Towards Uncovering How Large Language Model Works: An Explainability Perspective

Anonymous ACL submission

Abstract

Large language models (LLMs) have led to breakthroughs in language tasks, yet the internal mechanisms that enable their remarkable generalization and reasoning abilities remain opaque. This lack of transparency presents challenges such as hallucinations, toxicity, and misalignment with human values, hindering the safe and beneficial deployment of LLMs. This paper aims to uncover the mechanisms underlying LLM functionality through the lens of explainability. First, we review how knowledge is architecturally composed within LLMs and encoded in their internal parameters via mechanistic interpretability techniques. Then, we summarize how knowledge is embedded in LLM representations by leveraging probing techniques and representation engineering. Additionally, we investigate the training dynamics through a mechanistic perspective to explain phenomena such as grokking and memorization. Lastly, we explore how the insights gained from these explanations can enhance LLM performance through model editing, improve efficiency through pruning, and better align with human values.

1 Introduction

Large language models (LLMs) such as GPT-4 (OpenAI, 2023), LLaMA-2 (Touvron et al., 2023), Claude-3 (AnthropicAI, 2023), and Gemini (Team et al., 2023) have led to tremendous advancements in language understanding and generation, achieving state-of-the-art performance in a wide array of real-world tasks. Despite their superior performance across various tasks, the “how” and “why” behind their generalization and reasoning abilities are still not well understood. This lack of understanding poses several challenges. First, LLMs frequently generate hallucinations and factually incorrect output, which complicates efforts to improve their performance. Second, as LLMs become more powerful, problems surrounding potential toxicity, unfairness, and dishonesty threaten to undermine user trust. Third, the impressive generalization capabilities of LLMs suggest that they may be acquiring and leveraging knowledge in ways that fundamentally differ from traditional machine learning approaches. Therefore, there is an urgent need to delve deeper into the inner workings of LLMs to fully address these issues. Gaining insights into how these models operate is a crucial step towards developing robust safeguards and ensuring their responsible deployment.

In this paper, we provide a systematic overview of the existing literature that uncovers the working mechanisms of LLMs using existing explainability techniques (Figure 1). First, we provide a summary of findings on probing trained LLMs to understand how knowledge is composed in model architectures. To achieve this, the main explainability technique used is mechanistic interpretability. It focuses on the functionality of each model component and interprets how models operate at the level of neurons, circuits, and attention heads. Second, we examine how knowledge is encoded internally in intermediate representations. To this end, representation engineering is adopted to explain specific behavior of the model, such as dishonesty, by analyzing hidden representations (Zou et al., 2023). Third, we inspect the model training process to understand the development of generalization abilities during the training process. Finally, we review how insights from the aforementioned analysis help us improve models in terms of higher performance through model editing, better efficiency through pruning, and better human alignment.

Our work differs from existing survey articles on the explainability of LLMs (Zhao et al., 2023; Wu et al., 2024b; Luo and Specia, 2024), which either summarize explainability techniques or discuss their utilities. On the contrary, our goal is to review existing studies to uncover how LLMs function and identify the factors that contribute to their reasoning abilities via using explainabil-

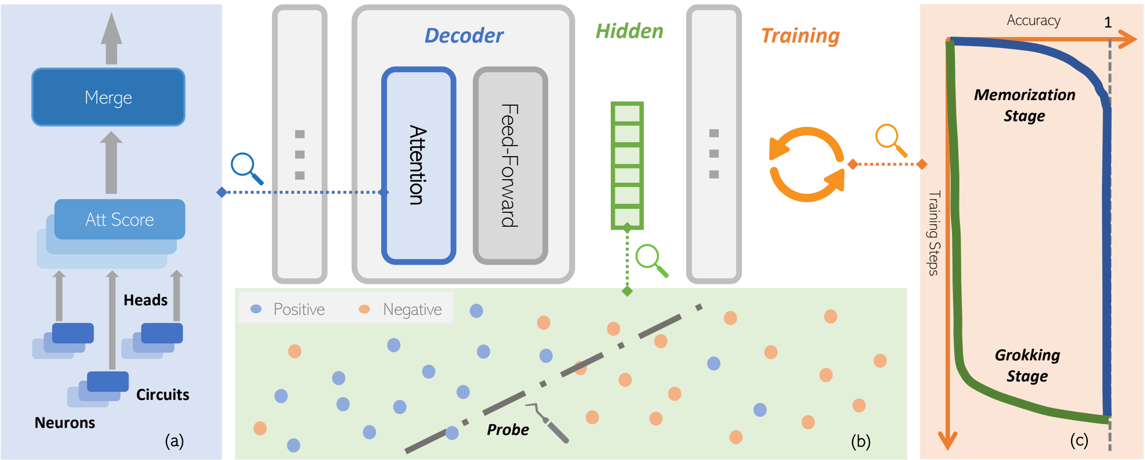

Figure 1: In this work, we review existing progress on how LLMs work, including: a) how knowledge is architecturally composed within model components; b) what knowledge is encoded in intermediate representations; and c) how generalization abilities are achieved during the training process.

ity techniques and monitoring the training process. We review the state-of-the-art insights on the inner workings of LLMs and explore how these insights can further enhance model performance and benefit humans. We believe that a deep understanding of how LLMs work is crucial for their safe and beneficial deployment in real-world applications.

2 How is Knowledge Composed in Model Architectures?

LLMs are built on extensive training datasets and intricate model architectures, which contribute to their remarkable emergent abilities (Wei et al., 2022). However, the exact mechanisms through which these models acquire and process vast amounts of knowledge remain unclear. Additionally, the contributions of individual model components to the overall function have been largely unexplored. To fully understand LLMs, recent studies have shifted to make use of mechanistic interpretability (more details are given in Section A at the Appendix) to reverse engineer LLMs at a more granular level such as neurons, attention heads, and connections.

2.1 Neurons

Neurons serve as the basic units for storing knowledge and patterns within LLMs. They are observed to be *polysemantic*, meaning that an individual neuron can be activated on multiple unrelated terms (Olah et al., 2020a; Bills et al., 2023). This characteristic presents a significant challenge in mechanistically understanding how models operate. Despite the challenge, recent work has explored the underlying causes of this polysemantic nature. Two key concepts have emerged as instrumental in unraveling its formation: *Superposition* (Olah et al., 2020a) and *Monosemanticity* (Bricken et al., 2023).

2.1.1 Superposition

Superposition describes the phenomenon where a feature can be spread across multiple neurons, meanwhile a neuron can also be mixed up with multiple features. Some researchers believe that this mechanism originated from an excessive number of features compared to the number of neurons (Olah et al., 2020a; Elhage et al., 2022). In exploring this concept through a toy example, i.e. a ReLU network, researchers have found that superposition allows for the representation of additional features. However, to mitigate interference, a nonlinear filter needs to be introduced (Elhage et al., 2022). When features are sparse, superposition effectively supports the representation of these features and allows computations such as the absolute value function (Elhage et al., 2022). Neurons within models can be either monosemantic or polysemantic.

Others argue that polysemanticity arises incidentally due to factors encountered during the training process such as regularization and neural noise (Lecomte et al., 2024). Mathematical demonstrations have shown that a constant fraction of feature collisions, introduced through random initialization, can always result in polysemantic neurons, even when the number of neurons exceeds the number of features (Lecomte et al., 2024).

Another study investigates polysemanticity through the lens of the “feature capacity”, denoting the fraction of embedding dimensions consumed by a feature in the representation space (Scherlis et al., 2022). By analyzing one-layer and two-layer toy models, this work indicates that features are represented based on their importance in reducing loss. More important features are allocated their own dimensions, while the less critical ones may be overlooked, and the rest will share embedding dimensions (Scherlis et al., 2022). Features only end up sharing dimensions when assigning additional capacity will not result in loss decreasing (Scherlis et al., 2022). Moreover, the relationship between superposition and feature importance has been demonstrated on LLMs (Gurnee et al., 2023). Experiments show that the early layers tend to represent many features in superposition, while the middle layers include dedicated neurons to represent high-level features (Gurnee et al., 2023).

2.1.2 Monosematicity

Monosemantic neurons, associated with a single concept, are much easier to interpret than polysemantic neurons. Investigating the factors that enhance monosemanticity is meaningful to model interpretation. A research using toy models reveals that changing the loss minimum could improve monosemanticity. Such loss minimum usually coexists with negative biases (Jermyn et al., 2022). However, in reality building a purely monosemantic model is infeasible due to the unmanageable loss (Bricken et al., 2023). Another line of studies seeks to disentangle superposition to reach a monosemantic understanding. The spare autoencoder emerges as a promising tool for this purpose, particularly through the method dictionary learning where features are predefined (Sharkey et al., 2022). The effectiveness of this approach largely depends on the comprehensiveness of the pre-defined dictionary. Bricken et al. (2023) using a one-layer transformer model with a 512-neuron MLP layer highlights this approach. The sparse autoencoder is trained on MLP activations from 8B data points, with autoencoder sizes ranging from 512 to 13,100 features. Larger autoencoders are able to achieve finer granularity in interpreting features, revealing details that cannot be discovered at the neuron level. These identified features can be used to manipulate the model's output, offering new ways to control and understand LLMs (Bricken et al., 2023).

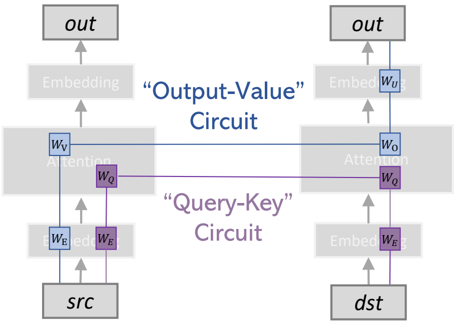

Figure 2: An illustration of a Transformer circuit, which is a key concept in mechanistic interpretability.

2.2 Circuits

Circuit is one of the core concepts in the field of mechanistic interpretability (see Figure 2). It was originally proposed to reverse engineer vision models, in which individual neurons and their connections are viewed as functional units (Olah et al., 2020a). Researchers have found that features in former layers of models act as fundamental units, such as edge detectors. These features are combined through weights to form a circuit unit. This viewpoint is partially evidenced by a few understandable neuron units (or circuits) performing specific functions, such as curve detectors (Cammarata et al., 2020) and high-low frequency detectors (Schubert et al., 2021). Several interesting phenomena have been observed in these circuits. For example, symmetric transformations of basic features, including copying, scaling, flipping, coloring, rotating, can be achieved with basic neurons known as "equivariance" or "motif" (Olah et al., 2020b).

Despite rich insights from vision models, transformer models, with their unique architecture featuring attention blocks, present new challenges. To address these, a mathematical framework specifically for *transformer circuits* has also been proposed (Elhage et al., 2021). This framework simplifies the complex architecture of LLMs by focusing on decoder-only transformer models that have no more than two layers, all made up entirely of attention blocks. Within this toy model, the transformer encompasses input embedding, residual stream, attention layers, and output embeddings. Attention layers read information from the residual stream and then write their output back into it. Consequently, communication is achieved through read and write operations at the layer level.

Each attention head works independently and in parallel, contributing its output to the residual stream. These heads consist of key, query, output, and value vectors, represented as $W_K, W_Q, W_O, \text{and} W_V$. There are two types of circuits: i) "query-key" (QK) circuits; ii) "output-value" (OV) circuits (Elhage et al., 2021), as shown in Figure 2. The QK circuits, formed by $W_Q^T W_K$, play a crucial role in determining which previously learned token to copy information from (Elhage et al., 2021). It is essential for models to recall and retrieve information from earlier context. Conversely, the OV circuits, composed of $W_O W_V$, determine how the current token influences the output logits (Elhage et al., 2021).

The result shows that transformers with no layer can model bigram statistics, predicting the next token from the source token. Adding one layer allows the model to capture both bigram and "skip-trigram" patterns. Interestingly, with two layers, transformer models give rise to a concept termed as "*induction head*" (Section 2.3). These induction heads exist in the second layer and beyond. Usually, they are composed of heads from their previous layer, which is useful in suggesting the next token based on the present ones (Elhage et al., 2021).

2.3 Attention heads

A special type of attention head called *induction head* is assumed critical in enabling in-context learning abilities within LLMs (Brown et al., 2020), due to the co-occurrence of induction heads and in-context learning (Olsson et al., 2022). Induction heads also refer to a kind of circuits that complete the pattern by prefix matching and copying previously occurred sequences (Olsson et al., 2022). They are composed of two heads: the first attention head is from the previous layer attending to previous tokens that are followed by the current token, which achieves prefix matching and provides the attend-to token(the token following current token). The second head, i.e. induction head, copies the attend-to token and increases its output logits. More specifically, this rule means that if models have seen similar patterns such as "[A*][B*]" given current token "[A]", these models are able to predict "[B]" (Olsson et al., 2022). Despite the single token used in the toy example, long prefix matching such as three consecutive tokens has also been observed in related work (Chan et al., 2022).

As a result, layers with induction heads possess more powerful in-context learning abilities than simple copying. In addition, multiple empirical studies have demonstrated 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; Chan et al., 2022). Although this theory offers a comprehensive explanation of the mechanisms behind transformer models with only two attention layers, further ablation studies are still needed to validate its effectiveness. It is also important to note that this framework is exclusively based on attention heads, without incorporating MLP layers.

3 What Knowledge is Encoded in Intermediate Representations?

In the previous section, we summarize existing studies on the architectural composition of knowledge within LLMs, with a focus on their structural components. We highlight how components of LLMs function differently. In this section, we introduce an in-depth review of the knowledge encoded by *representations of LLMs*, including world knowledge and factual knowledge captured within these models. We examine how factors such as the depth of layers and the scale of models influence this encoding process.

3.1 Probing World and Factual Knowledge

To investigate whether the representations of LLMs encode world knowledge and factual knowledge, probing techniques offer crucial insights into the structure and dynamics of these representations. Specifically, probing techniques can identify specific directions within the representation space, and these directions are essential for understanding certain behaviors and the encoding of knowledge (Zou et al., 2023; Liu et al., 2023).

Recent studies have demonstrated that LLMs can learn world models and encode them in their representations for certain tasks. One study successfully uses a set of non-linear probes to uncover world representations within models, specifically in the context of the game of Othello (Li et al., 2022). It demonstrates models' ability to track the board state, and make predictions without being explicitly to do so (Li et al., 2022). Furthermore, another work finds that linear representation structures can also perform well on predictions, simply by altering the expression of the board state at each timestamp (Nanda et al., 2023b). The linear and non-linear explanations reveal how models perceive the world naturally, which might be different from humans. Additionally, by analyzing representations of spatial datasets, one study reveals the model's ability to learn linear representations of space and time across multiple levels (Gurnee and Tegmark, 2023).

LLMs are also capable of encoding factual knowledge. Marks and Tegmark (2023) craft self-curated true/false datasets to study the geometry of representations of true/false statements derived from a model's residual stream. By applying principal component analysis (PCA), a clear linear structure emerges. The truth directions are leveraged to mediate the model's dishonest behaviors locally. Another research avenue explores vectors related to toxicity within MLP blocks through singular value decomposition (SVD). The identified dimensions are simply subtracted to efficiently achieve mitigation (Lee et al., 2024).

Function vectors have also been discovered within the attention heads of LLMs, which trigger the execution of a certain task across diverse inputs. For example, Todd et al. (2023) found that these function vectors are shown in various in-context learning tasks, and can execute related tasks despite zero-shot inputs. Also, causal interventions at the neuron level can help identify the individual neurons encoding spatial coordinates and time information (Gurnee and Tegmark, 2023).

Lastly, representations associated with undesirable behaviors of LLMs, such as dishonesty, toxicity, hallucinations, can also be extracted. Typically, a direction in the representation space is identified as contributing to a specific behavior. This direction will then be used to adjust the representations so that models' behaviors can be controlled (Zou et al., 2023). For example, Li et al. (2024) employs this technique to probe and enhance the truthfulness of models. Azaria and Mitchell (2023) also successfully distinguishes the truthfulness of statements by simply training a classifier on model representations. A recent work has been developed to identify hallucination tokens from the response by integrating a range of classifiers that are trained on each layer from separate hidden parts: MLPs and attention layers (CH-Wang et al., 2023).

3.2 Role of Layer Depth and Model Scale

The influence of layer depth and model scale on representations has been an interesting research direction. Empirically, research shows that a range of knowledge is well trained until the middle layers. For example, Gurnee and Tegmark (2023) demonstrate that space and time representations reach the best quality up to half of the layers in a range of open-source LLMs. Besides, the function vectors with strong causal effects are also collected from the middle layers of LLMs, while the effects are near zero in the deeper layers (Todd et al., 2023). Furthermore, another study shows that different levels of concepts are well learned in different layers, where simpler tasks are learned in the early layers while complex tasks can only be well learned in the deeper layers (Jin et al., 2024; Ju et al., 2024). However, the underlying reason why the middle layers perform so well remains unexplored.

It is generally believed that more capabilities are gained as the models scale up (Wei et al., 2022). Some recent studies have also supported this hypothesis in certain cases. For example, the space and time representations are more precise as the models scale up (Gurnee and Tegmark, 2023). But the inner mechanism leading to better performance when model scales remain unknown.

4 How is Generalization Ability Achieved During Training Process?

In the preceding sections, we analyze LLMs in a post-hoc manner, focusing on neurons, connections, attention heads, and representations to understand how knowledge is acquired within models. In this section, we discuss the dynamic training process of models to understand how the generalization ability is achieved during the training process. We will particularly examine two important phenomena observed in relation to generalization: grokking and memorization. Here, grokking indicates the phenomenon where models suddenly improve validation accuracy after overfitting. Investigating grokking can shed light on how generalization emerges during training. Moreover, examining memorization, where models rely on statistical patterns rather than causal relationships, can help disentangle the roles of generalization versus the roles of memorization in model behaviors.

4.1 Understanding Grokking

Grokking is a phenomenon in which models suddenly improve their validation accuracy after severely overfitting on over-parameterized neural networks (Power et al., 2022). The surge in validation accuracy is generally interpreted as a gain of generalization ability.

4.1.1 A Data Perspective

Experiments implemented on a two-layer decoder-only transformer network have shown that grokking is closely related to factors such as data, representations, and regularization. Smaller datasets require more optimization steps for grokking to occur (Power et al., 2022). Conversely, more samples can decrease the number of steps needed for generalization (Zhu et al., 2024). The minimal amount of data needed for grokking also depends on the 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 strengthening generalization capabilities (Liu et al., 2022a). A recent study proposes that massive datasets in LLMs make grokking less conceivable (Zhu et al., 2024).

4.1.2 Weight Norms

When examining the weight norms of the final layers in models that do not 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, i.e., training loss spike. The spike cooccurs with 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 the 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, describing 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.

4.1.3 Test Loss

Double descent captures the pattern where a model's test accuracy at the 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.

4.2 Memorization

Memorization often refers to the phenomenon that models predict with statistical features rather than causal relations. The study using slightly corrupted algorithmic datasets with two-layer neural models has revealed that memorization can coexist with generalization. And memorization can be mitigated by pruning relevant neurons or by regularization (Doshi et al., 2023). Although different regularization methods might not share learning goals, they all contribute to better 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 yet fully understood. Besides, the assumption that regularization is the key to this process is under debate, especially in 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 is favored in explanations, challenging the role of the weight norm (Kumar et al., 2023).

Interestingly, a study hypothesizes that memorization constitutes a phase of grokking (Nanda et al., 2023a). The study finds that grokking includes three distinct stages: memorization, circuit formation, and memorization cleanup (Nanda et al., 2023a). The study identifies an algorithm that utilizes Discrete Fourier Transforms and trigonometric identities to achieve modular addition through analyzing the model's weights. The circuits enabling this algorithm seem to evolve in a steady manner instead of randomly walking. However, our understanding towards the relationship between memorization and grokking is still limited.

5 How to Make Use of The Insights?

In the preceding three sections, we have explored how knowledge is architecturally composed within LLMs (Section 2), and how this knowledge is encoded in their representations (Section 3). Building on these insights, this section emphasizes on how we can leverage our in-depth understanding of LLMs to enhance their performance through editing, improve their efficiency via pruning, and better align them with human values and preferences.

5.1 Model Editing for Better Performance

Research has shown that it is possible to edit factual knowledge by modifying the weights of specific neurons in MLPs. One study successfully adopts this approach by altering neural computations related to recall of factual knowledge (Meng et al., 2022). Another study expands this method further to allow multiple edits at the same time (Meng et al., 2023). Although these methods are effective for targeted edits, their capabilities on updating relevant knowledge and preventing forgetting still require further investigation (Cohen et al., 2023).

Interestingly, a recent study indicates that the paragraphs memorized by a model can be pinpointed using high-gradient weights in attention heads of the lower layers (Stoehr et al., 2024). This research employs localization techniques to identify specific attention heads, which are then fine-tuned to unlearn the memorized knowledge. This approach holds promise in enhancing privacy protection in large language models, although a comprehensive evaluation is still needed.

Besides, facts are also encoded in the representation space, making representation a natural candidate to edit models' outputs. So far, most studies focus on modifying representations at inference time, while the influence of permanent modifications has barely been studied. A recent work provides a more precise way to edit model representations to change their output distributions (Hernandez et al., 2023). Instead of only adding the derived vectors into output representations, this study directly changes the embedding of a related entity so as to trigger targeted outputs. As a result, the position of the modified entity in the embedding space has changed, leading to causal influence on model generations.

5.2 Model Pruning for Better Efficiency

Causal tracing or causal mediation analysis from mechanistic interpretability serves as one of the fundamental techniques for studying neurons and attention heads. One study uses above method to reveal how the model processes the inputs, showing that the attention mechanism helps models extract query information into the final token in early layers, then result-related information will be incorporated into residual stream in the late MLP layers (Stolfo et al., 2023). This finding is meaningful for both pruning and fine-tuning when targeting specific queries.

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

Different from pruning neurons, the idea of representation engineering, that is directly manipulating representations without the need for optimization or additional labeled data, has also been demonstrated effective in model pruning. Some work attempts to fine-tune models with representation engineering and achieves a comparable and even better performance than state-of-the-art fine-tuning techniques (Wu et al., 2024a,c). One work employs forward passes from two topics and derives their difference vectors, which are used in inference time without additional fine-tuning (Turner et al., 2023). Wu et al. (2024a) also demonstrates the feasibility of fine-tuning models through editing representations. Unlike conventional parameter-efficient fine-tuning (PEFT), representation editing focuses on learning an additional group of trainable parameters to modify representations directly other than models' parameters. And the trainable parameters have been reduced to a factor of 32 compared to that of LoRA (Hu et al., 2021; Wu et al., 2024a). Another approach utilizes the distributed alignment search of Geiger et al. (2024) to find a set of linear subspace implementing interventions. This method outperforms most PEFT models on a range of tasks (Wu et al., 2024c).

5.3 Model Alignment to Human Values

From the mechanistic perspective, practical applications tend to evaluate model alignments with different tools. Inspired by induction heads, a recent work measures bias scores of attention heads in pre-trained LLMs, focusing on specific stereotypes. It implemented a method to ensure the accuracy of identifying biased heads by comparing the changes of attention score between biased and regular heads. Through masking identified biased heads, the study effectively reduces the gender bias encoded in the model (Yang et al., 2023). Besides, another work localizes attention heads that are responsible to lie with linear probing and activation patching. A set of intentionally designed prompts is used to instruct LLMs to be dishonest. Meanwhile, linear probes are trained to classify true/false activations of heads. Then, the selected activations are patched with those of honest behaviors to observe the changes of outputs. Multiple attention heads across five layers are causally located in Campbell et al. (2023).

Existing probing-based bias measurements rely heavily on carefully designed prompts known as prompt engineering (Tamkin et al., 2023). The effectiveness of these measurements is determined by the comprehensiveness of these prompts. However, the prompts are capable of capturing only recognized biases using a finite set of examples. This fails to provide an inclusive way to uncover biases that have been learned but are not explicitly known. Recently, representation engineering has emerged as a promising avenue for detecting such biases within embedding space. 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 line of work finds multiple representation vectors within MLPs that encourage models' undesired behaviors. These vectors are decomposed using singular value decomposition, allowing researchers to pinpoint specific dimensions that contribute to toxicity (Lee et al., 2024).

6 Conclusions and Looking Beyond

In this paper, we explore techniques to uncover the inner workings of LLMs through an explainability lens. We focus on two major paradigms of explainability: mechanistic interpretability and representation engineering. We provide a systematic overview of how these techniques can reveal the architectural composition of knowledge within LLMs and the encoding of knowledge in their internal representations. Furthermore, we inspect training dynamics through a mechanistic perspective to explain phenomena like "grokking" that can explain generalization abilities of LLMs. Lastly, we reviewed how insights from these explainability analyses can enhance LLM performance through model editing, improve efficiency via pruning, and better align models with human preferences.

Although there is some preliminary progress in uncovering the inner workings of LLMs, looking beyond, there exist several critical challenges and opportunities. First, LLMs have encoded a vast amount of real-world knowledge into their architectures and parameters. However, current research has only revealed a small fraction of the encoded knowledge. Future efforts should focus on developing scalable techniques that can effectively analyze and interpret the intricate knowledge structures embedded within LLMs. Second, LLMs have demonstrated remarkable reasoning abilities exhibiting human-like cognitive abilities. However, our current understanding of how these high-level reasoning abilities emerge from the interplay of architectural components and training dynamics is limited. More efforts are needed to reveal the intricate mechanisms that give rise to these advanced reasoning capabilities. Third, although the insights gained from mechanistic interpretability and representation engineering have enabled preliminary efforts in areas such as model editing, pruning, and alignment, the progress achieved thus far has been relatively modest. More work is required to fully leverage these insights and develop techniques that can substantially improve LLM performance.

Limitations

In this paper, we intend to integrate available techniques that enable us to learn the inner workings of LLMs. Despite the valuable perspectives provided, our study has several limitations. First, we do not explore the complete landscape of relevant XAI methods for understanding LLMs, due to space constraints. Other techniques like concept-based explanations, example-based explanations, and counterfactual explanations may also provide some useful insights into the inner workings of LLMs. These methods could potentially uncover additional aspects or offer complementary viewpoints that are not covered by the mechanistic interpretability and representation engineering approaches discussed in this paper. Furthermore, while we try to provide a comprehensive overview of the current state-of-the-art, the field of explainable AI for LLMs is rapidly evolving. New techniques, theories, and findings may emerge that could reshape or extend our understanding of how LLM works. Continuous monitoring and incorporating these developments will be crucial to maintaining a comprehensive and up-to-date perspective on this topic.

References

AnthropicAI. 2023. Introducing claude.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when its lying. *arXiv preprint arXiv:2304.13734*.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. `https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html`.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*.

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*.

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*.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2023. Evaluating the ripple effects of knowledge editing in language models. *arXiv preprint arXiv:2307.12976*.

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, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/toy_model/index.html.

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*.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. 2024. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pages 160–187. PMLR.

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.

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*.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

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*.

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2023. Inspecting and editing knowledge representations in language models.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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*.

Adam S Jermyn, Nicholas Schiefer, and Evan Hubinger. 2022. Engineering monosemanticity in toy models. *arXiv preprint arXiv:2211.09169*.

Mingyu Jin, Qinkai Yu, Jingyuan Huang, Qingcheng Zeng, Zhenting Wang, Wenyue Hua, Haiyan Zhao, Kai Mei, Yanda Meng, Kaize Ding, Fan Yang, Mengnan Du, and Yongfeng Zhang. 2024. Exploring concept depth: How large language models acquire knowledge at different layers?

Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. 2024. How large language models encode context knowledge? a layer-wise probing study. *arXiv preprint arXiv:2402.16061*.

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*.

Victor Lecomte, Kushal Thaman, Rylan Schaeffer, Naomi Bashkansky, Trevor Chow, and Sanmi Koyejo. 2024. What causes polysemanticity? an alternative origin story of mixed selectivity from incidental causes. In *ICLR 2024 Workshop on Representational Alignment*.

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*.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2022. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

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*.

Wenhao Liu, Xiaohua Wang, Muling Wu, Tianlong Li, Changze Lv, Zixuan Ling, Jianhao Zhu, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. 2023.

Aligning large language models with human preferences through representation engineering. *arXiv preprint arXiv:2312.15997*.

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.

Haoyan Luo and Lucia Specia. 2024. From understanding to utilization: A survey on explainability for large language models. *arXiv preprint arXiv:2401.12874*.

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.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

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. 2023a. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*.

Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023b. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*.

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.

Niklas Stoehr, Mitchell Gordon, Chiyuan Zhang, and Owen Lewis. 2024. Localizing paragraph memorization in language models. *arXiv preprint arXiv:2403.19851*.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052.

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 *Grokking Phenomenon*. In *Has it Trained Yet? NeurIPS 2022 Workshop*.

Eric Todd, Millicent L Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. 2023. Function vectors in large language models. *arXiv preprint arXiv:2310.15213*.

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*.

Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.

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*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Muling Wu, Wenhao Liu, Xiaohua Wang, Tianlong Li, Changze Lv, Zixuan Ling, Jianhao Zhu, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. 2024a. Advancing parameter efficiency in fine-tuning via representation editing. *arXiv preprint arXiv:2402.15179*.

Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du, et al. 2024b. Usable xai: 10 strategies towards exploiting explainability in the llm era. *arXiv preprint arXiv:2403.08946*.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2024c. Reft: Representation finetuning for language models.

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*.

Xuekai Zhu, Yao Fu, Bowen Zhou, and Zhouhan Lin. 2024. Critical data size of language models from a grokking perspective. *arXiv preprint arXiv:2401.10463*.

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*.

A 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 etc. Instead, it draws inspiration 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.

A.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.

A.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 look through several factors that we believe have played 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 dishonesty 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 has 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 models 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 focuses on 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.

A.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*.

B Mechanistic Interpretability v.s. Representation Engineering

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 multifaceted 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 implicitly demonstrate the quality of learning. On the other hand, through the lens of mechanistic interpretability, we can investigate 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 the 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 the 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.

C Research Challenges

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

C.1 The Validity of Existing Theories

While theories that attempt to explain the mechanisms behind the capabilities of transformer models are promising, their empirical support is not definitive. For example, understanding induction heads is key to explain transformer models 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 a 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.

C.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).

C.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 causal 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).

C.4 Conflicted Explanations

There are other observations in understanding observations, 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 lead 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.