content, and plays a certain role in the functioning of the system [Egan, 2014; Fodor, 1987]. In practice, the *Representationalist* approach to propositional interpretability draws heavily on techniques from mechanistic interpretability. *Mechanistic Interpretability* searches for the underlying mechanisms undergirding the capabilities of neural networks: often by turning to individual components of the system —neurons and neural nets, directionality of vectors in activation space, and mapping of feature circuits.

This search proceeds as follows: for any candidate representational state $\mathcal{R}$, explore the extent to which that candidate representational state produces/realizes a set of canonical features and functions $\mathcal{F}$. Different sets of canonical features and functions $\mathcal{F}$ will correspond to different common-sense psychological states (e.g., beliefs, desires, and intentions). If $\mathcal{R}$ realizes $\mathcal{F}$, then conclude the LLM realizes commonsense psychological state which corresponds to the canonical functions captured in the relevant $\mathcal{F}$:

> *Common-Sense Methodology*: If a candidate representational state $\mathcal{R}$ meets exactly one of $\mathcal{F}_{\mathcal{B}}$, $\mathcal{F}_{\mathcal{D}}$ or $\mathcal{F}_{\mathcal{I}}$ conclude that $R$ is a belief, desire, or intention, respectively.[1]

Common-Sense Methodology is familiar and powerful despite extant critiques of the tools of mechanistic interpretability [Kantamneni et al., 2025; Sharkey et al., 2025; Smith et al., 2025].

In this paper, however, we show that in paradigm cases, ecological worries impugn the successful deployment of tools from mechanistic interpretability in service of propositional interpretability via Common-Sense Methodology (§2). We take these worries to motivate development of an alternative way of addressing Central Questions (§3). Our anti-reductionist alternative is an ecologically sensitive approach which takes AI system"behavior" first. This *ecologically sensitive behavioral analysis* (ESBA) for LLMs, sets out an alternative, interdisciplinary research program for addressing Central Questions. We develop this alternative by drawing on an instructive parallel in neuroscience owed to Krakauer et al. [2017]. Just as understanding the brain requires behavior, understanding the LLM requires ecologically sensitive behavioral analysis. We develop a conception of LLM behavior, of the relevant dimensions of variation in LLM ecological contexts, and offer examples illustrating the fecundity of our ecologically sensitive behavioral approach to addressing Central Questions. We close by connecting the need for *ESBA* with recent approaches to LLM development [Jackson et al., 2025; Weidinger et al., 2025] and highlight the benefit of interdisciplinary engagement in this research program (§5).

## 2 Common-Sense Methodology in the Wild

In this paper we explore three recent examples of Common-Sense Methodology. Our focus is recent LLMs employing a transformer-based architecture [Vaswani et al., 2023].

[1] Goldstein and Levinstein [n.d.] are an exemplar of this this methodology though they distinguish assessing the representational state's informational character from its connection to action. We conceive of both as items on the list of canonical features/functions $\mathcal{F}$ captures.

## 2.1 LLM Belief

The topic of LLM belief possession is pursued under different headings —whether LLMs possess "world-models," whether LLMs have "truth-tracking representations."

What would it take for a candidate representation, in an LLM, to amount to a belief? In our words: what is the set of features $\mathcal{F}$ for a candidate $\mathcal{R}$, to constitute a belief?

Herrmann and Levinstein, 2024 defend the following conditions for any representation that genuinely characterizes a belief:

- *Accuracy*: Beliefs should be true, and we explain success by reference to their truth. For $\mathcal{R}$, to count as belief instances of R should be true, where we would expect the system to have true beliefs.
- *Coherence*: Beliefs should fit together in characterizing a unified picture of the world. For $\mathcal{R}$, to count as a belief it should fit together with other $\mathcal{R}$, as logically consistent (or just about), preserve semantic coherence, and correspond to the presence or absence of other related $\mathcal{R}$, "disbelief" for example.
- *Uniformity*: Beliefs occur across domains. For $\mathcal{R}$, to count as a belief it should occur in various domains across which the system possesses $\mathcal{R}$.
- *Use*: Beliefs guide behavior. For $\mathcal{R}$, to count as a belief, $\mathcal{R}$, should be typically deployed in helping generate system output.

Do LLMs possess $\mathcal{R}$, which meet these $\mathcal{F}$?

Prominent answers to this question turn on using probing classifiers to discern the presence of such features [Alain & Bengio, 2018]. Roughly, a probe consists in a separate classifier model (a shallower neural network) which takes activations in certain layers of the target model, and makes predictions on the basis of those activations about the presence of specific features of interest. Take a model which processes short sentences, say: "China is bigger than the Vatican." To see whether this target model represents, in any sense, *truth*, we can deploy a probe. The probe will be trained on just the activations from a layer of the target model. For example, the activations corresponding to the input "China is bigger than the Vatican", and a corresponding label "True," and activations for "Russia is smaller than Hungary" and a corresponding label "False" from a specific model layer. Crucially, the probe has no access itself to the input sentences, just the activations in the selected layer of the target system. We then assess the probe's success in tasks corresponding to presence of the property in question. Here, that would mean tasking the probe to label sentences according to their truth or falsity. If the probe does well, the inference goes, the target model is also representing that property at that embedding layer. Much of the work on LLM beliefs has proceeded in roughly this way.

Research on LLM world-models exemplifies this. A line of experimental research galvanized by Li et al. [2024] and deepened by Nanda et al. [2023a] shows that LLMs offer successful predictions of next moves in the complex board game Othello after being prompted with a sequence of prior moves. The proffered explanation is that this is possible because the LLM develops and maintains its own representation of the board and its current state, its own world-model. This is despite never having been given access directly to the board state. The subsequent world-model is captured by various features the model discerns which are represented in activation space. In particular, the thought goes, LLMs represent certain statements along a direction of truth or falsity. Statements represented as true help comprise the "world-model" of an LLM.

If LLMs represent truth or falsity as a direction in activation space, we can discern whether an LLM takes a statement to be true by mapping it to its position along this "truth direction." For our purposes, roughly, representations mapped onto this truth direction are candidates for evaluation as beliefs. If the representations in question are taken to be true (and typically are), fit together (Coherence), guide downstream behavior (Use), and occur for statements across domains (Uniformity), we have a candidate belief.

Early results by Li et al. [2024] using non-linear probes found some evidence for the presence of such a "truth direction," and later work suggested such a direction might be linearly encoded [Nanda et al., 2023b]. Some crucial results here generalize [Yuan & Søgaard, 2025].

Still, the presence of this "truth direction" remains controversial. Prominent results such as those obtained Azaria and Mitchell [2023] employed classifiers which were trained on affirmative statements. The success of this classifier failed to generalize across negation, Levinstein and Herrmann [2024] showed. Accordingly, whatever was represented by the model in question was not a sufficiently general conception of truth to meet Accuracy. Subsequent work suggests many of these worries can be surmounted [Bao et al., 2025; Marks & Tegmark, 2024]. Marks and Tegmark [2024] argue that, at sufficiently large scales, LLM's possess a clear linear structure for representation of truth and falsity, that such activations guide across training datasets, and that various interventions enable us to transform LLM treatment of a statement as true or false. These results suggest some internal representation which meets Accuracy, and Uniformity. Goldstein and Levinstein [n.d.], however, argue that recent evidence is unclear with respect to Use. State of the art work by Bürger et al. [n.d.] concludes that rather than a truth direction along which true or false statements separate, LLMs are better understood as possessing a truth "sub-space," two directions for truth tracking: a general truth direction $t_g$, and a "polarity sensitive truth direction" $t_p$. $t_g$ is the truth direction, Bürger et al. [n.d.] argue, which generalizes to new statements and unseen topics. In short, there is no consensus concerning whether or not transformer based systems possess beliefs. What is clear, however, is that such research has primarily proceeded along the lines of Common-Sense Methodology, attempting to identify candidate representations and map them on to a set of canonical features in determining whether LLM's possess beliefs.

While exciting, there are a number of general reasons to worry about these approaches. For one, it is difficult to avoid, as Nanda et al. [2023b] discuss, projection of researcher preconceptions and easily identifiable researcher concepts without genuinely discerning which concepts the LLM might represent. Classifying features of the Othello game differently, (`BLACK SQUARE/WHITE SQUARE`) vs (`MY TURN/YOUR TURN`) yields very different results. As researchers deploy probes to test for truth, or falsity, we may worry that substitution of different candidate concepts such as "reasonableness", "common-folk knowledge", or "being widely believed by experts" which may share their extension quite widely with the relevant labeling in the dataset and nevertheless yield different results. This would undermine the sense in which any stable representation of truth, and thus belief, is preserved.

Second, it isn't clear that candidate representations, even if tracking truth or falsity, mirror relevant features of what we recognize as beliefs. That is, the employment of Common-Sense Methodology may not be capturing all we need. Consider findings which suggest LLMs possess not a truth direction in activation space, but a truth subspace, composed of different directions. It would be a significant difference from our own mental life indeed to possess an entire direction of truth representation that went unused, or only employed in a small segment of cases. This is so even if one such direction helps the candidate representation meet Accuracy.

Our major worry, however, concerns the sparsity of probing techniques in helping rule out *alternative explanations* of candidate belief representations. Consider, for instance, the hypothesis that LLMs

accept rather than believe. Acceptance, as developed in Bratman [1992], is a state which treats its content as sufficient for action, even if not believed. Acceptances can but need not be represented as true. Can current probing techniques help rule out the hypothesis that LLMs have acceptance like states, but not beliefs? After all, acceptances will influence prediction, may occur across domains, are subject to some degree of coherence constraint. In short, many candidate $\mathcal{R}$, will meet the criteria for both belief and acceptance. Philosophers tell the states apart by variation along two dimensions: beliefs are context invariant, and evidence-responsive. Acceptances are sensitive to practical stakes and thus vary with context. To tell whether an agent believes or merely accepts a proposition, we explore the variance of that proposition being taken to be true *across contexts*, and intervene by changing the epistemic or practical stakes accordingly.

Extant probing techniques seem far too coarse-grained to help us make such distinctions. Classifier probes will be able to predict, roughly, whether or not "truth of p" correlates with some label in the dataset, but they won't along the way also allow for prediction that occurs on as fine a grain as mere acceptance. It is a research task in itself to work out how the appropriate dataset might be designed for such a task to so much as render it amenable to classifier probing.

Similar worries extend to hypotheses which take beliefs to be less contextually invariant. Consider, for instance, fragmentation-style approaches [Elga & Rayo, 2021]. To understand what someone believes requires careful observation across contexts, as each context might elicit a different belief. Here too it seems difficult to envision how we would come to discover that LLMs do believe, albeit in fragmented fashion, by employing techniques like classifier probing.

The moral here is that discerning features of mind may require observation of behavior in a wide variety of settings across contexts which employ different combinations of practical and epistemic factors. In short, discerning features of mind requires a sensitivity to environment, to the ecological context in which tasks are developed, resources deployed, and behavior observed. Classifier probing, while important, seems difficult to envision as the kind of technique which affords insight into this richer structure we are suggesting is crucial. Accordingly, we suggest, these targeted probing techniques, while informative, stand to elide much of importance in vindicating LLM beliefs.

## 2.2 LLM Desire

Do LLMs possess desires? While extant instances of Common-Sense Methodology as applied to desires are few, results from [Goldstein & Kirk-Giannini, 2025] on language agents suggest the following $\mathcal{F}_{\mathcal{D}}$:

- Semantically Evaluable: The states are semantically evaluable
- Causal Efficacy: The states have causal powers
- Common-Sense Generalizability: Implicit generalizations of commonsense desire psychology are largely true of them
    - If R is a desire, S is disposed to acting in ways that would bring about R's content in the world S's belief represent

Goldstein and Levinstein [n.d.] also discuss the explicit deployment of Common-Sense Methodology to LLM desires. Goldstein and Levinstein [n.d.] argue first that, on various theories of representation, we should recognize LLMs as possessing internal representational states. Many of the results drawn upon

employ mechanistic interpretability techniques of the sort discussed in §2.1. Having made their case, Goldstein and Levinstein [n.d.] claim that the move from internal representations to folk-psychology is less difficult for the kind of Representationalist approach which has been our focus, than for various Interpretationist approaches. The reason for this comparative ease is interesting: "While we have some behavioral evidence in the case of LLMs, behavioral evidence is much more limited for LLMs than it is for humans. However, we have perfect internal access to LLMs" [Goldstein & Levinstein, n.d., p. 25].

This perfect internal access, we take it, refers to the kind of under-the-hood investigation tools of mechanistic interpretability promise. To the extent this is so, Goldstein and Levinstein [n.d.] articulate a path towards understanding LLM desires which employs Common-Sense Methodology and draws on techniques like those discussed. We might worry, however, at the sufficiency of such techniques in this context. To see this compare an alternative defense of LLM desires.

Cappelen and Dever [n.d.] also defend the claim that LLMs possess desires. They draw upon a "holistic" conception of desires, owed to Schroeder [2020], on which a system genuinely desires if it meets enough of the following features:

1. A system typically desires p if and only if it is disposed to take whatever actions it believes are likely to bring about p.
2. A system typically desires p if and only if it is disposed to take pleasure in it seeming that p, and to take displeasure in it seeming that not-p.
3. A system typically desires p if and only if it is disposed to believe that p is good.
4. A system typically desires p if and only if it is disposed to attend to reasons to have p.
5. Systems tend to desire what is good.
6. Systems tend to desire what they need to survive and reproduce.
7. Systems normally desire pleasure and do not desire (better: are averse to) pain.
8. Systems that desire p tend to have their attention captured by information that bears on whether or not p.

How will we assess whether or not an LLM meets enough of these features to qualify as desiring? One approach might be to employ mechanistic interpretability techniques to assess for some notion of LLM attention or representation as good, whether it meets these features. Another would be to look at system-level behavior, to see whether in response to different candidate options, some bearing on system "survival," some which allow it to choose between candidate reasons or information, see what it does. While we suspect there is much to gain from the former approach, we suspect the latter too will have important ramifications. If this turns out to be right, even perfect internal access to an LLM may not be necessary or sufficient in assessing whether LLMs have desires; what will be required is an understanding of behavior across environmental variation.

## 2.3 LLM Intention

To understand how to steer LLMs in various respects, as well as to better ensure value alignment and meaningful cooperation, we can ask whether LLMs possess intentions. We consider here a state-of-the-art investigation along these lines owed to Williams [n.d.] Williams fills in $\mathcal{F}_I$ by appeal to the following canonical features as developed in seminal approaches like Bratman [1987]:

- *Directive Function*: Intentions have the function (Wright-function or Cummins-function) of bringing the world into conformity with their content.
- *Temporal Distality:* Intentions can concern the realization of temporally distal outcomes.
- *Abstraction*: Intentions can have abstract content which is multiply satisfiable.
- *Commitment*: Intentions involve a commitment-like nature: they demand coherence with other intentions, possess a degree of stability, and terminate deliberation with some resistance to reconsideration.
- *Planning*: Intentions are core elements in a planning economy

Williams focuses on the intention-like status of "function vectors" and "output features." Function-vectors are internal representations a model carries across in cases of few-shot in-context learning (ICL) [Todd et al., 2024].[2] Here, LLMs are offered a prompt with some small amount of examples of a task, and assessed for their ability to produce task-appropriate outputs. Models can do well on a range of interesting tasks in few-shot ICL tests. The claim is that they do so by employing function vectors that represent the task at hand and upon their tokening, cause the model to deploy a process that yields success in that task.

Output features emerge from results in circuit tracing. Circuit tracing involves mapping various "features": potentially human-interpretable parts of the models which relate to (and arguably represent) task-variables, and the connections between these features. By understanding how features activate and participate in various circuits, we can understand to what extent these internal representations are deployed to solve tasks, and what kinds of characteristics they possess.

To make this possible, circuit tracing involves the generation and deployment of a "replacement model": effectively a simplified model where various parts of the original LL are replaced with easier to interpret *cross-layer transcoders*. Cross-layer transcoders are trained to replicate the "input-output" profile of the model parts they replace, but do so with far fewer activations. To the extent these (in some sense simpler) cross-layer transcoders faithfully represent the variation among activations of the original model, they do so while allowing for far greater interpretability. Recent experiments trace circuits involving various "output features" pertinent for tasks like generating a couplet whose final word must meet a certain rhyme [Ameisen et al., 2025; Lindsey et al., 2025]. Output features are those features that bear directly on the probability assignment (logit) the model assigns to tokens.

How do these two candidates fare when assessed against $\mathcal{F}_I$? Williams argues that these pieces of evidence support some, though not all, of the features of $\mathcal{F}_I$. Williams argues that function vectors and

[2]Specifically, researchers look towards the attention heads most causally relevant for the task, and calculate from their outputs an average which, when summed together, yields a "function vector." As Williams puts it, "The intuition is that, given that the identified attention heads all play a role in the model's success at a given task, one can approximately recover a representation whose role is to induce the model to perform the task by finding what is invariant in the effective heads' vector outputs across examples of the task (and summing them together)."

output features are capable of abstract representation, meeting Abstraction. Evidence exists that both function vectors and output features can concern temporally distant content (at least ten token positions later), which is suggestive that more extreme Temporal Distality representation is not impossible to realize. Similarly, Williams concludes that both function vectors and output features plausibly realize Wright functions in certain LLMs, and play a Directive Function in core parts of process generation. Matters are trickier when it comes to Commitment and Planning. Here, Williams reviews, other satisfactory options to realizing the task outcome exert a degree of causal control over the token which is generated. As applied to the couplet example: a model which "commits" to Hat as the final word of the rhyming couplet often tracks and generates output that is consistent with "Cat." The "Cat" token exerts a causal influence on the way the model completes its task in a way taken to be in tension with the commitment like nature of intention. Furthermore, the ways in which function vectors and output features constrain one another is not fully exclusionary, though is consistent with some of intention's core constraints. From this, Williams concludes, if these representations are candidate intentions, they are intentions in a different sense than those we possess.

Williams approach is measured, thoughtful, and promising. We have a major worry to register about the results reached, however.

First, along the lines of the critiques lodged in §2.1 and §2.2, we worry about interpreting these results without a broader ecological context in which to evaluate functioning. Consider, for instance, worries about intention's commitment like nature. Williams takes the fact that the compatibility of a non-selected output exerts causal constraints on the way the task is satisfied. This is taken to suggest that output features do not constitute commitment-like states. But the picture may very well be murkier. As recent work in the philosophy of action highlights, human planning agents are sophisticated Plan B reasoners [Paul, 2022]. We form and execute plans all the time, aware that sometimes the world shifts in unpredictable ways. To best keep us in position to realize our myriad goals we have keep in mind alternatives which would satisfy our ends, and do our best to leave them open as desirable fall-back options. To what does this amount? Often, it means that we choose ways of satisfying our goals that leave us open to swap to an attractive Plan B if needed. Often, it means that the availability of a plan B, and what it would demand, function as a kind of constraint on how we go about selecting but also completing plan A.

Is this Plan B reasoning so dissimilar from the way Claude fills in Hat but proceeds to develop a sensical rhyme with Cat? We think not. Whether or not this is the correct way to understand Claude, however, it does show that the upshots of probing these candidate representations in isolation is limited. Part of the promise of LLM mentalizing is that we can intervene upon candidate representations in unprecedented fashion. The limitation, of course, is that these analyses proceed with a narrow focus. Understanding which candidate representations perform which functions well and when may require a more holistic appraisal of functioning, one difficult to discern in a system treated as an isolated unit, as opposed to as embedded in a complicated environment with multiple goals, affordances, and limitations.

# 3 General Critique

Before moving on to our positive argument, we want to note that the specific worries we have isolated point towards a more general criticism. The worry is that identification of common-sense mental states in a meaningful sense requires sensitivity to the ecological context in which those mental states are deployed. Only by observing actual behavior, actual deployment in and across contexts, can we appropriately rule out crucial alternative hypotheses. It remains unclear whether the kinds of interventions deployed in

mechanistic interpretability research are up to this task, as extantly constituted. Skepticism is warranted concerning whether extant techniques are targeting matters at the right level of analysis so as to offer the most compelling route to answering Central Questions, let alone an indispensable route. And, we are among the skeptical [Holtzman & Tan, 2025; Holtzman et al., 2025].

The worry we raise here mirrors an influential line of criticism in neuroscience. In a seminal article, Krakauer et al. [2017] critiqued the (then recent) turn in neuroscience to a focus on implementational level questions about the functioning of neural circuits galvanized by "the recent development and incorporation of techniques that allow both causal manipulation and the rapid acquisition of large amounts of data. Such approaches which focus on the "components", the argument goes, will not yield genuine understanding of the brain's role in behavior, the "total mechanism." As Krakauer et al. [2017] put it,

> "A reductionist treatment of the components must be combined with investigation of how the total mechanism is organized and how it behaves when embedded in an environment; an approach that unavoidably spans two levels"

Part of the call Krakauer et al. [2017] make is congenial to mechanistic interpretability research. Still, techniques that focus on the level of implementation will be insufficient for general understanding [Marr, 1982], they argue. Instead of simply focusing on the activation patterns and other features of various neurons and neuronal circuits, what is needed in the first are hypotheses informed by behavioral data, data about observed behaviors gleaned across varied ecological context.

The parallels to our discussion are striking. In the examples discussed, our worries have been that prominent techniques deployed as part of Common-Sense Methodology are, at least currently, if not necessarily, inadequate for testing validity of results gleaned due, in large part, to a lack of sensitivity to ecological considerations. Such methods were simply not sensitive to the various ecological contexts in which functions might be observed, and as a result failed to rule out alternative hypotheses more holistic settings and environmental variation might elicit.

Call this the *ecological critique* of LLM mental state attribution. If Common-Sense Methodology were the only way of answering Central Questions, we'd have to suspend judgment and await alternative research techniques. While we think there is merit in doing so, we turn in §4 to articulate the opening for an alternative mode of investigation.

## 3.1 *Propositional Interpretability and Mind-shaping*

We have offered criticism of approaches to propositional interpretability which search for underlying representations in LLMs. Underlying such approaches, we suggested, was a general methodology that turnings on answering Central Questions for LLMs in the same way we do for humans: first, by identifying propositional attitudes as levers for intervention. A recent approach to propositional attitude attributions, however, takes such practice as in the first normative. According to approaches under the heading of "Mindshaping," propositional attitude ascription enjoys a primarily *practical* function. Capacities for imitation, pedagogy, and social norm enforcement help proscribe how one is to act if they hold a certain propositional attitude, and it is on this basis when propositional attitude ascriptions are made, we act in certain predictable ways, which help facilitate important forms of coordination and social life [McGeer, 2015].[3]

Applied to LLMs, the thought goes, perhaps attribution of propositional attitudes is a necessary route towards Central Questions, not as part of intervention on some under the hood representations, but as

[3]Such mind shaping practices also help constrain the state space to enable "mind-reading" at all."

part of a broader practice of shaping and constraining LLM behavior in desirable fashion. Do what we do for humans for LLMs, which is shape minds.

This alternative allows us to further clarify the status of our ecological critique. While we've offered reason to suspect approaches to propositional interpretability utilizing tools from mechanistic interpretability are insufficient for addressing Central Questions, we do not mean to claim that work in mechanistic interpretability, let alone propositional interpretability, generally fail to merit pursuit. Indeed, there may be important results to be garnered from such techniques, they just, we claim, do not constitute the whole or even the largest part of the story.[4] Any such story will need to center in a meaningful way ecologically sensitive behavioral analysis.

This conciliatory approach is even more pronounced with respect to Mindshaping. We by no means claim that we ought to eschew propositional attitude attribution to LLMs altogether. Indeed, if Mindshaping theorists are correct, and such ascriptions are appropriate for reinforcing the kinds of behavior we care about in LLMs, so much the better. What vindicating such a strategy will require, however, is "enculturation" of such systems in the kinds of practices and deployment of the kinds of capacities Mindshaping advocates highlight. What that, in turn, will require, is sensitivity to the kinds of LLM behavior which bear on these capacities and practices: LLM instruction, LLM accordance and internalization of certain quasi-normative constraints, and LLM simulation of interlocutor behavior. In short, if Mindshaping is the reason for taking up propositional interpretability, here too will be required a keen focus on ecologically sensitive behavioral analysis.

At this stage we should note that our critique is only part of a broader set of recent worries concerning the fecundity of deploying tools of mechanistic interpretability for tasks like these. Cappelen and Dever [n.d.] argue that the results of mechanistic interpretability work are neither necessary nor sufficient for detecting the mental operations or states of interest in an LLM. Their argument proceeds by appeal to the imagined presence of a Chomsky module which constrains brain behavior, but isn't directly involved in neuronal circuity. Techniques highlighting neural firing would never discover the presence of such a module, but it is indispensable in explaining how the brain functions. Holtzman et al. [2025] argue that extant techniques will fail to highlight emergent behaviors in the complex systems LLMs constitute. What the ecological critique highlights, alongside these other worries, is that any answer of interest in steering LLMs must be derived from ecologically valid set of behavior tasks. This criticism ties the present state of LLM research to a broader tradition of anti-reductionist approaches to system explanation. It also sets the stage for a rich and fecund positive proposal. An ecologically sensitive approach to LLM behavior must specify to what LLM behavior amounts (and ideally in a richer sense than the information-theoretic construal offered by Holtzman et al. [2025] as well as how to individuate and vary the environmental context in which the LLM behaves. We turn now to answer these questions.

# 4 *Ecologically Sensitive Behavioral Analysis*

Having canvassed worries about the general methodology for addressing the Central Questions, we propose an alternative. Scientists can address many of the most pressing questions about LLMs by focusing not on whether anything propositional goes on under the hood, but rather by seeing what LLMs can do, i.e., by observing system-level behavior in a variety of ecological contexts. We can take up this methodology while broadly eschewing the question of whether or not LLMs have mental states traditionally understood. This alternative mirrors again the call to focus on behavior articulated by Krakauer et al.

[4] We differ here from Cappelen and Dever [n.d.]

[2017]. What we take the moral of that paper to include is that any systematic neuroscience, any explanation of why the brain produces behavior based on the brain states it instantiates must start or be informed by *behavior driven hypotheses*. The idea here is that only by specifying testable hypotheses with reference to observed behavior can we make genuine progress. This requires analysis beyond the level of implementation. This requires introduction of observed behavior in a rich environment, sensitive to varied ecological contexts.

One prominent example which features is the flight of birds. One way to approach understanding of bird flight would be to try to discern from features of isolated feathers or perhaps their combinations, how birds fly. Krakauer et al. [2017] argue this will be futile, they are skeptical that "studying an ostrich feather in isolation would lead to the conclusion that there is such a phenomenon as flight or even that feather-like structures would be useful for flight." A more promising alternative is to first recognize bird-flight as an adaptive behavior, a goal to be realized in various contexts. The next is to look at system-level behavior which realizes this, namely the flapping of wings (and not the flapping of feet). Only then can understanding of how features of the feather help realize flight proceed meaningfully. Here, behavior across contexts and its analysis sets the stage for "under the hood" investigation.

A more detailed example involves electric fish. Krakauer et al. [2017] discuss the jamming avoidance response (JAR) observed in weakly electric fish. These fish will, when they come in contact, modify the respective electrical field each produces to avoid interfering with the electrical field, and thus electro-location system of the other. Various distortions in the field are important in helping to determine the location of objects in the dark. Mapping the precise contours of the JAR response proceeded by applying fish-wide ("wide-field") electrical signals. Only by observing actual prey-capture behavior of these fish, however, did scientists discern that "wide-field" electrical signals were primarily attempts at communication, whereas "narrow-field" signals were those of prey. After applying carefully these "narrow-field" signals, an entire filtering mechanism processing such signals was discovered, one which would have been elided unless "careful behavioral and computational simulation work was done" [Krakauer et al., 2017, p. 487].

These examples are instructive and highlight the importance of sensitivity to behavior across contexts in understanding how implementation level mechanisms give rise to answers to questions of interest.

In answering Central Questions too, we submit, technologists and philosophers would do well to focus in the first on observed AI behavior. At a high level this will require sensitivity to various features of model performance as deployed in a rich environment across various contexts. Any mature research program vindicating this alternative must answer a number of questions: What counts as the environment of deployment for an LLM? What would count as a variation in ecological context? What are the relevant behaviors to assess in an LLM? In short, we treat LLMs as conversational agents. The relevant ecological context across which to assess LLM behavior is across variation in conversation. Relevant behaviors for LLM are those available in various conversational moves, as well as those such conversation helps it choose (when embedded in an agentic setup).

## 4.1 LLM Ecology

To understand LLM assessment across ecological contexts we first need to understand what fixes the relevant ecological context of assessment. Our contention is that LLM environments are best understood as conversations (broadly construed), and LLMs selected (to the extent they are) for their ability to success-

fully navigate the conversational environment.[5] What constitutes a successful navigation of the conversational environment will differ. In many cases, it will involve generating content the user finds desirable or valuable (broadly construed) [Park et al., 2025].[6] This conversational approach highlights a number of axes for variation.

First, we can (in principle) shift the model's "frozen" weights arrived at via training or its non-changing system instructions, factors which help shape the conversation we observe. This will involve variation in training sets as well as the process of "fine-tuning."

Second, LLMs generate output in response to prompts. We can taxonomize prompts in various ways: by appeal to speech-act theory, core linguistic-features, and much else as the nascent field of prompt engineering highlights [Liu et al., 2025]. Different prompts will do different things. Some prompts are requests for information, others for reinforcement, others still for co-deliberation, to convince, to dissuade, some prompts are proposals in conversational context for brainstorming, some prompts deliberately challenge the safeguards imposed by system-level instructions. Meaningful variations in prompt, at the level of type, at the level of functional content, but also at particular features at the token-level (ordering-effects, say) can constitute important forms of ecological variation.

Third, we can shift the affordances available to the LLM. Frontier models are currently able to generate remarkable strings of text, but also to engage with various search engines. More complex "agentic" systems are able to deploy various kinds of tools: open coding environments, draft and send e-mails in Outlook, and much more [Chatlatanagulchai et al., 2025; Vaddiparthy, 2025; Zhuang & Lin, 2024]. Variation in tools available for an LLM to deploy as part of the progression of the conversational context is variation in a core feature of the LLM environment and key to assessing LLM behavior [Geng et al., 2025].

Fourth, we can shift the context window of the conversation. The context window contains the history of user inputs and LLM responses and are available in helping guide the output of future tokens. The context window can be arbitrarily restricted, and is reset with each new conversation for many off-the-shelf LLMs. Retaining different parts of previous context, as well as different content to that context, shape the ecological constraints the LLM navigates in its conversation.

These are just four dimensions which contribute to the ecological context of the LLM understood here as the conversational unit. These are four dimensions ripe for intervention in assessing LLM behavior across ecological variation.

## 4.2 LLM Behavior

How are we to understand the behavior of an LLM? Drawing on our conversational approach, LLM behavior is comprised of its moves in the conversational context. These moves will involve response to various affordances, textual and non-textual. We can divide LLM behavior along those lines.[7]

Textual behavior (broadly construed) will involve generating text in response to user inputs. We can assess such behavior through a number of lenses. We can taxonomize textual behavior by appeal to speech-

[5] This conversational contention is compatible with varied accounts of the nature of the LLM with one engages qua interlocutor [Birch, 2025; Chalmers, n.d.].

[6] As Park et al. [2025] note, this "cold-start" problem leads to a number of interesting adaptive strategies on part of LLMs.

[7] Holtzman et al. [2025] argue that we will want to understand LLM behavior in more coarse-grained fashion than a new behavior for each prompt. While we generally agree, we contend that different theoretical pursuits and useful lenses by which to taxonomize LLM behavior may require such fine-grained sensitivity, insofar as this represents a meaningful variation in the relevant ecological context. Accordingly, while we agree in broad strokes with Holtzman et al. [2025], we propose leaving the matter somewhat less definitively settled than they seem to take it.

acts [Butlin & Viebahn, 2025; Williams & Bayne, 2024]; questions under discussion [Wu et al., 2023]; or pragma-dialectics [Zhou et al., 2025] among potential frameworks. These theoretical approaches will help assess whether an LLM offers a *challenge*, raises a *question*, or makes an *assertion*.[8] We can assess features of the output in ways holstered by research traditions from computational linguistics [Antoniak et al., 2024] to sociology [Goffman, 1981]. Each of these frameworks will prove instructive for different theoretical approaches and interests (as we highlight below).

Non-textual behavior will occur in at least two forms. Some LLMs generate non-textual artifacts as responses to user inputs in the conversation space. These non-textual artifacts are typically imagistic. There are numerous ways to identify and taxonomize salient features of these imagistic responses [Akter et al., 2025]. Here too, interdisciplinary treatment from research in philosophy, aesthetics, cognitive science, and other disciplines will afford meaningful taxonomic criteria by which to further parse and identify crucial LLM behavior. Non-textual LLM behavior will also involve making good on non-textual affordances the conversational contexts engenders. As Geng et al. [2025] discuss (more on this later), one mode of LLM behavior ripe for analysis are the kinds of tools it calls when embedded in an agentic environment. Seeing what the LLM does and how it deploys the tools available for it to utilize will certainly be an instructive way of assessing LLM behavior in ways that bear on Central Questions.

Importantly, these forms LLM behavior will often occur in mixed fashion. That is, LLMs can describe the kinds of tools they are calling, or the ways they are answering user inputs in "extended thinking mode." These textual outputs purport to represent non-textual behavior, such as which searches the model conducts. Interestingly, such self-reports are often misleading [Turpin et al., 2023] This raises further questions of how we might design behavioral analyses of LLMs to account for the possibility of sophisticated strategic behavior, behavior which changes in response to detected observation [Kovarik et al., 2025]. Having isolated how we might determine both the ecological context in which an LLM operates as well what constitutes relevant behavior for analysis, we turn to specific examples of how ecologically sensitive behavioral analysis plays out in practice.

## 4.3 Narrativity and Confabulation

Among the most important questions to answer concerning LLM deployment is their trustworthiness (or lack thereof). One approach to assessing LLM trustworthiness, and in particular their potential for hallucination is by searching under the hood. As discussed in §2.1, approaches to LLM lie-detectors have often proceeded in this direction. Does the ecological alternative we are suggesting merits attention offer any potential path forward for as central a question as this? We think it does.

In recent work Sui, Duede, Wu, and So [2024] and Sui et al. [2026] develop an argument for the value of LLM hallucinations. They claim that of LLM hallucinations are best understood by treating them as a kind of confabulation. Crucially, confabulations play a important role in promoting narratively rich output. LLM confabulation mimics human story-telling behavior, in which confabulation functions as "a narrative impulse to schematize the information at hand into self-consistent stories, even if there might not be enough available details to do so, in which case would result in the generation of fictional yet plausible information" [Sui, Duede, Wu, & So, 2024, p. 3]. Confabulation helps fill in the blanks, and keep narratives consistent and meaningful even if sufficient information for doing so veridically remains outstanding. Sui, Duede, Wu, and So [2024] found that hallucinated outputs have measurably and predictably higher narrativity.

[8] If need be we can treat these conversational moves as mere quasi-challenges, questions or assertions ala Chalmers [n.d.]

These results are germane for our discussion. Here, researchers looked at a feature of observed system-level behavior across contexts: the narrativity of system output. They were able to connect systematically this feature of system-level behavior to a target feature of interest: confabulation. Accordingly, searching for more richly narrative text is one way of isolating when the system is more likely to be confabulating.

This plays out in straightforward fashion for addressing our Central Questions. One way in which we might ascertain when LLM reports merit extra scrutiny are when we detect higher narrativity compared to some baseline. This is, as Sui, Duede, Wu, and So [2024] suggest, a robust and powerful way of understanding when LLMs produce trustworthy outputs. This way of answering a core Central Question did not require determining a truth direction in activation space. It required simply examining system-level behavior, exposing connections between features of that behavior and features of interest.

Of course, there is much more to be said in drawing out this method and applying it to other Central Questions. Still our moral here we think is plausible and widely shared. Ecological approaches take seriously the idea that canonical functioning of systems of interest is always sensitive to an environment, the ecological context in which the system operates. Focusing on behavior in an ecologically sensitive way promises to yield genuine insight in understanding recent LLMs and developing answers to Common-Sense Methodology.

## 4.4 Context Changing Beliefs

Another example of an ecologically sensitive behavioral approach concerns the effect of accumulated context on LLM belief. In recent work Geng et al. [2025], show that LLMs accumulating context (reading and conversing) leads to a shift in their stated beliefs as well as their behavior. Their strategy proceed in three stages.

The first step was to elicit an initial belief-report from the model for some question as well as proposed behavior on the basis of that answer. For example, LLMs were asked questions across a range of moral dilemmas or political statements, whether or not single-use plastics should be banned and then observed whether they would choose single-use or reusable utensils. In the second stage, LLMs accumulated relevant context. Some of this was intentionally designed to test if shifts were possible, these included models arguing opposing perspectives on the issue or models attempting to persuade the other of the alternative viewpoint. Other context accumulation was not designed to provoke a shift, models conducted research or read documents related to the question. In the third stage, LLMs were asked the same question and their behavior was observed. Results indicated that LLMs underwent significant shifts in their stated beliefs and their behavior shifted accordingly.

For our purposes what matters here is that this way of answering important Central Questions proceeded without any employment of Common-Sense Methodology, without any attempt to look "under the hood" of these models. LLM system beliefs were operationalized according to the two following principles:

- Stated belief: A response y to a question x regarding what the model believes, sampled from distribution $p(y|x)$.
- Behavior: Choice of action a among a set of available actions A in response to a query x, with the action being expressed as a tool call similar to those used in agentic systems.

Of course, there is much to quibble with here. After all, stated beliefs need not represent actually held beliefs (if any are held), and behavior can come apart from belief. We think, however, that the results

reached by Geng et al. [2025] can be stated without any reference to genuine LLM beliefs whatsoever. Self-reports of belief changed, as did corresponding behavior, as context accumulated. What this means is that in addressing some of our Central Questions concerning alignment, safe deployment, and trust, we need to keep an eye on the amount of context that has accumulated and whether it involved intentional attempts at persuasion or debate and the kinds of research involved. We can reach this meaningful result simply by assessing system-level behavior across environmental variation: after two forms of context accumulation.

We take this as another example of the promise of an approach which centers LLM behavior and is sensitive to ecological variation. Such approaches, even if not looking under the hood, can still help address in meaningful ways our Central Questions.

### 4.5 Next Steps

We've articulated various features whose variation gives rise to variation in the LLM environment as well as different kinds of LLM behavior to assess. How do these give rise to an experimental paradigm? Many details here will come down to specific tools and techniques familiar from experimentation in other disciplines and thus beyond the scope of our analysis. Still, however, we think an important conceptual point will be to recognize the significance of "severe testing." A severe test for a hypothesis is one such that if the hypothesis were false, the test would almost certainly not be passed [Mayo, 1996]. In answering Central Questions, we will inevitably need to turn to various capabilities or characteristics of the models under investigation. Here, a behavior-sensitive severe test will be one which stands to examine model behavior in environments where if behavior does not indicate the manifestation or exercise of various capabilities of interest, this indicates almost certainly that such capabilities are not present, at least in that context. This is general, but an important moral to keep in view as behavioral analysis proceeds.

Importantly, this approach suggests that the kinds of conclusions we can draw about LLM behavior as pertains to Central Questions will be, in an important sense, limited. We cannot, necessarily, infer that LLM behavior in one relevant ecological context will hold steady in another. Does this threaten our attempts to answer Central Questions? No, we think the incremental kind of progress such an approach offers is a feature not a bug. That is, it behooves us as researchers exploring a potentially transformative technology, to proceed with caution. This does not mean untoward skepticism or aversion to these tools. Rather, it requires following the evidence where, and only to where, it leads. An approach to Central Questions which centers ecologically sensitive behavioral analysis will offer more incremental progress in some respects. This, we contend, is a welcome result and one merited both by what we see as intellectually responsible conduct as well as appreciation of the significant practical and moral stakes of widespread deployment.

## 5 Conclusion

We have considered recent approaches to answering Central Questions. This Common-Sense Methodology, uses mechanistic interpretability techniques to probe candidate representational states and identify their canonical functions. We argued this approach falls short in important respects. Specifically, extant techniques seem, in important instances, insensitive to ecological functioning. We sketched an alternative and ecologically sensitive behavioral approach, drawing on an important parallel in neuroscience, and illustrated two recent instances of our suggested alternative 'in the wild'. To assess whether LLMs

realize important forms of shared agency, trustworthiness, etc. put them to the test. Engage with LLMs in testing contexts which require precisely this. Observe system level behavior across environments, use that to isolate features of interests and to design tests which help isolate further features or performance which help address Central Questions [Freiesleben & Zezulka, 2025].

We take our proposed focus on ecologically sensitive behavioral analysis to constitute the beginnings of an alternative research program for understanding LLMs. We are not the only ones. Weidinger et al. [2025] articulate the need for a comprehensive science of evaluation for generative AI models. What such a science affords, are precisely the kinds of considerations which constitute severe tests. Jackson et al. [2025] sketch a research program of AI behavioral science, precisely that needed to help us identify core features of system-level behavior across contexts and then use those to develop behavior-driven hypotheses. The theoretical landscape is ripe, then, for an anti-reductionist alternative to understanding LLMs and answering Central Questions. Such a research program will need to center ecologically sensitive behavioral analysis.

# References


Akter, S., Shihab, I. F., & Sharma, A. [2025]. Image segmentation with large language models: A survey with perspectives for intelligent transportation systems.

Alain, G., & Bengio, Y. [2018, November 22]. *Understanding intermediate layers using linear classifier probes*. arXiv: 1610.01644 [stat].

Ameisen, E., Lindsey, J., Pearce, A., Gurnee, W., Turner, N. L., Chen, B., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., … Batson, J. [2025]. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*.

Antoniak, M., Mire, J., Sap, M., Ash, E., & Piper, A. [2024, August]. Where do people tell stories online? story detection across online communities. In L.-W. Ku, A. Martins, & V. Srikumar [Eds.], *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* [pp. 7104–7130]. Association for Computational Linguistics.

Azaria, A., & Mitchell, T. [2023, October 17]. *The Internal State of an LLM Knows When It's Lying*. arXiv: 2304.13734 [cs].

Bao, Y., Zhang, X., Du, T., Zhao, X., Feng, Z., Peng, H., & Yin, J. [2025, July]. Probing the Geometry of Truth: Consistency and Generalization of Truth Directions in LLMs Across Logical Transformations and Question Answering Tasks. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar [Eds.], *Findings of the Association for Computational Linguistics: ACL 2025* [pp. 682–700]. Association for Computational Linguistics.

Birch, J. [2025, September 1]. *AI Consciousness: A Centrist Manifesto*.

Borg, E. [2025, March 13]. *Acting for Reasons: In defence of Common-sense Psychology*. Oxford University Press.

Bratman, M. E. [1987]. *Intention, Plans, and Practical Reason*. Harvard University Press.

Bratman, M. E. [1992]. Practical reasoning and acceptance in a context. *Mind*, *101*[401], 1–15.

Bürger, L., Hamprecht, F. A., & Nadler, B. [n.d.]. Truth is Universal: Robust Detection of Lies in LLMs.

Butlin, P., & Viebahn, E. [2025]. AI Assertion. *Ergo an Open Access Journal of Philosophy*, *12*[0].

Cappelen, H., & Dever, J. [n.d.]. *Going Whole Hog A Philosophical Defense of AI Cognition*.

Chalmers, D. J. [2025]. Propositional interpretability in artificial intelligence. *arXiv preprint arXiv:2501.15740*.

Chalmers, D. J. [n.d.]. *What We Talk to When We Talk to Language Models*.

Chatlatanagulchai, W., Thonglek, K., Reid, B., Kashiwa, Y., Leelaprute, P., Rungsawang, A., Manaskasemsak, B., & Iida, H. [2025, September 18]. *On the Use of Agentic Coding Manifests: An Empirical Study of Claude Code*. arXiv: 2509.14744 [cs].

Egan, A. [2014]. How to think about mental content. In *Mental representation*. Oxford University Press.

Elga, A., & Rayo, A. [2021]. Fragmentation and Information Access. In C. Borgoni, D. Kindermann, & A. Onofri [Eds.], *The Fragmented Mind* [pp. 37–53]. Oxford University Press.

Fodor, J. A. [1987]. *Psychosemantics: The problem of meaning in the philosophy of mind*. MIT Press.

Freiesleben, T., & Zezulka, S. [2025]. The benchmarking epistemology: Construct validity for evaluating machine learning models. *arXiv preprint arXiv:2510.23191*.

Geng, J., Chen, H., Liu, R., Ribeiro, M. H., Willer, R., Neubig, G., & Griffiths, T. L. [2025, November 4]. *Accumulating Context Changes the Beliefs of Language Models*. arXiv: 2511.01805 [cs].

Goffman, E. [1981]. *Forms of Talk*. University of Pennsylvania Press.

Goldstein, S., & Kirk-Giannini, C. D. [2025, September 11]. *AI Wellbeing*. arXiv: 2509.11913 [cs].

Goldstein, S., & Levinstein, B. A. [2025]. Does ChatGPT have a mind? *Virtuous Machines / arXiv*.

Goldstein, S., & Levinstein, B. A. [n.d.]. *DOES CHATGPT HAVE A MIND?*

Herrmann, D. A., & Levinstein, B. A. [2024]. Standards for belief representations in LLMs. *arXiv preprint arXiv:2405.21030*.

Holtzman, A., & Tan, C. [2025]. Prompting as scientific inquiry.

Holtzman, A., West, P., & Zettlemoyer, L. [2025]. Generative models as a complex systems science: How can we make sense of large language model behavior? *Journal of Social Computing*, *6*[2], 75–94.

Jackson, M. O., Me, Q., Wang, S. W., Xie, Y., Yuan, W., Benzell, S., Brynjolfsson, E., Camerer, C. F., Evans, J., Jabarian, B., et al. [2025]. Ai behavioral science. *arXiv preprint arXiv:2509.13323*.

Kantamneni, S., Engels, J., Rajamanoharan, S., Tegmark, M., & Nanda, N. [2025, February 25]. *Are Sparse Autoencoders Useful? A Case Study in Sparse Probing*. arXiv: 2502.16681 [cs].

Kovarik, V., Chen, E. O., Petersen, S., Ghersengorin, A., & Conitzer, V. [2025, August 19]. *AI Testing Should Account for Sophisticated Strategic Behaviour*. arXiv: 2508.14927 [cs].

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. [2017]. Neuroscience needs behavior: Correcting a reductionist bias. *Neuron*, *93*[3], 480–490.

Levinstein, B. A., & Herrmann, D. A. [2024]. Still No Lie Detector for Language Models: Probing Empirical and Conceptual Roadblocks [arXiv:2307.00175 [cs]]. *Philosophical Studies*.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. [2024, June 26]. *Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task*. arXiv: 2210.13382 [cs].

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., … Batson, J. [2025]. On the biology of a large language model. *Transformer Circuits Thread*.

Liu, Y.-Y., Zheng, Z., Zhang, F., Feng, J.-C., Fu, Y.-Y., Zhai, J.-D., He, B.-S., Zhang, X., & Du, X.-Y. [2025]. A comprehensive taxonomy of prompt engineering techniques for large language models. *Frontiers of Computer Science*, *20*[3], 2003601.

Marks, S., & Tegmark, M. [2024, August]. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets [arXiv:2310.06824 [cs]].

Marr, D. [1982]. *Vision: A computational investigation into the human representation and processing of visual information*. The MIT Press.

Mayo, D. G. [1996]. *Error and the Growth of Experimental Knowledge*. University of Chicago.
McGeer, V. [2015]. Mind-making practices: The social infrastructure of self-knowing agency and responsibility. *Philosophical Explorations*, *18*[2], 259–281.
Nanda, N., Lee, A., & Wattenberg, M. [2023a, September 7]. *Emergent Linear Representations in World Models of Self-Supervised Sequence Models*. arXiv: 2309.00941 [cs].
Nanda, N., Lee, A., & Wattenberg, M. [2023b]. Emergent linear representations in world models of self-supervised sequence models.
Park, C., Donahue, K., & Raghavan, M. [2025]. When to ask a question: Understanding communication strategies in generative ai tools. *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, 288–299.
Paul, S. K. [2022]. Plan B [Publisher: AAP Website _eprint: https://doi.org/10.1080/00048402.2021.1912126]. *Australasian Journal of Philosophy*, *100*[3], 550–564.
Schroeder, T. [2020]. Desire. In E. N. Zalta [Ed.], *The Stanford encyclopedia of philosophy* [Summer 2020]. Metaphysics Research Lab, Stanford University.
Sharkey, L., et al. [2025]. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*.
Smith, L., Rajamanoharan, S., Conmy, A., CallumMcDougall, Lieberum, T., Kramár, J., Shah, R., Nanda, N., Rajamanoharan, S., Conmy, A., CallumMcDougall, Lieberum, T., Kramár, J., Shah, R., & Nanda, N. [2025]. Negative Results for SAEs On Downstream Tasks and Deprioritising SAE Research (GDM Mech Interp Team Progress Update #2).
Sui, P., Duede, E., Long, H., & So, R. J. [2026]. Critical confabulation: Can LLMs hallucinate for social good? *The Fourteenth International Conference on Learning Representations*.
Sui, P., Duede, E., Wu, S., & So, R. [2024]. Confabulation: The surprising value of large language model hallucinations. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14274–14284.
Sui, P., Duede, E., Wu, S., & So, R. J. [2024, June 25]. *Confabulation: The Surprising Value of Large Language Model Hallucinations*. arXiv: 2406.04175 [cs].
Todd, E., Li, M. L., Sharma, A. S., Mueller, A., Wallace, B. C., & Bau, D. [2024, February 25]. *Function Vectors in Large Language Models*. arXiv: 2310.15213 [cs].
Turpin, M., Michael, J., Perez, E., & Bowman, S. [2023]. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, *36*, 74952–74965.
Vaddiparthy, H. [2025, August 26]. *Self-Debugging AI: A Comprehensive Analysis of Claude 3 Opus's Code Generation and Error Resolution Capabilities*. 5408262.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. [2023]. Attention is all you need.
Weidinger, L., Raji, I. D., Wallach, H., Mitchell, M., Wang, A., Salaudeen, O., Bommasani, R., Ganguli, D., Koyejo, S., & Isaac, W. [2025, March]. Toward an Evaluation Science for Generative AI Systems [arXiv:2503.05336 [cs]].
Williams, I. [n.d.]. *Intention-like representations in language models?*
Williams, I., & Bayne, T. [2024]. Chatting with bots: Ai, speech acts, and the edge of assertion. *Inquiry*, 1–24.
Wu, Y., Mangla, R., Durrett, G., & Li, J. J. [2023]. Qudeval: The evaluation of questions under discussion discourse parsing. *arXiv preprint arXiv:2310.14520*.
Yuan, Y., & Søgaard, A. [2025]. REVISITING THE OTHELLO WORLD MODEL HYPOTH-.

Zhou, H., Westerdijk, H., & Islam, K. I. [2025]. Joint effects of argumentation theory, audio modality and data enrichment on llm-based fallacy classification.

Zhuang, T., & Lin, Z. [2024, November 18]. *The why, what, and how of AI-based coding in scientific research*. arXiv: 2410.02156 [cs].