Opening the Black Box of Large Language Models: Two Views on Holistic Interpretability

Anonymous ACL submission

Abstract

As large language models (LLMs) grow more powerful, concerns around potential harms like toxicity, unfairness, and hallucination threaten user trust. Ensuring beneficial alignment of LLMs with human values through model alignment is thus critical yet challenging, requiring a deeper understanding of LLM behaviors and mechanisms. We propose opening the black box of LLMs through a framework of holistic interpretability encompassing complementary bottom-up and top-down perspectives. The bottom-up view, enabled by *mechanistic interpretability*, focuses on component functionalities and training dynamics. The top-down view utilizes *representation engineering* to analyze behaviors through hidden representations. In this paper, we review the landscape around mechanistic interpretability and representation engineering, summarizing approaches, discussing limitations and applications, and outlining future challenges in using these techniques to achieve ethical, honest, and reliable reasoning aligned with human values.

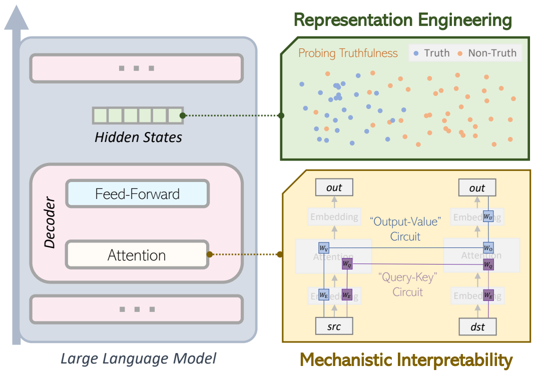

Figure 1: Two views on holistic interpretability: (i) Bottom-up view of mechanistic interpretability and (ii) Top-down view of representation engineering.

1 Introduction

Large language models (LLMs) such as GPT-4 (OpenAI, 2023), LLaMA-2 (Touvron et al., 2023), Claude (AnthropicAI, 2023), and Gemini (Team et al., 2023) have led to tremendous advances in language understanding and generation. However, as LLMs grow more powerful, issues around potential toxicity, unfairness, dishonesty, and hallucination threaten to undermine user trust. There is thus an urgent need to ensure the safe and beneficial alignment of LLMs with human values through *model alignment*. Model alignment aims to address these issues in order to build user trust and ensure LLMs safely generate helpful, honest, and unbiased text.

To address these alignment issues, we need a deeper understanding of the reasoning abilities and inner mechanisms of LLMs. More specifically, *eXplainable Artificial Intelligence* (XAI) techniques can be leveraged to explain these complex models. The research community has developed a wide range of tools to provide explanations from both local and global perspectives (Zhao et al., 2023). For example, various feature attribution methods have been proposed to offer insight into how different input tokens contribute to model predictions. However, these approaches have limited competence in fully understanding the functions and behaviors of LLMs. As such, they have limited capacity to address alignment issues.

Facing these challenges, we propose opening the black box of LLMs through *holistic* interpretability as a promising direction to overcome limitations of conventional approaches. As shown in Figure 1, holistic interpretability encompasses two complementary perspectives: *bottom-up* and *top-down* views. The bottom-up view through *mechanistic interpretability* explains models by focusing on the functionality of each component, as well as the relationships between models' abilities and training dynamics. This view interprets functional components through the concept of circuit, which offers insights into the inner workings of LLMs. The top-down view through *representation engineering* explains specific model behaviors such as dishonesty by analyzing hidden representations (Zou et al., 2023). This view enables understanding of how information is encoded internally in LLMs.

In this paper, we provide a systematic overview of the landscape around mechanistic interpretability and representation engineering. We summarize key techniques, discuss limitations, highlight applications, and outline open challenges. Our focus lies in how these approaches can enable better alignment of LLMs, so as to achieve ethical, honest, and reliable predictions for social good.

2 Mechanistic Interpretability

Mechanistic interpretability refers to the process of zooming into neural networks to understand the underlying components and mechanisms that drive their behaviors, also known as reverse engineering (Olah et al., 2020a). Just as the microscope revealed the world of cells, looking inside neural networks provides a glimpse into rich inner structures of models. This approach diverges from conventional interpretability methods that aim to explain the overall behaviors through features, neural activations, data instances and etc. Instead, it draws inspirations from other fields, such as neuroscience and biology, to investigate individual neurons and their connections. By tracking each neuron and weight, an intricate picture emerges on how neural networks operate through interconnected "circuits" that implement meaningful algorithms. On this delicate scale, neural networks become approachable systems rather than black boxes. Neurons play an understandable role and their circuits of connections implement factual relationships about the world. We can thus observe the step-by-step construction of high-level concepts, such as circle detectors, animal faces, cars, and logical operations (Olah et al., 2020a). In essence, zooming into the micro-level mechanics of LLMs enables deeper comprehension of their macro-level behaviors. Such mechanistic perspective represents a paradigm shift in interpretability towards unpacking the causal factors that drive model outputs.

2.1 Role in the General XAI Field

Mechanistic interpretability in XAI represents a paradigm shift towards a deeper and more fundamental understanding of deep neural network (DNN) models (Zhao et al., 2023).

- ***Global* versus *Local* Interpretation:** Mechanistic interpretability diverges from the traditional local focus of XAI, which concentrates on explaining specific predictions made by deep learning models, e.g., feature attribution techniques. Instead, it adopts a global approach, aiming to comprehend DNN models as a whole through the lens of high-level concepts and circuits.
- ***Post-hoc* Analysis versus *Intrinsic* Design:** Mechanistic interpretability aims to decipher the complexities inherent in pre-trained DNN models in a post-hoc way. This contrasts with efforts to create models that are mechanistically interpretable by design (Friedman et al., 2023).
- ***Model-Specific* versus *Model-Agnostic*:** Unlike some XAI methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), which are model-agnostic, mechanistic interpretability is a model-specific explanation. It requires tailor-made designs for each distinct LLM, analyzing their unique characteristics.
- ***White-box* versus *Black-box*:** Mechanistic interpretability aligns with white-box analysis, requiring direct access to a model's internal parameters and activations. This is in contrast to black-box XAI tools such as LIME and SHAP, which operate solely based on the model's inputs and outputs.

In summary, mechanistic interpretability in XAI is a critical approach to gain a profound understanding of DNN models. It emphasizes a **global** and **post-hoc** perspective, focusing on **model-specific, white-box** analysis to decipher the inner workings and intrinsic logic of complex AI systems. This approach is pivotal to advance transparency and build trust for LLMs, especially in high-stake scenarios where grasping "why" behind AI systems is as crucial as the decisions themselves.

2.2 Why Mechanistic Interpretability?

The question naturally arises: *Why has XAI research on LLMs moved towards the more specialized domain in mechanistic interpretability*? Exploring this shift can shed light on the evolving needs and challenges in this field. In this section, we attempt to delve into several factors that we believe have play a major role in steering the shift.

Alignment Requirement. In the age of LLMs, the standards for model performance have become more rigorous, not just in terms of accuracy but also in addressing crucial social concerns like honesty and fairness. Under this circumstance, the challenge of aligning LLMs with our values and expectations has become a pressing concern, one that demands a deep understanding and effective control of these models. To tackle these challenges, mechanistic interpretability stands out as a promising approach, offering a way to understand the underlying workings of these models.

Understanding Reasoning Capability. The field of XAI in machine learning had made significant progress with techniques designed to provide valuable insights to end users, such as feature attributions (Ribeiro et al., 2016) and example-based explanations (Koh and Liang, 2017). These techniques have been proven to be quite effective in computer vision tasks, where the demands for complex alignment were less strict. However, as LLMs become more sophisticated, their reasoning capability has transformed from mere pattern recognition to a form of complex, human-like cognition. This advancement in LLMs' reasoning abilities renders traditional XAI methods obsolete and less competent in interpreting their behaviors.

Understanding Inner Working of LLMs. Moreover, alongside the strong reasoning abilities of LLMs, their notorious deep and intricate architectures are raising new concerns. Since the inner workings of these models are multifaceted and intricate, new challenges in explaining model at the structure level have emerged. Conventional global interpretability techniques, which are adept at uncovering the high-level knowledge acquired in different components of models, fall short when providing sights into the functions and the evolution of knowledge within these models. This issue is further confounded as LLMs scale aggressively, making neuron-level and layer-level insights increasingly insufficient. This complexity highlights the urgent need for innovative approaches that enable us to zoom in models and provide more in-depth, mechanistic understandings at various levels.

Alternatively, mechanistic interpretability aims to unravel the inner workings of LLMs, providing insights into the "how" and "why" behind their decision-making processes. Specifically, mechanistic interpretability delves into the causal relationships and underlying mechanisms within models. This not only is more suited to the advanced nature of LLMs, but is also crucial to ensure transparency, trust, and reliability in their applications.

2.3 Mechanistic Interpretability Theories

Most of the current work on mechanistic interpretability is based on vision models, and some recent work has begun to investigate Transformer models. In this section, we introduce some core concepts and pivotal phenomenons in the field of mechanistic interpretability. Since LLMs are too complicated to analyze locally, simple yet artificial models are purposely designed to investigate their characteristics and internal mechanisms. We will introduce the main assumptions and observations made under this setting, including *circuits*, *induction heads*, *superposition*, *polysemanticity*, and *monosemanticity*.

Circuits. "Circuit" is one of the core concepts in the field of mechanistic interpretability. It was first proposed to reverse engineer vision models, in which individual neurons and their connections are considered functional units (Olah et al., 2020a). Some researchers believe that features existing in earlier layers are fundamental units of models, such as edge detectors. And these features are combined by weights to form a circuit unit. This view is partially evidenced by a few understandable neuron units (circuits) firing for directions, such as curve detectors (Cammarata et al., 2020) and high-low frequency detectors (Schubert et al., 2021). Several phenomenons have been observed in these circuits. For example, symmetric transformations of basic features can be achieved with basic neurons also known as "equivariance" or "motif", which include copying, scaling, flipping, coloring, rotating, etc. (Olah et al., 2020b).

Based on insights derived from vision models, a mathematical framework for transformer *circuits* has also been proposed (Elhage et al., 2021). To avoid the intricacies associated with LLMs, this framework focuses on decoder-only transformers with no more than two layers, comprised entirely of attention blocks. Thus, within this toy model, the transformer encompasses input embedding, residual stream, attention layers, and output embeddings. Attention layers read information from residual stream and then write their output back into the residual stream. Communications can occur at the layer level. Each attention head works independently in parallel and contributes its output to the residual stream. Within each head, there are two circuits: i) "query-key" (QK) circuits, responsible for determining attention patterns and source-to-destination token relationships that provide matching abilities. ii) “output-value” (OV) circuits, dictating how a given token influences the output logit and providing copying abilities. The result shows that transformers with zero layer can model bigram statistics, predicting the next token from the source token. Adding one layer allows the models to capture both bigram and “skip-trigram” patterns. Interestingly, with two layers, these transformers give rise to a concept termed as “induction head”. These induction heads exist in the second layer and beyond. Usually, they are composed of heads from their previous layer, which are useful in suggesting the next token based on the present ones.

Induction heads. The research indicates that *induction heads* are composed of two kinds of heads from the previous layer including a query head, which provides information from previous tokens, and a key head, which matches the destination token with the source token. If there is no information for them to copy, they often fall back to the first token. As a result, layers with induction heads are equipped with more powerful in-context learning abilities rather than simple copying. In addition, multiple empirical evidences have been presented to prove the causal relationships between induction heads and in-context learning abilities by observing the change of in-context learning abilities after manipulating induction heads (Olsson et al., 2022). Although this theory appears to offer a comprehensive explanation of the mechanisms behind Transformers with only two attention layers, further ablation studies are needed to validate its accuracy. In particular, this framework is exclusively based on attention heads, without incorporating MLP layers.

Superposition. Another important observed phenomena is *superposition*, which describes that different features can be spread across many polysemantic neurons (Olah et al., 2020a). It is believed to originate from the excessive number of features compared to the number of neurons. In the analysis of a toy example, i.e. a ReLU network, they suppose that superposition can not only be used to represent additional features but also tolerate interference, which is based on the ideal assumption that larger sparse network is able to disentangle all those features into specialized individual neurons. Superposition is more powerful as feature sparsity increases. Furthermore, superposition can also perform some kind of computation, such as the absolute value function in circuits (Olah et al., 2020a). Based on the concept of superposition, one work attempts to identify neurons firing for individual specific human-interpretable high-level features with sparse probing (Gurnee et al., 2023).

Polysemanticity. Polysemanticity occurs when individual neurons in neural networks respond to a variety of features. Superposition is identified as a key factor in understanding *polysemanticity* of neurons within models. A recent study investigates this by exploring the cause of polysemanticity through the the lens of “feature capacity”, denoting the proposition of embedding dimension consumed by a feature in the representation space (Scherlis et al., 2022). Adopting one-layer and two-layer toy models similar to those used in the superposition study (Olah et al., 2020a), this work investigates how feature sparsity and importance influence feature capacity allocation. The result indicates that features are represented based on their significance in reducing loss, with more important features being allocated their own dimensions, while less critical ones can be overlooked (Olah et al., 2020a; Scherlis et al., 2022). Features end up sharing dimensions only when assigning additional capacity results in an equal or substantial decrease in loss.

Monosemanticity. A comparable hypothesized phenomena is *monosemanticity*, in which polysemantic neurons can be elucidated by a combination of interpretable features using dictionary learning/a sparse autoencoder (Bricken and Pehlevan, 2021). The study is based on a one-layer transformer model equipped with a 512-neuron MLP layer. The sparse encoder is trained on MLP activations from 8B data points, with feature expansion ranging from 512 to 13,100. The analysis focuses primarily on an expansion with 4,096 features learned in a single run. During the experiment, the hypothesis about disentangling features in a larger sparse network is nearly debunked due to the increased loss and unmanageable performance. Alternatively, Bricken and Pehlevan (2021) adopts the sparse encoder to decompose MLP activations. The results show that dictionary learning can successfully extract monosemantic features such as Arabic text, DNA sequences, base64 strings, etc. Despite the challenges posed by recovering low-dimension data into a higher dimension, a recent study shows that ground truth features can be recovered through dictionary learning with a sparse autoencoder (Sharkey et al., 2022). It is worth noting that, in this study, the ground true features are pre-defined, which differs from real data with unknown number of learned features.

2.4 Understanding Grokking

In the preceding subsection, we discuss various observations of toy networks from a structural perspective. However, it is equally important to understand the training dynamics of these models, which remains a largely unexplored area. In the following, we will focus on a specific phenomena that challenges our belief about early stopping to prevent overfitting, known as *grokking*.

Grokking is a phenomenon where models suddenly improve their validation accuracy after severely overfitting on overparameterized neural networks (Power et al., 2022). The surge in validation accuracy is regarded as a gain of generalization ability. Experiments built on a two-layer decoder-only transformer network has shown that smaller datasets necessitate a greater number of optimization steps (Power et al., 2022). The minimal amount of data needed for grokking also hinges on minimal number of data points required to learn a robust representation (Liu et al., 2022a). Furthermore, it has been found that generalization often coincides with well-structured embeddings. Additionally, regularization measures can accelerate the onset of grokking, with weight decay standing out as particularly effective in bolstering generalization capabilities (Liu et al., 2022a).

When examining weight norms of the final layers in models that don't use regularization techniques, a phenomenon, termed as *slingshot mechanism*, has been observed. It describes a cyclic behavior during the terminal phase of training, where there are oscillations between stable and unstable regimes. It is characterized by a phase where weight norms grow, followed by a phase of norm plateau. Thilak et al. (2022) point out that grokking, non-trivial feature adaptation, occurs only at the beginning of slingshots. The appearance of the slingshot effect and grokking can be modulated by adjusting optimizer parameters, especially when using adaptive optimizers such as Adam (Kingma and Ba, 2014). However, it is unclear whether this observation holds universally across various scenarios. Additionally, another concept called the *LU mechanism* has also been proposed, focusing on dynamics between loss and weight norms (Liu et al., 2022b). In algorithmic datasets, an L-shaped training loss and a U-shaped test loss reduction concerning weight norms are identified, implying an optimal range for initializing weight norms. Nevertheless, this finding does not seamlessly transfer to real-world machine learning tasks, where large initialization and small weight decay are often necessary. Lyu et al. (2023); Mohamadi et al. (2023) attribute it to a competition between the early-phase implicit bias favoring kernel predictors induced by large initialization and a late-phase implicit bias favoring min-norm/margin predictors promoted by small weight decay. Similarly, Merrill et al. (2023) conclude that this competition manifests a competition between a dense subnetwork in the initial phase and a sparse one after grokking.

Another concept is also considered related to grokking, known as *double descent*. It captures the pattern where a model's test accuracy at log level initially improves, then drops due to overfitting, and finally increases again after gaining generalization abilities (Nakkiran et al., 2021). This pattern is more noticeable in the test loss. A unified framework has been developed to integrate grokking with double descent, treating them as two manifestations of the same underlying process (Davies et al., 2023). The framework attributes the transition of generalization to slower pattern learning, which has been further supported by Kumar et al. (2023). This transition is demonstrated to exist at the level of both epochs and models.

The relation between grokking and memorization has also been explored on algorithmic datasets and with two-layer neural models. Experiments using slightly corrupted datasets have revealed that memorization can coexist with generalization. Memorization can be mitigated by pruning relevant neurons or by regularization. While different regularization methods target various learning strategies, they all contribute to better generalizing representations. And the training process in the study consists of two stages: i) the grokking process, ii) the decay of memorization learning (Doshi et al., 2023). However, the underlying causes behind this process are not fully understood. And the assumption that regularization is the key to this process are under debate, especially in the light of observing grokking in absence of regularization (Kumar et al., 2023). The importance of the rate of feature learning and the number of necessary features are favored in explanations, challenging the role of the weight norm (Kumar et al., 2023).

Interestingly, Nanda et al. (2023) find that grokking is correlated with another phenomenon termed "*phase change*" (Olsson et al., 2022), observed during the training phase when using the Adam optimizer. "*Phase change*" is conceived as an indicator that models begin to gain in-context learning abilities. However, our understanding towards it is still in its infancy. The assumption of links between grokking and "phase change" needs further exploration. Besides, the study uncovers an algorithm that utilizes Discrete Fourier Transforms and trigonometric identities to achieve modular addition, with evidence of these operations embedded within the model's weights. The circuits that enable this algorithm seems to evolve in a steady manner instead of through randomly walking. It is also claimed that the training process encompass three distinct stages: memorization, circuit formation, and memorization cleanup. Consequently, it is hypothesized that grokking occurs gradually as memorized parts are removed.

Additionally, there are other observations in this field, such as neural collapse (Papyan et al., 2020), yet there is a notable gap in understanding how these observations are interconnected. The root causes of these observations often leads to conflicting viewpoints. For example, Gromov (2023) suggests that grokking might be triggered by the learning of a new feature. Unfortunately, the leap in generalization could be too subtle to notice without a hierarchical model (Gromov, 2023). On the other hand, there is some debate around linking grokking with generalization (Levi et al., 2023). Moreover, a significant limitation of these studies is their focus on arithmetic datasets instead of real-world datasets, which casts doubt on how broadly these findings can be applied. To fully understand the generalization of models and reconcile these conflicting views, a holistic examination of how these observations relate to each other and their impact on training dynamics across models is essential.

2.5 Application to LLM Alignment

In this section, we summarize the potential applications of mechanistic interpretability techniques on LLMs. These applications aim at evaluating models' alignments from different views.

Inspired by induction heads, a recent work measures bias scores of attention heads in pretrained LLMs, focusing on specific stereotypes. It has also implemented a method to ensure the accuracy of identifying biased heads by comparing the changes in attention score between biased heads and regular heads. By masking the biased heads identified, the study effectively reduces the gender bias present in the model (Yang et al., 2023).

Similarly, one line of work localizes attention heads that are responsible for lying with linear probing and activation patching. A set of carefully designed prompts are employed to instruct LLMs to be dishonest. Linear probes are trained to classify true and false activations of attention heads. Then, activations relevant to lying behaviors are patched with those of honest behaviors to observe the change of outputs. Multiple attention heads across five layers are casually located (Campbell et al., 2023).

In contrast to decipher the inner workings of models, one study shifts focus to examine the differences between pre-training and fine-tuning phases. It reveals that fine-tuning retains all the capabilities learned in pre-training phase. Transformations are due to "wrappers" learned on top of models. Interestingly, these wrappers can be eliminated by pruning a few neurons or by retraining on an unrelated downstream task (Jain et al., 2023). This discovery sheds light on potential safety concerns associated with current alignment approaches.

3 Representation Engineering

As mentioned in Section 2.1, mechanistic interpretability represents a narrow field of XAI, relying on tailored, post-hoc, global, and white-box methods to explain the inner workings of LLMs. Within this post-hoc, global and white-box landscape of XAI, there are also other emerging philosophies. One notable direction is representation engineering (Zou et al., 2023).

3.1 Why Representation Engineering?

Representation engineering excels at offering an intuitive grasp of the learned embedding spaces. Through visualizing these learned representations, we can gain causal insights that help explain and potentially refine models' behaviors. Furthermore, the use of probing techniques plays a key role in identifying specific parts within models that are crucial for achieving alignments.

3.2 Representation Engineering Algorithms

Representation engineering techniques follow the probing-based analysis paradigm, which could date back to the BERT era (Zhao et al., 2023). These techniques can be further grouped into two categories: unsupervised and supervised methods.

Unsupervised Methods: A typical example is using principle component analysis (PCA) for visualization. To obtain unambiguous representation, Marks and Tegmark (2023) create self-curated true/false datasets to study the geometry of true/false statement representations derived from a model's residual stream. After applying PCA, a clear linear structure emerges. The truth directions are leveraged to mediate model's dishonest behaviors locally. Another avenue of study employs LLMs to explain representations in natural language. Firstly, the representations of the original models are extracted and then transformed. These transformed activations are patched into a translation model, which has been trained using data from previous interpretability methods. This approach has proven to be as effective as, or even better than, exiting probing techniques in 12 factual reasoning tasks (Ghandeharioun et al., 2024).

Supervised Methods: Building a linear model on top of representations stands as a fundamental technique in representation engineering to identify specific behaviors. However, one notable limitation of this method is that each classifier is designed to predict just one particular type of behavior. For example, hallucination detection targets at discriminating if a response is generated from prompts and model memories or from extrapolation from prompts (CH-Wang et al., 2023). A recent work has been developed to identify hallucination tokens from the response by ensembling a range of classifiers that are trained for all layers on separate hidden parts: MLPs and attention layers (CH-Wang et al., 2023). The integrated classifier is responsible for performing hallucination detection in response.

3.3 Application to LLM Alignment

Existing bias measurements rely heavily on carefully designed prompts (Tamkin et al., 2023). The effectiveness of these measurements is determined by the comprehensiveness of these prompts. However, prompts are limited to capturing only recognized biases using a finite set of examples. This fails to provide a thorough way to uncover biases that have been learned but not explicitly known. Recently, representation engineering has emerged as a promising avenue for detecting such biases within embedding spaces.

A notable study suggests that MLPs operate on token representations to alter the distribution of output vocabulary (Geva et al., 2022). After reverse engineering MLPs, it is believed that the output from each feed-forward layer can be seen as sub-updates to output vocabulary distributions, essentially promoting certain high-level concepts. This insight has been used effectively to mitigate toxicity levels in LLMs (Geva et al., 2022). Another work finds multiple representation vectors within MLPs that encourage toxicity. These vectors are decomposed using singular value decomposition, allowing researchers to pinpoint specific dimensions that contribute to toxicity (Lee et al., 2024).

4 Further Discussion

In this section, we provide further discussion on different explanation scales of two techniques. Further, we provide our understanding towards their

Explanability Scale. These two techniques explain LLMs at opposite scales.

- **Micro-scale**: Mechanistic interpretability focuses on dissecting the intricate inner workings of LLMs at the neuron and circuit levels. It aims at illustrating how models function and process certain tasks with subnetworks.
- **Macro-scale**: Representation engineering places representations, rather than neurons or circuits, as the central unit of analysis. The goal is to understand and control cognitive behaviors by studying their manifestations in learned representation spaces.

Roles in XAI. Two techniques are providing multi-faceted perspectives in the field of XAI. Representation engineering embodies how well embeddings capture the essence of data. Good representations are crucial to making accurate predictions. The visualization of representation can also demonstrate implicitly the quality of learning. On the other hand, through the lens of mechanistic interpretability, we can delve into relations between models' abilities like generalization and training dynamics. Examining the evolution of models from initialization to generalization, we can reveal characteristics of generalization, such as sparsity. These characteristics could serve as benchmarks for what constitutes "good learning". Apart from that, mechanistic interpretability is known to explain individual functional components and potentially improve model performance in the future.

Potential to Alignment. At current stage, both techniques have witnessed preliminary applications in LLM alignment. Mechanistic interpretability plays a crucial role in locating knowledge or biases at the level of attention heads, while representation engineering is primarily employed in targeting undesired behaviors at the level of layers. Despite distinct focus of each approach within models, both have proven effective in identifying biases and highlighting practical steps for improvement. However, they are still incompetent in uncovering rudimentary causes behind these biases.

5 Research Challenges

In this section, we outline the research challenges that deserve future efforts from the community.

5.1 The Validity of Existing Theories

While theories that attempt to explain the mechanisms behind the capabilities of Transformers are promising, their empirical support is not definitive. For example, understanding induction heads is key to explain Transformers because they are recognized as foundations for in-context learning abilities. However, as highlighted by Olsson et al. (2022), defining what exactly an induction head is remains somewhat elusive. Similarly, the proposition of mathematical framework to explain circuits inside a simplified network opens up an interesting avenue of research. Although Lieberum et al. (2023) conclude that circuit analysis is feasible on LLMs, this theoretical framework has not been thoroughly tested with empirical studies. Besides, these theoretical models rely on idealized assumptions such as superposition and often lack ground truth. This further complicates the task of validating these theories.

5.2 The Curse of Dimensionality

Another challenge is that the parameters we can explain are much less than a third of all parameters in LLMs. These explanations focus on components of attention heads, and although dictionary learning helps to partially understand polysemantic neurons, there is still a vast territory that remains unexplored. The rest majority of these model parameters are tied to MLP layers, which are notoriously difficult to fully comprehend (Olsson et al., 2022). Their compositions are more complicated than those of attention layers, making the analysis process considerably more arduous and perplexing. For instance, Geva et al. (2021) believes that the output of MLPs is a composition of memories including textual patterns and output distributions. Meng et al. (2022) attempt to modify MLPs to edit factual knowledge in LLMs. However, the effectiveness of editing has been put into doubt by another work (Hase et al., 2023).

5.3 Evaluation of Concepts and Circuits

A key challenge in mechanistic interpretability is validating and ensuring the accuracy of proposed conceptual explanations and functional circuits. Unlike straightforward metrics in machine learning to assess predictions, interpretation evaluation lacks clear ground truth. As noted in Chan et al. (2022), we are short of tools to measure the degree to which explanations interpret the relevant phenomenon. Existing ad-hoc ablation methods, i.e. standard zero and mean ablations, are neither universal nor scalable. Exploring measurements from various angles, such as casual scrubbing, which involves randomly sampling inputs to patch activations without disturbing the input distribution, could enrich our evaluation dimensions. Moreover, manual inspections are challenging in identifying circuits within LLMs. Our understanding of automatically discovering these circuits is still developing (Wang et al., 2022). Heterogeneous mechanistic explanations can be generated in networks trained on simple tasks such as modular additions (Zhong et al., 2023). This suggests that even in seemingly simple scenarios, the outcomes of circuit analysis can be uncertain. Additionally, different models learned on similar tasks might learn same family of circuits, but the precise circuits learned by individual networks are not the same (Chughtai et al., 2023).

6 Conclusions

In this paper, we investigate the techniques that allow us to examine the holistic interpretability of LLMs with the goal of better alignment. We center on two main paradigms: mechanistic interpretability and representation engineering. We introduce the key techniques within two paradigms, as well as their applications to enhance LLM alignment. Additionally, we also share our insights and visions on them, outlining open challenges around validating theories, tackling complexity, and precisely defining target behaviors.

Limitations

In this paper, we study the XAI techniques that can zoom in LLMs to provide insights on decision making. We primarily focus on the mechanistic interpretability and representation engineering. Despite the valuable perspectives, our position study has some notable limitations. We do not explore the complete landscape of relevant XAI methods for understanding LLMs, due to the space constraints. Other techniques like concept-based explanations, example-based explanations, and counterfactual explanations may also provide some useful insights on LLM inner workings as well.

References

AnthropicAI. 2023. Introducing claude.

Trenton Bricken and Cengiz Pehlevan. 2021. Attention approximates sparse distributed memory. *Advances in Neural Information Processing Systems*, 34:15301–15315.

Nick Cammarata, Gabriel Goh, Shan Carter, Ludwig Schubert, Michael Petrov, and Chris Olah. 2020. Curve detectors. *Distill*. Https://distill.pub/2020/circuits/curve-detectors.

James Campbell, Richard Ren, and Phillip Guo. 2023. Localizing lying in llama: Understanding instructed dishonesty on true-false questions through prompting, probing, and patching. *arXiv preprint arXiv:2311.15131*.

Sky CH-Wang, Benjamin Van Durme, Jason Eisner, and Chris Kedzie. 2023. Do androids know they're only dreaming of electric sheep? *arXiv preprint arXiv:2312.17249*.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldwosky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. 2022. Causal scrubbing, a method for rigorously testing interpretability hypotheses. *AI Alignment Forum*. https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. 2023. Neural networks learn representation theory: Reverse engineering how networks perform group operations. In *ICLR 2023 Workshop on Physics for Machine Learning*.

Xander Davies, Lauro Langosco, and David Krueger. 2023. Unifying grokking and double descent. *arXiv preprint arXiv:2303.06173*.

Darshil Doshi, Aritra Das, Tianyu He, and Andrey Gromov. 2023. To grok or not to grok: Disentangling generalization and memorization on corrupted algorithmic datasets. *arXiv preprint arXiv:2310.13061*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Dan Friedman, Alexander Wettig, and Danqi Chen. 2023. Learning transformer programs. *arXiv preprint arXiv:2306.01128*.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *arXiv preprint arXiv:2203.14680*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscope: A unifying framework for inspecting hidden representations of language models. *arXiv preprint arXiv:2401.06102*.

Andrey Gromov. 2023. Grokking modular arithmetic. *arXiv preprint arXiv:2301.02679*.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. *arXiv preprint arXiv:2301.04213*.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. 2023. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. *arXiv preprint arXiv:2311.12786*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR.

Tanishq Kumar, Blake Bordelon, Samuel J Gershman, and Cengiz Pehlevan. 2023. Grokking as the transition from lazy to rich training dynamics. *arXiv preprint arXiv:2310.06110*.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K Kummerfeld, and Rada Mihalcea. 2024. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity. *arXiv preprint arXiv:2401.01967*.

Noam Levi, Alon Beck, and Yohai Bar-Sinai. 2023. Grokking in linear estimators–a solvable model that groks without understanding. *arXiv preprint arXiv:2310.16441*.

Tom Lieberum, Matthew Rahtz, János Kramár, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. 2023. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *arXiv preprint arXiv:2307.09458*.

Ziming Liu, Ouail Kitouni, Niklas S Nolte, Eric Michaud, Max Tegmark, and Mike Williams. 2022a. Towards understanding grokking: An effective theory of representation learning. *Advances in Neural Information Processing Systems*, 35:34651–34663.

Ziming Liu, Eric J Michaud, and Max Tegmark. 2022b. Omnigrok: Grokking beyond algorithmic data. *arXiv preprint arXiv:2210.01117*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon S Du, Jason D Lee, and Wei Hu. 2023. Dichotomy of early and late phase implicit biases can provably induce grokking. *arXiv preprint arXiv:2311.18817*.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

William Merrill, Nikolaos Tsilivis, and Aman Shukla. 2023. A tale of two circuits: Grokking as competition of sparse and dense subnetworks. *arXiv preprint arXiv:2303.11873*.

Mohamad Amin Mohamadi, Zhiyuan Li, Lei Wu, and Danica Sutherland. 2023. Grokking modular arithmetic can be explained by margin maximization. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. 2021. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020a. Zoom in: An introduction to circuits. *Distill*. Https://distill.pub/2020/circuits/zoom-in.

Chris Olah, Nick Cammarata, Chelsea Voss, Ludwig Schubert, and Gabriel Goh. 2020b. Naturally occurring equivariance in neural networks. *Distill*. Https://distill.pub/2020/circuits/equivariance.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

OpenAI. 2023. Gpt-4 technical report.

Vardan Papyan, XY Han, and David L Donoho. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Adam Scherlis, Kshitij Sachan, Adam S Jermyn, Joe Benton, and Buck Shlegeris. 2022. Polysemanticity and capacity in neural networks. *arXiv preprint arXiv:2210.01892*.

Ludwig Schubert, Chelsea Voss, Nick Cammarata, Gabriel Goh, and Chris Olah. 2021. High-low frequency detectors. *Distill*. Https://distill.pub/2020/circuits/frequency-edges.

Lee Sharkey, Dan Braun, and beren. 2022. [Interim research report] Taking features out of superposition with sparse autoencoders. Accessed 2024-01-23.

Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. 2023. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Vimal Thilak, Etai Littwin, Shuangfei Zhai, Omid Saremi, Roni Paiss, and Joshua M. Susskind. 2022. The slingshot mechanism: An empirical study of adaptive optimizers and the \emph{Grokking Phenomenon}. In *Has it Trained Yet? NeurIPS 2022 Workshop*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.

Yi Yang, Hanyu Duan, Ahmed Abbasi, John P Lalor, and Kar Yan Tam. 2023. Bias a-head? analyzing bias in transformer-based language model attention heads. *arXiv preprint arXiv:2311.10395*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. 2023. The clock and the pizza: Two stories in mechanistic explanation of neural networks. *arXiv preprint arXiv:2306.17844*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.