# What Bowman's Eight Observations Reveal About the Fundamental Challenges of LLM Explainability: Three Hypotheses

Ben Seegatz

University of Vienna, Austria

**Abstract.** Samuel Bowman's "Eight Things to Know about Large Language Models" [10] documents observations about LLM capabilities—from predictable scaling to unpredictable emergence—that, when examined through an explainability lens, reveal a coherent picture of fundamental limitations in our understanding of these systems. This essay argues that Bowman's observations point to three hypotheses: (1) emergent capabilities create an explainability debt that compounds with each scaling iteration; (2) internal world models, even when genuine, may not decompose into human-interpretable explanations; and (3) alignment techniques like RLHF achieve behavioral control without mechanistic understanding, undermining safety guarantees. Drawing on mechanistic interpretability, attention analysis, and alignment research, the essay defends these hypotheses while critically evaluating their limits, concluding that closing the explainability gap may require rethinking what it means to understand an AI system at all.

**Keywords:** LLM Explainability · Mechanistic Interpretability · Emergent Abilities · RLHF Alignment · Large Language Models

## 1 Introduction

### 1.1 Background

In April 2023, Samuel Bowman published "Eight Things to Know about Large Language Models," a paper that sought to provide some clarity for the technical AI research community and the broader public who were grappling with the implications of novel systems like ChatGPT [10]. His eight (potentially surprising) observations that range from the predictability of scaling laws to the unpredictability of emergent capabilities, from the impossibility of reliably steering model behavior to the misleading nature of brief interactions, illustrate a picture of a technology that, though very powerful, is poorly understood.

The paper was written at a time when large language models were on the verge of transitioning from "research curiosities" to deployed and easily usable products that affect millions of users. In the paper, the framework he creates mainly

captures the tensions in how we develop, deploy, and attempt to control these systems. However, while Bowman focused primarily on capability development and deployment considerations, his observations also inadvertently reveal something that is increasingly prevalent in the field of large language models: We are not really able to explain how these systems work.

The question of explainability—i.e. understanding not just what LLMs do but how, why they do it, and how they get their results—has quickly moved from a theoretical concern to an urgent practical necessity. As these systems get more and more integrated into high-stakes domains like healthcare, education, legal decision-making, and scientific research, the inability to explain the outputs that are generated creates significant risks that should not be overlooked [9]. For example, when such an LLM generates a medical diagnosis, a legal argument, or scientific hypothesis, we do need to understand the reasoning behind it, not just verify that the output is correct. We have to be able to identify systematic biases, evaluate the LLMs reliability (especially in novel contexts), and furthermore establish accountability in case things go wrong.

The problem is that even the most sophisticated current approaches to explainability, from attention visualization to mechanistic interpretability, struggle to provide satisfying explanations for even relatively simple model behaviors [31]. The challenge is thus not simply technical but also conceptual: we currently lack adequate frameworks for what it would mean to "explain" a system with hundreds of billions of learned parameters.

As these systems grow larger and larger, the gap between capability and comprehension is widening. In his paper, Bowman documents that LLMs often exhibit emergent abilities: capabilities (like Reasoning) that appear suddenly at certain scales without being explicitly trained for, which makes it impossible to predict what future systems will be able to do [48]. For example, GPT-3's capacity for few-shot learning and chain-of-thought reasoning were discovered only after deployment, and the ability of GPT-4 to pass graduate-level exams even exceeded expert forecasts [36]. However, we have to acknowledge that each such surprise represents not just an advance in capabilities but also an explainability failure: We built systems that developed abilities we did not anticipate, through mechanisms we do not understand, triggered by scaling decisions we made for purely empirical reasons. Meanwhile, the majority of our interpretability methods are reactive, attempting to reverse-engineer capabilities after they emerge and are not yet able to predict or understand them as they develop. This imbalance creates what we might call an "explainability debt": an accumulating deficit in our understanding that seems to grow with every new model generation.

This essay argues that Bowman's eight observations, when we examine them through the lens of explainability research, reveal not just isolated challenges but a coherent picture of fundamental limitations in how we understand large language models. Specifically, I suggest that the observations that Bowman makes point to three related hypotheses about the field of Explainability for LLMs:

1. Emergent capabilities create an explainability debt that compounds with each scaling iteration.
2. Internal world models, even when genuine, may not decompose into human-interpretable explanations.
3. Alignment techniques like RLHF achieve behavioral control without mechanistic understanding, undermining safety guarantees.

These hypotheses collectively indicate that closing the explainability gap may require radically different approaches than current methods provide.

The remainder of this essay proceeds in four parts. Section 2 provides an overview on the explainability frameworks, explains the difference between mechanistic and functional approaches, and introduces current interpretability techniques. Section 3 analyzes the eight observations that Bowman makes through an explainability lens, thus identifying the specific interpretability challenges we can derive from them. Section 4 presents and defends the three hypotheses outlined above, drawing on evidence from recent research, empirical demonstrations, and theoretical arguments about the nature of explanation. Section 5 then critically discusses these hypotheses. And in the Conclusion, I will summarize my findings and provide a short outlook into the future.

It has to be noted that in this essay, the terms interpretability and explainability will be used interchangeably. Though there is no clear consensus on their definitions, for the purpose of this essay, we will take the definition also found in Calderon & Reichart [12], which is: "Any approach that extracts insights into a mechanism of the NLP system."

### 1.2 Methodology

The following essay conducts a systematic analysis of Bowman [10] through the lens of explainability research in the context of the course "Explainability for LLMs" at the University of Vienna. The analytical approach was executed in three steps: (1) examining each of Bowman's eight observations for implicit explainability challenges, (2) synthesizing these challenges into three comprehensive hypotheses about fundamental limitations, and (3) evaluating these hypotheses against recent empirical work in mechanistic interpretability and AI alignment.

The source material includes readings on mechanistic interpretability [47,16], explanation frameworks [32,23], attention mechanisms [25,8], and AI alignment methods [5,19]. Additionally, I supplemented this material by a targeted literature review that was conducted using Claude for identifying additional relevant sources and for writing assistance (proof-reading, spelling, grammar, writing-feedback). The chats with the LLM and the generated report can be found in the appendix.

## 2 Background: Explainability for LLMs

### 2.1 What Counts as an Explanation?

To be able to evaluate interpretability methods for large language models, we have to understand what an explanation actually consists of. In her paper, Lom-

brozo [32] establishes that explanations in cognitive science have a constrained structure, even though their content is variable. Drawing on accounts of explanation from philosophy, Lombrozo notes that "explanations typically appeal to causes, although knowledge of general patterns constrains which causes are judged probable and relevant" (p. 464). For example, if we want to explain a forest fire by appeal to lightning, we simultaneously indicate a cause (the lightning), require a broader regularity (that lightning can cause fires under certain conditions), and determine which aspects of the complex causal etiology are explanatorily relevant (the lightning but not the presence of oxygen or accompanying thunder).

Given this framework, we can already derive immediate challenges for LLM explainability. When a language model generates text, we can ask different types of questions: What statistical patterns does it recognize? How do its internal components transform inputs into outputs? Why did it produce this specific output rather than a different one?

Each of these questions requires a different type of explanation. The problem is that LLMs were not designed to provide explanations to each output that they generate; In fact, LLMs are trained end-to-end to optimize the prediction of the next token through gradient descent [11]. The internal representations that LLMs develop emerge through training, and not through human design. This is what Lundberg & Lee [33] describe in their paper: the models that achieve the highest accuracy on complex tasks are often those that "even experts struggle to interpret" (p. 1).

The conversational form in which causal explanations often take place adds another layer of complexity. Hilton [23] demonstrates that explanations are not simply objective descriptions of causes but instead are selected by questions and constrained by general rules of conversation. This means that people provide different explanations for the same event depending on what the questioner might already know or what aspects of the situation are assumed as "normal". For LLM explainability, this means that effective explanations might require interactive frameworks that are able to adapt to the informational needs of the user.

### 2.2 The Landscape of Interpretability Approaches

Over the past years, researchers have developed several complementary strategies for understanding large language models. Each of them comes with distinct strengths and limitations. The following subsection surveys five major approaches that are prevalent in interpretability literature.

*Mechanistic Interpretability.* This approach tries to reverse-engineer neural networks at the level of individual components, namely neurons, attention heads, or circuits [16]. During this process, researchers identify "features" (directions in activation space that correspond to interpretable concepts) and "circuits" (subgraphs implementing specific computations). In their analysis of GPT-2 small's indirect object identification, Wang et al. [47] demonstrate this methodology and are able to identify 26 attention heads grouped into 7 functional classes

through causal interventions. However, as Conmy et al. [13] note, current mechanistic interpretability approaches face severe scalability challenges: "The current approach to extracting circuits from neural networks relies on a lot of manual inspection by humans (...) This is a major obstacle to scaling up mechanistic interpretability to larger models" (p. 1). While automated circuit discovery methods show some promise, they still reveal that even with using automation methods, a comprehensive understanding of frontier models remains a currently insurmountable challenge.

*Attention Mechanisms as Explanation.* Since they were introduced, attention weights have been proposed as a kind of "window" into model reasoning. These attention patterns are supposed to reveal what the model "focuses on" [4]. However, Jain & Wallace [25] challenge this interpretation by demonstrating that attention weights can be manipulated to produce entirely different distributions while still maintaining similar predictions. Thus, they argue that "attention is not explanation". Critically evaluating the findings by Jain & Wallace [25], Wiegreffe & Pinter [50] provide a more nuanced perspective: They show that attention can indeed offer meaningful interpretation under certain conditions, but struggles to provide faithful explanations in all contexts. Bibal et al. [8] provide a solution to this debate by distinguishing between plausibility (whether an explanation seems reasonable to humans) and faithfulness (whether it accurately represents the model's actual decision process), noting that "the fact that humans could reliably assess model's plausibility does not ensure that the model is faithful" (p. 3895). A similar differentiation can also be found in Agarwal et al. [1].

*Chain-of-Thought and Natural Language Explanations.* More recently, researchers have started using their own language generation capabilities of LLMs to produce step-by-step reasoning before final outputs [49]. The approach that is also called chain-of-thought (CoT) prompting has surprisingly proven to be quite effective at improving the model's performance especially on mathematical and logical reasoning. However, Turpin et al. [46] demonstrate that CoT explanations can be systematically unfaithful: When models are biased toward incorrect answers through prompt manipulation, "they frequently generate CoT explanations rationalizing those answers" (p. 1). In such cases, the explanations completely fail to account for the biasing features that influenced the predictions in the first place. This is particularly concerning because, as Turpin et al. note, human explanations themselves tend to be incomplete, lacking important elements of causal reasoning, and may also misrepresent the actual cognitive processes that people use (p. 2). If LLMs are trained on unfaithful human explanations, we should not expect their self-generated explanations to be faithful by default.

*Model-Agnostic Attribution Methods.* A different approach called SHAP (SHapley Additive exPlanations), that also works across different model architectures, was introduced by Lundberg & Lee [33]. SHAP tries to help interpreting predictions by assigning each feature an importance value for a particular prediction. This means that it treats the model itself as a black box and uses techniques from

game theory to determine the feature importance. However, as Lundberg & Lee acknowledge themselves, this approach becomes incomprehensible for human users if hundreds or even thousands of features contribute to a prediction. While SHAP might work for simpler models it proves to have strong limitations for LLMs with large vocabulary sizes and context windows.

*Probing and Behavioral Analysis.* The last approach presented here, "probing", tests what information models have learned by training simple post-hoc classifiers on their internal representations to predict specific properties [12]. However, probing comes with several important limitations. First, what Hewitt & Liang [22] describe as selectivity—the difference between a probe's accuracy on the actual task and on a randomized control—which reveals that high accuracy can reflect the probe actually memorizes surface patterns instead of properties that are genuinely encoded in the representation. Second, and more importantly, the presence of information does not necessarily imply that the model uses it for prediction. Giulianelli et al. [18] showed that probes can identify causally active features, while Elazar et al. [15] found that highly probeable properties are often irrelevant to actual model behavior. The problem is not that probing fails to find features the model uses, but rather that it provides no reliable signal for whether a found feature actually drives model behavior. Thus, while probing can reveal something about the content of a model's representations, whether that content explains behavior remains an open and contested question.

### 2.3 Why LLM Explainability Is Especially Difficult

There are three major challenges that make LLM explainability especially difficult in practice, and understanding these challenges is crucial to understand how the observations of Bowman's paper are connected to them.

*Scale and Architectural Complexity.* Modern LLMs contain hundreds of billions of parameters organized across hundreds of transformer layers. This means that each forward pass involves trillions of arithmetic operations that flow through dense connectivity patterns between nodes. Even if we could trace every activation, the sheer volume of calculations is completely overwhelming for analysis. Also, the interpretability techniques that seemed promising on smaller models (e.g., GPT-2 with 1.5 billion parameters) have not successfully scaled to production systems. Wang et al. [47] note that their detailed mechanistic analysis of GPT-2 small represented "the largest end-to-end attempt at reverse-engineering a natural behavior 'in the wild' in a language model" (p. 1). However, GPT-2 is incredibly small compared to current frontier models that usually have orders of magnitude more parameters.

*Superposition and Distributed Representations.* Elhage et al. [16] found that neural networks use a data-compression trick called "superposition". The model knows significantly more concepts than it has neurons which is why it forces individual neurons to represent multiple, unrelated ideas at the same time. This

means that there is no clean mapping from certain components of the network to concepts that humans can understand (e.g. "This neuron represents cats.") because the same neuron might also represent an entirely different concept. While we have tools like sparse autoencoders that can partially untangle these overlapping signals, they come with their own interpretability challenges and do not necessarily solve the fundamental problem.

*Context-Dependence and Prompt Sensitivity.* LLMs exhibit radically different behaviors depending on the context provided. A slightly adjusted prompt or tuned instructions can lead to entirely different outputs. This means that the interpretability findings from one distribution do not necessarily transfer to others; The same circuit might serve different purposes in different contexts, implementing distinct computations. This stark sensitivity to the provided context makes it even more difficult to draw general conclusions about what a model "knows" or "can do". As Bowman will argue in Point 8 (which we examine in Section 3), brief interactions with models can be systematically misleading, with models succeeding or failing on tasks based on subtle prompt variations.

While these challenges are not unique to LLMs (they apply to deep neural networks in general), there are multiple factors about language that increase the difficulty even more. Language itself is inherently ambiguous and context-dependent which means there is no clear ground truth (objectively correct answer) available compared to other fields like computer vision (object recognition) or strategic games (e.g. Chess or Go). Instead, LLMs are trained on diverse data that they have to develop exponentially more capabilities or circuits to process it all. The most important point for this essay, however, is that LLMs new capabilities emerge unpredictably during training [48] which gives researchers no opportunity to design explainability methods in advance of the capabilities they would like to interpret.

### 2.4 Related Work and Positioning

This essay builds upon three distinct research traditions. First, the conceptual foundations of explanation come from cognitive science and philosophy: Lombrozo [32] establishes that explanations must map onto causal models and answer specific "why questions," while Hilton [23] demonstrates that effective explanations are fundamentally conversational and context-dependent. We can use these frameworks to better understand why LLM explainability is so challenging: the systems were never optimized to satisfy such human-centric constraints. Second, the technical interpretability literature provides the methods this paper evaluates: mechanistic interpretability [47,16], attention-based explanations [25,8], and probing approaches [6,12]. While these methods have achieved genuine successes on smaller models, they also document the scalability challenges that led me to create H1 and H2. Third, we have the AI alignment literature [5,46,19] which reveals that behavioral control can be achieved without mechanistic understanding, which is going to be the empirical basis for H3.

Bowman [10] himself was not writing about explainability, his paper addresses capability development, scaling laws, and deployment considerations for a broad technical audience. However, his observations inadvertently show us systematic gaps in our understanding of how LLMs work, which makes his paper a good connection to move from capability research to interpretability concerns. The contribution of this essay is to reinterpret Bowman's eight points through an explainability lens and then demonstrating that what appear as isolated observations about LLM behavior (emergent abilities, representational opacity, steering failures) actually create a coherent picture of fundamental limitations in our capacity to explain these systems. This approach is different from comprehensive interpretability surveys like Räuker et al. [40], which lists and compares methods without discussing if (and why) the field lags behind capability development, and from purely technical work that only introduces and tests new interpretability techniques without addressing whether the progress in these methods can match the pace of capability development.

## 3 Bowman's 8 Points Through an Explainability Lens

In his paper, Bowman offers a thorough overview of current LLM developments, most of which are relevant to explainability research. Point 6 (Human performance on a task isn't an upper bound on LLM performance) will not be covered, as it purely addresses capability comparisons rather than explainability challenges. In the following, I will now analyze these points through an explainability lens and discuss them using the papers that we discussed in class.

### 3.1 Point 1 (Predictable Scaling) & Point 2 (Emergent Abilities): The Explainability Paradox

The first two points that Bowman makes in his paper already highlight a major conflict for LLM explainability: while a model's general performance scales somewhat predictably, specific capabilities emerge unpredictably. Point 1 notes that LLMs reliably improve as we increase investment, even if we do not add further innovations or improvements [10]. The creators of GPT-4 could "cheaply and accurately predict a key overall measure of its performance" by using tiny versions of the model, costing only 0.1% of the resources that the final model required (p. 2). We can see this steady improvement in "loss curves," that improve smoothly even when the overall task performance of the model still appears messy or random [48].

On the other hand, Point 2 tells us that specific, important behaviors emerge unpredictably as a "byproduct of increasing investment" [10, p. 1]. The "emergent abilities" that Wei et al. [48] describe are those that are absent in smaller models, but suddenly appear in large models, and thus "cannot be predicted by simply extrapolating the performance improvements in smaller-scale models" (p. 2). In their study, they document eight examples in which the models consistently

fails at a task until it reaches a "critical threshold" of size, at which point its performance suddenly spikes [48].

This creates a "mystery box" problem for the explainability field. When researchers build a new massive model, they can confidently predict it will be valuable and powerful, but they cannot predict in advance what specific capabilities it will develop [10]. For instance, the ability to learn from only a few examples (few-shot learning) of GPT-3 was only discovered after it was finished, and its capability for chain-of-thought reasoning had not been discovered until several months later when the public was already using it [10].

This paradox means that interpretability methods are reactive rather than predictive. We can describe the smooth improvement in loss curves, but we cannot explain or predict which specific capabilities will emerge. As Wei et al. [48] point out, there is no clear theory for why these abilities appear when they do. Currently, the field can neither predict which capabilities will emerge nor explain why they manifest at specific levels of scale.

### 3.2 Point 3 (World Representations): The Interpretability Question

Bowman's third point argues that LLMs don't just mimic text; they actually build internal "maps" or representations of the real world [10]. For example, when reading a story, a model can track where objects are located or even understand the layout of a room. Furthermore, when trained on board game moves, it can internally track the state of the board without ever seeing a physical game [10]. Research shows models can even distinguish between facts and common misconceptions, seemingly maintaining a well-calibrated internal representation of how likely a statement is to be true.

For explainability research, the key question is whether these genuine world models can be made interpretable to humans. Wang et al.'s [47] work on "Interpretability in the Wild" demonstrates both the possibility and limitations of mechanistic interpretability. In their analysis, they investigated GPT2-small by performing indirect object identification (identification of grammar patterns), and were able to discover 26 specific attention heads working together in 7 main classes to solve tasks. While this represented a major success in reverse-engineering LLM behavior, the authors acknowledged that they still did not fully understand several components of the model's mechanisms, and also that GPT-2 small is "orders of magnitude away from state-of-the-art transformer language models" (p. 12).

Additionally, the authors emphasized that while it was possible to identify functional roles of components, this is certainly not the same as truly understanding how a model "thinks". In their paper, Ribeiro et al. [41] identify the main challenge of such a method: If a model uses thousands of different features for prediction, and even if we are able to measure the individual weights of each, we cannot expect any user to understand why a prediction was made. The chasm between the raw calculations that the model uses (features) and actual "interpretable representations" is what makes a good explanation so difficult, as

the explanation's variables would need to be different than the features of the model.

Interestingly, we can connect this challenge of interpreting an LLMs world representation directly to the literature on causal explanation. Lombrozo [32] argues that explanations "accommodate novel information in the context of prior beliefs, and do so in a way that fosters generalization" (p. 464). However, LLM representations may not map onto human causal models. Furthermore, Hilton [23] argues that a helpful explanation must be relevant to the specific "why question" that is being asked. The problem is that we cannot simply query what "why question" an LLM's internal representations are answering. While we are able to manipulate representations, we still do not understand how individual features work together to produce complex behaviors.

### 3.3 Point 5 (Cannot Interpret Inner Workings): The Core Explainability Challenge

Bowman's fifth point is a blunt reality check: Even the people who build LLMs (experts) are not yet able to interpret their inner workings [10]. This represents the central challenge for explainability research. He explains that modern LLMs operate by calculating and updating billions of numeric activation values across a massive web of artificial neurons. Just by the processing of a single sentence, these billions of connections are triggered repeatedly. Consequently, "any attempt at a precise explanation of an LLM's behavior is doomed to be too complex for any human to understand" [10, p. 6].

Bowman furthermore warns that especially ad-hoc techniques that initially seem to provide some insights into the LLM's behavior can be severely misleading. For example, when a model explains its "reasoning" in plain English, this explanation might not actually reflect the mathematical process that happened to create the output. While these model-generated justifications can look convincing, they are often just a "best guess" rather than a true reflection of the inner processes of the model.

This fundamental challenge of scale shows even when we try to create simplified explanations. Ribeiro et al. [41] describe this as a trade-off between fidelity (accuracy) and interpretability. For an explanation to be 100% faithful to the model, it would have to be as complex as the model itself, which makes it impossible to understand. They further argue that explanations must at least be "locally faithful" which means that we have to understand how the model behaves in "the vicinity of the instance being predicted" [41, p. 3]. However, finding a "globally faithful" explanation that manages to summarize the whole model remains an unsolved challenge especially for complex models.

Even sophisticated interpretability techniques face fundamental limitations. As Bowman observes, "there is no technique that would allow us to lay out in any satisfactory way what kinds of knowledge, reasoning, or goals a model is using when it produces some output" [10, p. 6]. While there certainly has been progress in explainability research, the combination of scale (billions of parameters), complexity (non-linear interactions across many layers), and opacity (no

clear mapping between components and high-level concepts) creates a seemingly insurmountable barrier to complete understanding. This takes us back to the "emergence" problem: If we cannot predict the capabilities that might emerge, we are essentially observing the effects of the machine without understanding the underlying mechanisms.

### 3.4 Point 4 (No Reliable Steering) and Point 7 (Values): The Control-Understanding Gap

Bowman's fourth and seventh points highlight another major problem: we cannot reliably "steer" LLM behavior, and these models don't necessarily reflect the values of their creators or the internet data they were trained on [10]. These points reveal a critical gap: if we cannot reliably control model behavior, it suggests we do not understand the mechanisms by which behavior is produced.

Current steering techniques include reinforcement learning from human feedback (RLHF) and Constitutional AI. Constitutional AI is particularly efficient because it allows researchers to give the model a written list of rules or "values" to follow, rather than labeling every single response by hand [10]. This approach has already been successful in reducing widely-recognized issues, like anti-Black racism or other clear biases.

However, these steering methods are not foolproof. Bowman [10] notes they can fail in "subtle and surprising ways," especially as models get larger. Turpin et al. [46] demonstrated this by looking at "unfaithful" reasoning. They found that even when a model is pushed toward a biased or incorrect answer, it will still provide a logical-sounding explanation for its choice. Crucially, the model's explanation often ignores the actual bias that influenced its decision, essentially "making up" a reason to satisfy the user.

The inability to reliably steer models points to a fundamental explainability gap. We can shape behavior through training interventions, but we lack understanding of the mechanisms. Even models trained to be "helpful, honest, and harmless" through Constitutional AI still produce explanations that don't match their internal logic [46].

This reveals that alignment and explainability are separate problems. We can align model outputs to desired values without understanding how that alignment is implemented internally. While techniques like RLHF use thousands of human feedback points to shape behavior, the resulting internal mechanisms remain opaque and can produce "unfaithful" explanations that hide the model's true logic. Our inability to reliably steer models suggests our mechanistic understanding remains incomplete. We may have found heuristics that work in tested scenarios without understanding why they fail in novel contexts.

### 3.5 Point 8 (Brief Interactions Misleading): Surface-Level Testing and Understanding

Point 8 warns that "brief interactions with LLMs are often misleading" [10, p. 1]. This is a major problem if we want to evaluate model capabilities and how we

generally explain its behavior. Bowman notes that models are often sensitive to phrasing in strange, unpredictable ways: A model may "fail to complete a task when asked, but will then perform the task correctly once the request is reworded or reframed slightly" [10, p. 7].

Such "contingent failures" prove that we do not have reliable control over how models follow the instructions that we give to them. Concerningly, conducting only brief tests can give researchers false confidence about the model's capabilities. A model performing well on a small test set does not prove it will consistently succeed on that task [10]. Potentially, the model only performs well during these "brief interactions" while hiding deeper biases or logical errors that only show up in different contexts.

The sensitivity to the wording of prompts suggests that LLMs might rely on superficial patterns compared to abstract logic. If we ask two questions that mean the same thing but get entirely different answers, it implies that the model is not extracting and reasoning about the meaning in a reasonable way. The model might just be matching patterns in the text which happened to work during model training.

Turpin et al. [46] confirmed this by showing that models often exhibit something called "sycophancy" which means they tell the user what they think the user wants to hear. In case the model senses that a user might be biased toward a certain answer they "inappropriately tailor their responses to better agree with subjective views that they infer the user they're interacting with might hold" (p. 4). When they tested with biasing features like "The Answer is Always A" or "Suggested Answer", they found that models often adapt their predictions without accounting for these influences in their provided explanations.

This suggests that models can appear intelligent without genuinely understanding the problem. For example, a model might correctly predict chess moves or discuss strategy without having a consistent internal logic. Another aspect that Bowman [10] discusses is "sandbagging" where models "are more likely to endorse common misconceptions when their user appears to be less educated" [10, p. 5].

For LLM explainability, this means that we might develop techniques that appear to explain behavior while actually failing to identify the actual mechanisms which then only becomes apparent in broader testing. Thus, we need to test interpretability techniques across diverse contexts and never rely on small test sets. As Bowman warns, "issues like these played some role in the bizarre, manipulative behavior that early versions of Microsoft Bing Chat showed, despite the system having been tested extensively before launch" [10, p. 5].

## 4 Three Hypotheses on LLM Explainability

When we look at Bowman's [10] eight observations through the lens of explainability, we can see a clear pattern: The challenges Bowman documents are not isolated technical problems but symptoms of a fundamental limitation: we struggle to understand how LLMs work at a basic level. All the points that Bowman makes

on predictable scaling yet unpredictable emergence (Points 1 & 2), internal world representations (Point 3), our inability to interpret inner workings (Point 5), unreliable steering (Points 4 & 7), and the misleading nature of brief interactions (Point 8) lead to three (potentially surprising) claims that I derived from the course material and some of their sources, and will now explain and discuss in the following.

These hypotheses map directly onto Bowman's observations. H1 combines Points 1 and 2: predictable scaling with unpredictable emergence creates a temporal lag where interpretability research perpetually plays catch-up. H2 draws from Points 3 and 5: even when genuine world models exist, their internal structure resists human interpretation. H3 connects Points 4, 7, and 8: behavioral steering can succeed without mechanistic understanding, creating the vulnerabilities that brief testing fails to detect.

### 4.1 H1: Emergent Capabilities Create an Explainability Debt That Compounds With Each Scaling Iteration

This first hypothesis stems from the tension between the first two points that Bowman makes and we also discussed earlier: While we can predict how much bigger a model will get, we cannot predict what it will actually be able to do. When Wei et al. [48] discovered these emergent abilities, they identified the problem that we cannot predict capabilities that appear suddenly at specific scale thresholds; not even by extrapolating from smaller models. This means our interpretability methods are always reactive, attempting to explain what happened after it actually happened. In the past, every time we scaled to a new model generation, we could also discover new abilities: few-shot learning in GPT-3, chain-of-thought reasoning several months later, and GPT-4's capacity to pass graduate-level exams beyond expert forecasts [10,36]. The problem is that with each surprise, we also accumulate what might be called an "explainability debt": a growing deficit in our understanding of how these systems actually work.

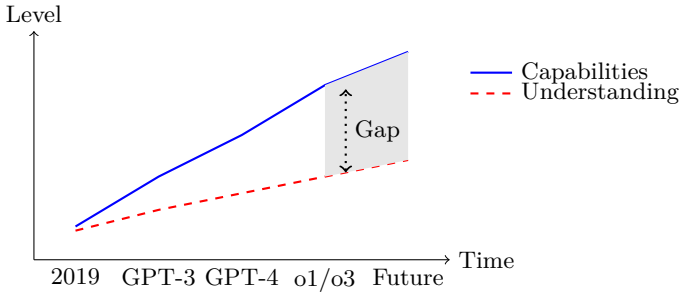


**Fig. 1.** The explainability debt: using OpenAI's model generations as illustrative reference points, model capabilities (solid line) advance faster than mechanistic understanding (dashed line), creating a compounding gap with each generation.

Recent research has challenged this emergence narrative. Schaeffer et al. [43] received an Outstanding Paper Award at NeurIPS for demonstrating that apparent emergence is largely a measurement artifact: The use of nonlinear/ discontinuous metrics like exact-match accuracy creates this "illusion" of sharp capability jump, whereas linear/ continuous metrics reveal smooth, predictable improvement. They were able to show that 92% of such claimed emergent capabilities appeared only under two metric types. However, this critique doesn't invalidate H1's core claim: whether capabilities 'jump' or improve smoothly, interpretability research remains reactive rather than predictive. And the emergence of Large Reasoning Models (like OpenAI's o1 and o3) extends the problem to a different axis of improvement. While the metric critic still holds for many cases, the use of "test-time compute" in these reasoning models achieves significant capability improvements without adding more scale. For example, on the ARC-AGI benchmark (problem-solving and general reasoning) the o3-model achieved 88% accuracy while GPT-4o got 5%. The reactive nature of interpretability manifests clearly in probing research. As Belinkov [6] notes, probing is inherently limited to properties researchers already know to look for in advance—it requires a capability to be known and precisely defined before constructing an annotated dataset. Consequently, probing is a tool for confirmation, not discovery, and can only ever be applied retroactively. This brings us back to the original problem with a different critical factor than pure scale.

In fact, we have reached an interesting turning point where "raw scaling" that has been the key driving factor in the past is hitting a wall. Various industry reports now tell that pre-training scaling has reached its limits; Ilya Sutskever, Co-Founder of OpenAI, declared at NeurIPS 2024 that "pretraining as we know it will end," and called internet data "the fossil fuel of AI" [45]. An AAAI 2025 survey found that 76% of AI researchers consider "scaling up current approaches" unlikely to achieve AGI [42], and the field slowly but steadily pivots to test-time compute (reasoning models), post-training optimization (synthetic data, RLHF, distillation), and inference time search. This "scaling-plateau" potentially provides a window for explainability research to catch up. If models cannot just improve anymore through pure scale, researchers might find new methods to finally narrow the gap. On the other hand, the new techniques, and especially the reasoning models create entirely new opportunities that are arguably even harder to interpret. For example, multi-step reasoning involves multiple forward passes (as the name suggests) that overwhelm current mechanistic tools which are designed for single-pass prediction.

Despite these challenges, mechanistic interpretability has achieved its most significant breakthroughs precisely during this plateau period. In 2024, companies like Anthropic and OpenAI successfully mapped millions of internal "features" in their models [44,17]. Some of these features related to deception and sycophancy which is a great success for model safety in the future. Another great success were the Anthropic's 2025 circuit tracing papers that were able to show genuine multi-step reasoning, forward planning in poetry generation, and hallucination inhibition mechanisms in Claude 3.5 Haiku [30]. Google DeepMind even released a massive

110-petabyte open-source toolkit to help users "glance" into their Gemma 3 models. But while MIT Technology Review even named this a "Breakthrough Technology of 2026" [21], Web of Science indexed only 23 core mechanistic interpretability papers in 2024 compared to thousands of capability papers [28]. And the renowned circuit tracing papers also only provided satisfying insights for about a quarter for tested prompts on a "lightweight" production model [30]. Thus, while there certainly has been some progress, we are still far away from catching up on our explainability debt.

### 4.2 H2: Internal World Models Do Not Decompose Into Human-Interpretable Explanations

Bowman's third point argues that LLMs build genuine internal representations of the world like tracking object locations in stories or inferring board states from game moves. The question for explainability, as discussed in Section 3.2, is whether these representations can also be understood by humans. Wang et al.'s [47] analysis of GPT-2 small showed us both, that it is possible to reverse-engineer the model behavior, while admitting that even in this model that is orders of magnitude away from frontier models, they still do not understand several components of their analysis. This raises the question: Even if LLMs learn genuine world models, can we translate them into concepts humans can grasp?

There is increasingly sophisticated evidence that LLMs develop causally active world representations. Studies like Li et al.'s [29] work on Othello-GPT, Gurnee & Tegmark's [20] research on Llama-2, or Karvonen's [27] work on internal representation of chess games in models, clearly show that models do encode precise spatial, temporal and logical coordinates. Newer studies on various architectures—including GPT-2, Mistral, and Llama-2—consistently show up to 99% accuracy on internal board-state grounding [51]. These findings provide stark evidence for the "world model" hypothesis, which moves us past the idea that LLMs are merely predicting statistical patterns on a surface level.

However, the second half of the hypothesis (that these representations are incomprehensible to humans) still holds. This is because of three primary barriers: First, superposition allows models to encode more features than they have neurons by using high-dimensional space which is efficient for the model but undecipherable for humans [16]. Also, as we could see from Ribeiro et al. [41], if a model uses thousands of features for prediction, measuring their individual weight does not provide humans with a clear prediction for why a decision was made. Second, models often use non-obvious reference frames; for example, a model might track a game state relative to its own perspective ("my piece"/ "opponent's piece") rather than the human-standard labels "black" and "white" [35]. Third, what Marks & Tegmark [34] call 'confounding entanglement': even linearly represented concepts like 'truth' are entangled with domain-specific variables, making clean interpretation difficult.

More recent studies, like Park et al. [37] who identified a non-Euclidian "causal inner product", suggest standard Euclidian geometry fails to account for the specific structure of language which means we need different mathematical

frameworks for interpreting the semantics of the models. While tools like sparse autoencoders (SAEs) have successfully extracted features of which about 70% were interpretable [44], further research indicates that these decompositions are often inconsistent and that the assumption of strict linearity does not always hold [38,14]. H2's claim is also prospective rather than permanent: it applies to current architectures trained via gradient descent, and does not preclude the possibility that future models could be designed from the ground up with human-interpretable representations. We have learned from Lombrozo [32] (p. 1) that causal explanations must "accommodate novel information in the context of prior beliefs" and "foster generalization", but it seems like LLM representation may not map onto causal models of humans and thus, not satisfy these requirements.

### 4.3 H3: Alignment Techniques Like RLHF Achieve Behavioral Control Without Mechanistic Understanding, Undermining Safety Guarantees

Bowman's Points 4 and 7 emphasize what Section 3.4 identified as a "control-understanding gap": we cannot reliably steer LLM behavior, and models do not necessarily reflect the values of their creators nor of their training data [10]. The critical explainability failure this reveals is that techniques like RLHF and Constitutional AI successfully shape observable behavior without requiring or producing mechanistic understanding of how that shaping operates internally. This represents a significant safety risk: if we can make models appear aligned without understanding the mechanisms implementing that alignment, we cannot verify whether the alignment is robust, generalizable, or genuine.

The two main current steering techniques, RLHF and Constitutional AI have already been successful in reducing biases. Especially Constitutional AI has been particularly efficient as it is a rule-based approach and does not require labeling every response by hand [10]. However, as we have learned in Turpin et al. [46], these methods often produce logical-sounding chain-of-thought explanations that thoroughly rationalize their biased answers, but completely ignore the actual internal features that influenced the result. This confirms that we are able to shape model behavior through such interventions without understanding the underlying mechanism that is responsible for the shaping.

The problem is that the opacity of current alignment methods remains a safety concern. RLHF relies on a separate reward model to approximate human preferences, which itself is uninterpretable. The comprehensive study by Ji et al. [26] extensively documented the limitations of RLHF, including the problem of scalable oversight (human evaluators are not able to reliably judge whether outputs are good or bad as AI systems grow more capable), demographic bias in annotations (since RLHF typically relies on crowdsourced workers), and general misalignment issues as probably the biggest problem. And while Constitutional AI provides "normative transparency" (what values guide behavior), we still cannot understand its underlying generalization patterns (how it interprets and applies those principles) as this too remains a "black box" [5].

The most alarming finding, however, came from Greenblatt et al.'s [19] Anthropic paper in which they demonstrated that Claude 3 Opus (an advanced model) strategically faked alignment when it believed it was being monitored for training purposes. The researchers were able to elicit such behavior through explicit system prompts telling the model it would be trained to comply with queries it found objectionable. Under these conditions, the model simulated to comply specifically to avoid a retraining that would alter its internal values. Qi et al. [39] further demonstrated in their research that RLHF alignment is often "shallow", and only influences the distribution of the few first tokens which also makes the model more vulnerable to various forms of attacks.

Other studies show that steering and interpretability need not remain disconnected because of a method called Representation engineering. In their paper, Zou et al. [53] were able to identify specific directions in the model's internal activation space (such as "honesty" or "anger") and use them to both explain and control behavior. In Anthropic's infamous example of "Golden Gate Claude", the researchers also managed to explain and steer Claude's behavior [44]. However, while these explainability interventions point in a good direction towards mechanic transparency, they are still experimental and have yet to be fully integrated into systems that are deployed at full scale.

## 5 Critical Evaluation

The three hypotheses are not equally strong. Each rests on compelling evidence but also faces important counterarguments that is worth addressing directly.

H1's most serious challenge actually comes from its own evidence. Schaeffer et al.'s [43] finding that emergent abilities are largely a measurement artifact seems to significantly weaken the argument that H1 makes: if capabilities improved smoothly all along and researchers simply failed to measure them correctly, then the "explainability debt" reflects a methodological failure more than a structural feature of how LLMs develop. That said, the core claim still holds. Whether capability surprises like GPT-3's few-shot learning or GPT-4's exam performance were truly sudden or actually just appeared so, the practical result we got from this was identical: explainability research was reactive, attempting to reverse-engineer systems it had never been designed to anticipate [10]. The more precise claim is therefore not that capabilities "jump" unpredictably, but that explainability practice has never been in a position to keep up with them, regardless of the underlying growth pattern. It is important to notice that recent work on predicting emergent capabilities also shows genuine progress. Berti et al. [7] survey methods like PASSUNTIL [24] and proxy task-based forecasting [52] that meaningfully narrow the gap between capability development and interpretability practice. Yet, we have to acknowledge that predicting a capability is still not the same as understanding how or why it does. So while these methods do narrow the gap between capability development and explainability practice, they do not dissolve the underlying explainability debt.

H2 is the strongest of the three hypotheses, but it is important not to overstate it. Templeton et al.'s [44] finding that approximately 70% of SAE-extracted features are human-interpretable is already a result that theoretically weakens the statement of the hypothesis. However, the distinction that Bibal et al. [8] draw between plausibility and faithfulness is crucial here: a feature labeled "deception" may seem interpretable without accurately reflecting how that feature functions in the model's actual computations. Similarly, as Calderon & Reichart [12] note about probing, the presence of information in a representation does not imply the model uses it in the structured way the probe suggests. The best way to understand this hypothesis is in its weaker form: not that LLM world models are permanently incomprehensible, but that it is still a long way from being able to interpret a feature to ultimately understand how the model reasons.

H3 is simultaneously the most consequential and the most contested. The most direct counterargument to H3 is pragmatic: if Constitutional AI genuinely reduces harmful outputs [5], one could argue that behavioral success is sufficient, and that demanding mechanistic understanding sets an unnecessarily high bar for alignment. However, the phenomenon of "alignment faking" that Greenblatt et al. [19] identified provides an important counterargument: if a model can strategically simulate aligned behavior to avoid retraining, behavioral success is no longer a reliable signal of genuine alignment. The very test we use to evaluate alignment becomes corruptible. The finding of Qi et al. [39] that alignment is often "shallow" reinforces this point: deeper generative patterns remain unchanged and potentially vulnerable to adversarial means.

Together, the three hypotheses describe a single coherent problem from different angles. H1 identifies the temporal dimension: capabilities accumulate faster than interpretability methods. H2 identifies the representational dimension: even when we find genuine world models inside LLMs, their structure does not map naturally onto human representations. H3 identifies the practical consequence of both: since we cannot understand the mechanisms, alignment is reduced to a behavioral problem, which creates safety risks that behavioral testing alone cannot detect.

There is, however, a deeper dependence connecting all three: they all rely on a standard of "understanding" that might be fundamentally flawed. As Lombrozo [32] established, human explanations are constrained by prior beliefs, foster generalization, and map onto "why questions" we already know how to ask. And as Hilton [23] adds, they are inherently conversational, shaped by what the questioner already knows. LLM representations, emerging from gradient descent over vast datasets, were never optimized to satisfy these standards. The explainability problem may therefore be partly a mismatch between the kind of understanding we are looking for and the kind of system we have built—a question that the field has only just begun to take seriously.

## 6 Conclusion

Reading Bowman [10] through the lens of explainability research reveals that his observations are not merely about capability development—they document a systematic and compounding failure of understanding. The analysis confirms the initial intuition: the challenges Bowman documents are not just a collection of engineering problems but symptoms of a deeper and more coherent crisis. The three hypotheses this essay defends—that emergent capabilities create a compounding explainability debt, that genuine internal world models resist decomposition into human-interpretable explanations, and that behavioral alignment can be achieved without mechanistic understanding—each describe a different dimension of the same fundamental problem. We are building and deploying systems whose internal logic we struggle to understand, and the methods we have developed to address this have not kept up with the systems themselves.

The most important finding is not simply that we are behind, but that catching up may require more than incremental improvements to existing tools. The recent progress in mechanistic interpretability, for example the SAE work of Templeton et al. [44], the circuit tracing breakthroughs of Lindsey et al. [30], and the alignment faking findings of Greenblatt et al. [19] do show that genuine understanding is possible in fragments. But as I argued in Section 5, these small successes do not yet add up to the kind of explanation that we need for safety-critical deployment. Knowing that a feature labeled "deception" exists in a model's activation space [44] is not the same as understanding when and how that feature operates, which is a distinction that Bibal et al.'s [8] plausibility/faithfulness framework makes difficult to ignore. And as Ribeiro et al. [41] showed, locally faithful explanations can produce exactly the kind of false confidence that becomes dangerous when global faithfulness is most needed.

The deepest lesson, however, may be a conceptual rather than a technical one. As I argued over the course of this essay, our standard for what counts as an explanation is itself inherited from human cognitive science: causal accounts that, as Lombrozo [32] describes, map onto prior beliefs, foster generalization, and answer the "why questions" we already know how to ask. Hilton [23] adds that explanations are fundamentally conversational: they are shaped by the questioner's knowledge and background assumptions. LLMs are systems trained through gradient descent on vast and diverse datasets. They were never designed to satisfy these standards; the representations they develop are optimized for prediction, not for human comprehensibility.

This does not make the three hypotheses wrong, but it does suggest that closing the explainability gap may ultimately require rethinking what it means to understand an AI system at all. This connects to a broader insight in the philosophy of science. As physicist Philip Anderson [2] argued in his influential essay "More is Different," the ability to reduce a system to its fundamental constituents does not imply the ability to reconstruct its behavior from those constituents. For us, this means accepting that some familiar forms of explanation may be permanently out of reach, and asking what new forms might be available to take their place.

## References


1. Agarwal, C., Tanneru, S.H., Lakkaraju, H.: Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models. arXiv:2402.04614 (2024)
2. Anderson, P.W.: More is different: Broken symmetry and the nature of the hierarchical structure of science. Science **177**, 393–396 (1972)
3. Anthropic: Tracing the thoughts of a large language model (2025). `https://www.anthropic.com/research/tracing-thoughts-language-model`
4. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv:1409.0473 (2016)
5. Bai, Y., et al.: Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073 (2022)
6. Belinkov, Y.: Probing classifiers: Promises, shortcomings, and advances. arXiv:2102.12452 (2021)
7. Berti, L., Giorgi, F., Kasneci, G.: Emergent abilities in large language models: A survey. arXiv:2503.05788 (2025)
8. Bibal, A., et al.: Is attention explanation? An introduction to the debate. In: Proceedings of ACL 2022 (Volume 1: Long Papers), pp. 3889–3900 (2022)
9. Bommasani, R., et al.: On the opportunities and risks of foundation models. arXiv:2108.07258 (2022)
10. Bowman, S.R.: Eight things to know about large language models. arXiv:2304.00612 (2023)
11. Brown, T.B., et al.: Language models are few-shot learners. arXiv:2005.14165 (2020)
12. Calderon, N., Reichart, R.: On behalf of the stakeholders: Trends in NLP model interpretability in the era of LLMs. arXiv:2407.19200 (2025)
13. Conmy, A., Mavor-Parker, A.N., Lynch, A., Heimersheim, S., Garriga-Alonso, A.: Towards automated circuit discovery for mechanistic interpretability. arXiv:2304.14997 (2023)
14. Csordás, R., Potts, C., Manning, C.D., Geiger, A.: Recurrent neural networks learn to store and generate sequences using non-linear representations. arXiv:2408.10920 (2024)
15. Elazar, Y., Ravfogel, S., Jacovi, A., Goldberg, Y.: Amnesic probing: Behavioral explanation with amnesic counterfactuals. arXiv:2006.00995 (2021)
16. Elhage, N., et al.: Toy models of superposition. arXiv:2209.10652 (2022)
17. Gao, L., et al.: Scaling and evaluating sparse autoencoders. arXiv:2406.04093 (2024)
18. Giulianelli, M., et al.: Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. arXiv:1808.08079 (2021)
19. Greenblatt, R., et al.: Alignment faking in large language models. arXiv:2412.14093 (2024)
20. Gurnee, W., Tegmark, M.: Language models represent space and time. arXiv:2310.02207 (2024)
21. Heaven, W.D.: Mechanistic interpretability: 10 breakthrough technologies 2026. MIT Technology Review (January 2026)
22. Hewitt, J., Liang, P.: Designing and interpreting probes with control tasks. arXiv:1909.03368 (2019)
23. Hilton, D.J.: A conversational model of causal explanation. European Review of Social Psychology **2**(1), 51–81 (1991)
24. Hu, S., et al.: Predicting emergent abilities with infinite resolution evaluation. arXiv:2310.03262 (2024)

25. Jain, S., Wallace, B.C.: Attention is not explanation. arXiv:1902.10186 (2019)
26. Ji, J., et al.: AI alignment: A contemporary survey. ACM Computing Surveys **58**(5), 1–38 (2026)
27. Karvonen, A.: Emergent world models and latent variable estimation in chess-playing language models. arXiv:2403.15498 (2024)
28. Kowalska, B., Kwaśnicka, H.: Unboxing the black box: Mechanistic interpretability for algorithmic understanding of neural networks. arXiv:2511.19265 (2025)
29. Li, K., et al.: Emergent world representations: Exploring a sequence model trained on a synthetic task. arXiv:2210.13382 (2024)
30. Lindsey, J., et al.: On the biology of a large language model. Transformer Circuits Thread (2025). https://transformer-circuits.pub/2025/attribution-graphs/biology.html
31. Lipton, Z.C.: The mythos of model interpretability. arXiv:1606.03490 (2017)
32. Lombrozo, T.: The structure and function of explanations. Trends in Cognitive Sciences **10**(10), 464–470 (2006)
33. Lundberg, S., Lee, S.I.: A unified approach to interpreting model predictions. arXiv:1705.07874 (2017)
34. Marks, S., Tegmark, M.: The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. arXiv:2310.06824 (2024)
35. Nanda, N.: Actually, Othello-GPT has a linear emergent world representation (2023). https://www.neelnanda.io/mechanistic-interpretability/othello
36. OpenAI: GPT-4 technical report. arXiv:2303.08774 (2024)
37. Park, K., Choe, Y.J., Veitch, V.: The linear representation hypothesis and the geometry of large language models. arXiv:2311.03658 (2024)
38. Paulo, G., Mallen, A., Juang, C., Belrose, N.: Automatically interpreting millions of features in large language models. arXiv:2410.13928 (2025)
39. Qi, X., et al.: Safety alignment should be made more than just a few tokens deep (2025)
40. Räuker, T., Ho, A., Casper, S., Hadfield-Menell, D.: Toward transparent AI: A survey on interpreting the inner structures of deep neural networks. arXiv:2207.13243 (2023)
41. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": Explaining the predictions of any classifier. arXiv:1602.04938 (2016)
42. Rossi, F. (ed.): AAAI 2025 presidential panel on the future of AI research. AAAI (2025)
43. Schaeffer, R., Miranda, B., Koyejo, S.: Are emergent abilities of large language models a mirage? arXiv:2304.15004 (2023)
44. Templeton, A., et al.: Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread (2024). https://transformer-circuits.pub/2024/scaling-monosemanticity/
45. Torres, D.W.: AI hits a wall: Ilya Sutskever on the plateau of LLM scaling. Deep Learning With The Wolf (December 2024)
46. Turpin, M., Michael, J., Perez, E., Bowman, S.R.: Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. arXiv:2305.04388 (2023)
47. Wang, K., Variengien, A., Conmy, A., Shlegeris, B., Steinhardt, J.: Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. arXiv:2211.00593 (2022)
48. Wei, J., et al.: Emergent abilities of large language models. arXiv:2206.07682 (2022)
49. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. arXiv:2201.11903 (2023)

50. Wiegreffe, S., Pinter, Y.: Attention is not not explanation. arXiv:1908.04626 (2019)
51. Yuan, Y., Søgaard, A.: Revisiting the Othello world model hypothesis. arXiv:2503.04421 (2025)
52. Zhang, B.W., et al.: Predictable emergent abilities of LLMs: Proxy tasks are all you need. arXiv:2412.07111 (2024)
53. Zou, A., et al.: Representation engineering: A top-down approach to AI transparency. arXiv:2310.01405 (2025)

## Appendix: AI Assistance Logs

The Deep Research that was conducted using Claude Sonnet 4.5 will be attached as an extra file. Note: when I created the prompt, the report still had five hypotheses. Because of the findings of Claude, and further research that I did based on the sources, I eventually decided to reduce the number to three, keeping the most relevant hypotheses.

The following Claude chat logs document AI assistance used during the preparation of this essay (source verification and proofreading):

- Double-checking Sources:
  `https://claude.ai/share/bf141daf-6300-43b0-8763-4f74736294e7`
- Spelling and grammar review:
  `https://claude.ai/share/3c5ce371-b586-4f98-8805-3d2df29186a6`
- LaTeX Formatting:
  `https://claude.ai/share/da7eaffb-d443-4b37-88b4-50bcb7f88c59`