

# From Understanding to Utilization: A Survey on Explainability for Large Language Models

**Haoyan Luo**

Imperial College London

`h.luo23@imperial.ac.uk`

**Lucia Specia**

Imperial College London

`l.specia@imperial.ac.uk`

## Abstract

This survey paper delves into the burgeoning field of explainability for Large Language Models (LLMs), a critical yet challenging aspect of natural language processing. With LLMs playing a pivotal role in various applications, their “black-box” nature raises concerns about transparency and ethical use. This paper emphasizes the necessity for enhanced explainability in LLMs, addressing both the general public’s trust and the technical community’s need for a deeper understanding of these models. We concentrate on pre-trained Transformer-based LLMs, such as LLaMA (Touvron et al., 2023), which present unique interpretability challenges due to their scale and complexity. Our review categorizes existing explainability methods and discusses their application in improving model transparency and reliability. We also discuss representative evaluation methods, highlighting their strengths and limitations. The goal of this survey is to bridge the gap between theoretical understanding and practical application, offering insights for future research and development in the field of LLM explainability.

## 1 Introduction

In the rapidly evolving field of natural language processing, Large Language Models (LLMs) have emerged as a cornerstone, demonstrating remarkable proficiency across a spectrum of tasks. Despite their effectiveness, LLMs, often characterized as “black-box” systems, present a substantial challenge in terms of explainability and transparency. This opacity can lead to unintended consequences, such as generating harmful or misleading content (Gehman et al., 2020), and the occurrence of model hallucinations (Weidinger et al., 2021). These issues underscore the urgency for enhanced explainability, not just for understanding but for responsible and ethical application.

Explainability in LLMs serves two critical functions. For end users, it fosters trust by clarifying the

model’s reasoning in a non-technical manner, enhancing understanding of their capabilities and potential flaws (Zhao et al., 2023). For developers and researchers, it offers insights into unintended biases and areas for improvement, serving as a tool for refining model performance on downstream tasks (Bastings et al., 2022; Meng et al., 2023a; Li et al., 2023b). However, the scale of LLMs poses unique challenges for explainability. Larger models with more parameters and extensive training data are harder to interpret. Traditional explanation methods like SHAP values (Lundberg and Lee, 2017) become less practical for these large-scale models (Zhao et al., 2023). Moreover, a comprehensive understanding of LLM-specific phenomena, including in-context learning (Halawi et al., 2023; Hendel et al., 2023; Todd et al., 2023; Wang et al., 2023), along with addressing issues like model hallucinations (Ji et al., 2023; Chuang et al., 2023) and inherent biases (dev, 2023; An and Rudinger, 2023; Schick et al., 2021), is vital for the ongoing refinement in model design.

In this literature review, we focus on explainability methods for pre-trained Transformer-based LLMs, often termed as *base models*. These models often scale up in training data and have billions of parameters, examples include GPT-2 (Radford et al., 2019), GPT-J (Chen et al., 2021), GPT-3 (Brown et al., 2020), OPT (Yordanov et al., 2022), and LLaMA family (Touvron et al., 2023). In Section 2, we categorize and pose research questions based on our literature review. Based on this categorization, we review explainability methods in Section 3, followed by a discussion in Section 4 on how these insights are leveraged. We further discuss the evaluation methods and metrics in Section 5. Our goal is to synthesize and critically assess contemporary research, aiming to bridge the gap between understanding and practical application of insights derived from complex language models.

## 2 Overview

The field of large language models LLMs is rapidly advancing, making explainability not only a tool for understanding these complex systems but also essential for their improvement. This section categorizes current explainability approaches, highlights the challenges in ethical and controllable generation, and proposes research questions for future exploration.

**Categorization of Methods** We present a structured categorization for the explainability methods and their applications in Figure 1. Figure 1 presents a structured categorization of explainability methods for pre-trained language models (LMs). We divide these into two broad domains: Local Analysis and Global Analysis. Local Analysis covers feature attribution and transformer block analysis, delving into detailed operations of models. Global Analysis, on the other hand, includes probing-based methods and mechanistic interpretability, offering a comprehensive understanding of model behaviors and capacities. Beyond understanding, we also explore applications of these insights in enhancing LLM capabilities, focusing on model editing, capability enhancement, and controlled generation.

## 3 Explainability for Large Language Models

### 3.1 Local Analysis

Local explanations in LLMs aim to elucidate how models generate specific predictions, such as sentiment classification or token predictions, for a given input. This section categorizes local explanation methods into two types: feature attribution analysis, and the analysis into individual Transformer (Vaswani et al., 2017) components.

#### 3.1.1 Feature Attribution Analysis

Feature attribution analysis quantifies the relevance of each input token to a model’s prediction. Given an input text  $x$  with  $n$  tokens  $\{x_1, x_2, \dots, x_n\}$ , a pre-trained language model  $f$  outputs  $f(x)$ . Attribution methods assign a relevance score  $R(x_i)$  (Modarressi et al., 2022; Ferrando et al., 2022; Modarressi et al., 2023) to each token  $x_i$ , reflecting its contribution to  $f(x)$ . This category includes perturbation-based, gradient-based, and vector-based methods.

**Perturbation-Based Methods.** Perturbation-based methods, such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), alter

input features to observe changes in model output. However, this removal strategy assumes input features are independent and ignores correlations among them. Additionally, models can be over-confidence even when the predictions are nonsensical or wrong (Feng et al., 2018). They also face challenges in efficiency and reliability highlighted in (Atanasova et al., 2020), leading to their diminished emphasis in recent attribution research.

**Gradient-Based Methods.** One might consider gradient-based explanation methods as a natural approach for feature attribution. This type of method compute per-token importance scores (Kindermans et al., 2016) using backward gradient vectors. Techniques like gradient  $\times$  input (Kindermans et al., 2017) and integrated gradients (IG) (Sundararajan et al., 2017) accumulate the gradients obtained as the input is interpolated between a reference point and the actual input. Despite their widespread use, one main challenge of IG is the computational overhead to achieve high-quality integrals (Sikdar et al., 2021; Enguehard, 2023). Their attribution score has also shown to be unreliable in terms of faithfulness (Ferrando et al., 2022) and their ability to elucidate the forward dynamics within hidden states remains constrained.

**Vector-Based Methods.** Vector-based analyses, which focus on token representation formation, have emerged as a key approach. Approaches range from global attribution from the final output layer to more granular, layer-wise decomposition of token representations (Chen et al., 2020; Modarressi et al., 2022). Consider decomposing the  $i^{th}$  token representation in layer  $l \in \{0, 1, 2, \dots, L, L + 1\}^1$ , i.e.,  $x_i^l \in \{x_1^l, x_2^l, \dots, x_N^l\}$ , into elemental vectors attributable to each of the  $N$  input tokens:

$$x_i^l = \sum_{k=1}^N x_{i \leftarrow k}^l \quad (1)$$

The norm (Modarressi et al., 2022) or the L1 norm (Ferrando et al., 2022) of the attribution vector for the  $k^{th}$  input ( $x_{i \leftarrow k}^l$ ) can be used to quantify its total attribution to  $x_i^l$ .

Although several established strategies, such as attention rollouts (Abnar and Zuidema, 2020; Ferrando et al., 2022; Modarressi et al., 2022), focus

<sup>1</sup> $l = 0$  is the input embedding layer and  $l = L + 1$  is the language model head over the last decoder layer.

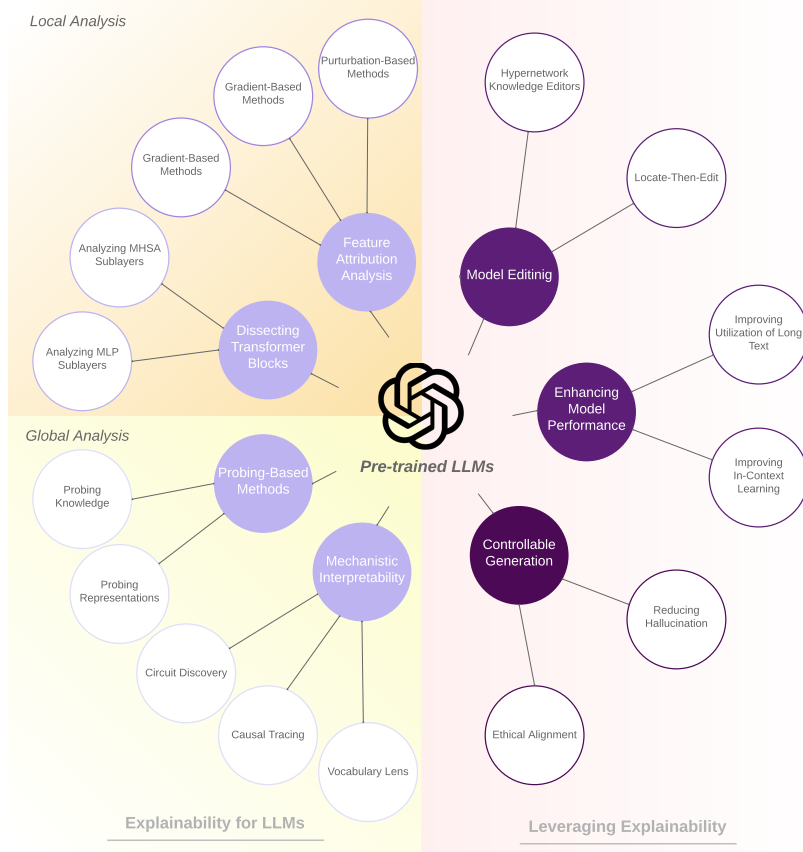


Figure 1: Categorization of literature on explainability in LLMs, focusing on techniques (Left, Section 3) and their applications (Right, Section 4).

on the global impact of inputs on outputs by aggregating the local behaviors of all layers, they often neglect Feed-Forward Network (FFN) in the analyses due to its nonlinearities. Recent works address this by approximating and decomposing activation functions and constructing decomposed token representations throughout layers (Yang et al., 2023; Modarressi et al., 2023). Empirical evaluations demonstrate the efficacy of vector-based analysis and exemplify the potential of such methods in dissecting each hidden state representation within transformers.

### 3.1.2 Dissecting Transformer Blocks

Tracking Transformer block’s *component-by-component internal processing* can provide rich information about their intermediate processing, given the stacked architecture of decoder-based language models (Kobayashi et al., 2023). In a transformer inference pass, the input embeddings are transformed through a sequence of  $L$  transformer layers, each composed of a multi-head self-attention (MHSA) sublayer followed by an MLP sublayer (Vaswani et al., 2017). Formally, the rep-

resentation  $x_i^l$  of token  $i$  at layer  $l$  is obtained by:

$$x_i^l = x_i^{l-1} + a_i^l + m_i^l \quad (2)$$

where  $a_i^l$  and  $m_i^l$  are the outputs from the  $l$ -th MHSA and MLP sublayers, respectively<sup>2</sup>. While studies have frequently analyzed individual Transformer components (Kobayashi et al., 2020; Modarressi et al., 2022), the interaction between these sublayers is less explored, presenting an avenue for future research.

**Analyzing MHSA Sublayers.** Attention mechanisms in MHSA sublayers are instrumental in capturing meaningful correlations between intermediate states of input that can explain the model’s predictions. Visualizing attention weights and utilizing gradient attribution scores are two primary methods for analyzing these sublayers (Zhao et al., 2023). Many studies have analyzed the linguistic capabilities of Transformers by tracking attention

<sup>2</sup>For brevity, bias terms and layer normalization (Ba et al., 2016) are omitted, as they are nonessential for most of analysis.

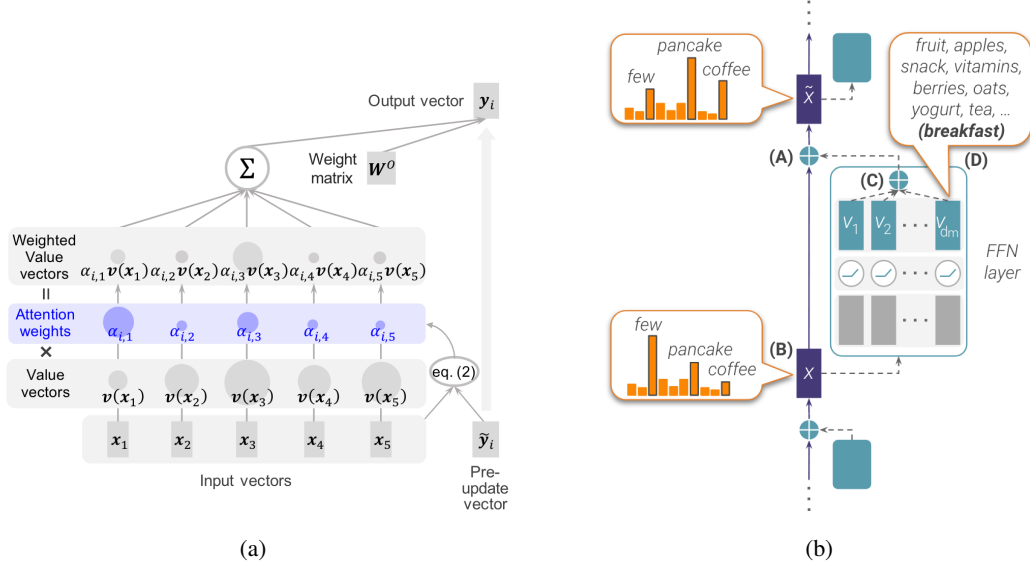


Figure 2: Studied role of each Transformer component. (a) gives an overview of attention mechanism in Transformers. Sizes of the colored circles illustrate the value of the scalar or the norm of the corresponding vector (Kobayashi et al., 2020). (b) analyzes the FFN updates in the vocabulary space, showing that each update can be decomposed to sub-updates corresponding to single FFN parameter vectors, each promoting concepts that are often human-interpretable (Geva et al., 2022).

weights. (Abnar and Zuidema, 2020; Katz and Belinkov, 2023; Kobayashi et al., 2023). For instance, attention mechanisms typically prioritize specific tokens while diminishing the emphasis on frequent words or special tokens, a phenomenon observable through norm-based analysis metrics, as illustrated in Figure 2(a) (Kobayashi et al., 2020). In the gradient analysis, some methods calculate gradients as partial derivatives of model outputs with respect to attention weights (Barkan et al., 2021), while others use integrated gradients, which are cumulative versions of these partial derivatives (Hao et al., 2021). Generally, these combined approaches, which integrate attention metrics with gradient information, tend to outperform methods using either metric in isolation.

**Analyzing MLP Sublayers.** More recently, a surge of works have investigated the knowledge captured by the FFN layers (Geva et al., 2022; Dai et al., 2022). These layers, consuming the majority of each layer’s parameter budget at  $8d^2$  compared to  $4d^2$  for self-attention layers (where  $d$  represents the model’s hidden dimension), function akin to key-value memories (Geva et al., 2021). Here, each "key" is associated with specific textual patterns identified during training, and each "value" generates a corresponding output vocabulary distribution (Geva et al., 2021). Figure 2(b) focuses in the FFN

outputs, illustrating how each update within these layers can be broken down into sub-updates linked to individual parameter vectors, often encoding concepts that are interpretable to humans (Geva et al., 2022). Additionally, there is an emerging interest in input-independent methods, which interpret model parameters directly, thus eliminating the need for a forward pass (Dar et al., 2023).

### 3.2 Global Analysis

In contrast to local analysis, which focus on elucidating individual model predictions, global analysis aims to understand and explain the knowledge or linguistic properties encoded in the hidden state activations of a model. This section explores two primary approaches to global analysis: probing methods that scrutinize model representations and mechanistic interpretability (Transformer Circuits, 2022), an emerging perspective that seeks to reverse engineer the inner workings of deep neural networks.

#### 3.2.1 Probing-Based Method

Self-supervised pre-training endows models with extensive linguistic knowledge, derived from large-scale training datasets. Probing-based methods are employed to capture the internal representations within these networks. This approach involves training a classifier, known as a probe, on the net-



work’s activations to distinguish between various types of inputs or outputs. In the following sections, we will discuss studies related to probing, categorized based on their objectives, whether it be probing for semantic knowledge or analyzing learned representations.

**Probing Knowledge.** LLMs trained on extensive text corpora, are recognized for their ability to encapsulate context-independent semantic and factual knowledge accessible via textual prompts (Petroni et al., 2019). Research in this area primarily focuses on formulating textual queries to extract various types of background knowledge from language models (Hewitt and Manning, 2019; Peng et al., 2022). Interestingly, probes can sometimes unearth factual information even in scenarios where language models may not reliably produce truthful outputs (Hernandez et al., 2023).

**Probing Representations.** LLMs are adept at developing context-dependent knowledge representations. To analyze these, probing classifiers are applied, typically involving a shallow classifier trained on the activations of attention heads to predict specific features. A notable study in this area involved training linear classifiers to identify a select group of attention heads that exhibit high linear probing accuracy for truthfulness (Li et al., 2023b). This research revealed a pattern of specialization across attention heads, with the representation of “truthfulness” predominantly processed in the early to middle layers, and only a few heads in each layer showing standout performance. Such insights pave the way for exploring more complex representations. For instance, research by (Li et al., 2023a) has revealed nonlinear internal representations, such as board game states, in models that initially lack explicit knowledge of the game or its rules.

### 3.2.2 Mechanistic Interpretability

Mechanistic interpretability seeks to comprehend language models by examining individual neurons and their interconnections, often conceptualized as circuits (Transformer Circuits, 2022; Zhao et al., 2023). This field encompasses various approaches, which can be primarily categorized into three groups: circuit discovery, causal tracing, and vocabulary lens. Each of these approaches offers distinct perspectives and insights into the mechanisms of language models.

**Circuit Discovery.** The circuit-based mechanistic interpretability approach aims to align learned model representations with known ground truths, initially by reverse-engineering the model’s algorithm to fully comprehend its feature set (Chughtai et al., 2023). A prominent example of this approach is the analysis of GPT-2 small (Radford et al., 2019), where a study identified a human-understandable subgraph within the computational graph responsible for performing the indirect object identification (IOI) task (Wang et al., 2022). In IOI, sentences like “When Mary and John went to the store, John gave a drink” are expected to be completed with “Mary”. The study discovered a circuit comprising 26 attention heads – just 1.1% of the total (head, token position) pairs – that predominantly manages this task. This circuit-based mechanistic view provides opportunities to scale our understanding to both larger models and more complex tasks, including recent explorations into In-Context Learning (ICL) (Halawi et al., 2023; Hendel et al., 2023; Todd et al., 2023; Wang et al., 2023).

**Causal Tracing.** The concept of causal analysis in machine learning has evolved from early methods that delineate dependencies between hidden variables using causal graphs (Pearl et al., 2000) to more recent approaches like causal mediation analysis (Vig et al., 2020). This newer method quantifies the impact of intermediate activations in neural networks on their output (Meng et al., 2023a). Specifically, (Meng et al., 2023a) assesses each activation’s contribution to accurate factual predictions through three distinct operational phases: a **clean** run generating correct predictions, a **corrupted** run where predictions are impaired, and a **corrupted-with-restoration** run that evaluates the ability of a single state to rectify the prediction (Meng et al., 2023a). Termed as *causal tracing*, this approach has identified crucial causal states predominantly in the middle layers, particularly at the last subject position where MLP contributions are most significant (Figure 3). This finding underscores the role of middle layer MLPs in factual recall within LLMs.

**Vocabulary Lens.** Recent work has suggested that model knowledge and knowledge retrieval may be localized within small parts of a language model (Geva et al., 2021) by projecting weights and hidden states onto their vocabulary space. To analyze

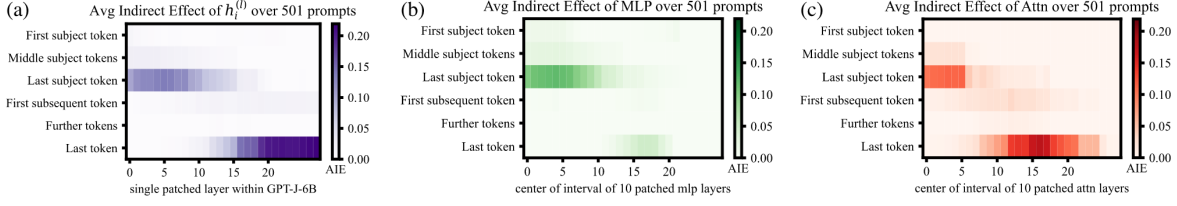


Figure 3: The intensity of each grid cell represents the average causal indirect effect of a hidden state on expressing a factual association. Darker cells indicate stronger causal mediators. It was found that MLPs at the last subject token and attention modules at the last token play crucial roles. (Meng et al., 2023a)

the components in vocabulary space, we read from each token component  $x_k^l$  at layer  $l$  at the last token position  $N$  ( $N$  is omitted here), by projecting with the unembedding matrix  $E$ :

$$p_k^l = \text{softmax}(E \ln(x_k^l)) \quad (3)$$

where  $\ln$  stands for layer normalization before the LM head. (Belrose et al., 2023) refines model predictions at each transformer layer and decodes hidden states into vocabulary distributions based on this method. Exploring this avenue further, (Geva et al., 2022) illuminated the role of transformer feed-forward layers in predictions, spotlighting specific conceptual emphases via FFN sub-updates. There is also a growing interest in input-independent methodologies, where model parameters are interpreted directly, bypassing a forward pass (Dar et al., 2023).

Augmenting projection-focused interpretations, (Din et al., 2023) first unveiled a feasible application for such projections, suggesting early exit strategies by treating hidden state representations as final outputs. (Geva et al., 2023) pinpointed two critical junctures where information propagates to the final predictions via projections and attention edge intervention. While much of the focus has been on how hidden states relate to model outputs, recent works have also highlighted the roles of individual tokens, revealing that their contributions through attention outputs are laden with rich semantic information (Ram et al., 2023; Katz and Belinkov, 2023).

## 4 Leveraging Explainability

In this section, we discuss how explainability can be used as a tool to debug and improve models. Although various approaches aim to improve model capabilities with fine-tuning or re-training, we focus on methods specifically designed with a strong foundation in model explainability.

### 4.1 Model Editing

Despite the ability to train proficient LLMs, the methodology for ensuring their relevance and rectifying errors remains elusive. In recent years, there has been a surge in techniques for editing LLMs. The goal is to efficiently modify the knowledge or behavior of LLMs within specific domains without adversely affecting their performance on other inputs (Yao et al., 2023).

**Hypernetwork Knowledge Editors.** This type of knowledge editors include *memory-based model* and editors with *additional parameters*. Memory-based models store all edit examples explicitly in memory based on the explainability finding of key-value memories inside the FFN (Section 3.1.2). They can then employ a retriever to extract the most relevant edit facts for each new input, guiding the model to generate the edited fact. SERAC (Mitchell et al., 2022), for instance, adopts a distinct counterfactual model while leaving the original model unchanged. Editors with additional parameters introduce extra trainable parameters within LLMs. These parameters are trained on a modified dataset while the original model parameters remain static. For example, T-Patcher (Huang et al., 2023) integrates one neuron (patch) for one mistake in the last layer of the FFN of the model, which takes effect only when encountering its corresponding mistake.

**Locate-Then-Edit.** The locate-then-edit paradigm first identifies parameters corresponding to specific knowledge and then modifies them through direct updates to the target parameters. The Knowledge Neuron (KN) method (Dai et al., 2022) introduces a knowledge attribution technique to pinpoint the “knowledge neuron” (a key-value pair in the FFN matrix) that embodies the knowledge and then updates these neurons. ROME (Meng et al., 2023a) and MEMIT (Meng

et al., 2023b) apply causal tracing (Section 3.2.2) to locate the editing area. Instead of modifying the knowledge neurons in the FFN, ROME alters the entire matrix. Building upon these two methods, PMET (Li et al., 2023c) involves the attention value to achieve better performance.

## 4.2 Enhancing Model Capability

While LLMs demonstrate versatility in various NLP tasks, insights from explainability can significantly enhance these capabilities. This section highlights two key tasks where explainability has shown considerable impact in recent work: improving the utilization of long text (Xiao et al., 2023; Liu et al., 2023; Pope et al., 2022) and enhancing the performance of In-Context Learning (ICL) (Hendel et al., 2023; Halawi et al., 2023; Wang et al., 2023).

### 4.2.1 Improving Utilization of Long Text

The optimization of handling long text aims to enhance the ability of LLMs to capture and effectively utilize content within longer contexts. This is particularly challenging because LLMs tend to struggle with generalizing to sequence lengths longer than what they were pretrained on, such as the 4K limit for Llama-2 (Touvron et al., 2023). (Beltagy et al., 2020), maintains a fixed-size sliding window on the key-value (KV) states of the most recent tokens. While this approach ensures constant memory usage and decoding speed after the cache is initially filled, it faces limitations when the sequence length exceeds the cache size (Liu et al., 2023). An innovative solution proposed by (Xiao et al., 2023) takes advantage of the MHSA explanations (Section 3.1.2) in LLMs, which allocates a significant amount of attention to the initial tokens. They introduce StreamingLLM, a simple and efficient framework that allows LLMs to handle unlimited text without fine-tuning. This is achieved by retaining the "attention sink," which consists of several initial tokens, in the KV states (Figure 4). The authors also demonstrate that pre-training models with a dedicated sink token can further improve streaming performance.

### 4.2.2 Improving In-Context Learning

In-context Learning (ICL) has emerged as a powerful capability alongside the development of scaled-up LLMs (Brown et al., 2020). ICL stands out because it doesn't require extensive updates to the vast number of model parameters and relies on

human-understandable natural language instructions (Dong et al., 2023). As a result, it offers a promising approach to harness the full potential of LLMs. With mechanistic interpretability (Section 3.2.2), (Wang et al., 2023) reveal that label words in the demonstration examples function as anchors, which can be used to improve ICL performance with simple anchor re-weighting method. (Halawi et al., 2023) study harmful imitation in ICL through vocabulary lens to inspect a model's internal representations (Section 3.2.2), and identify two related phenomena: *overthinking* and *false induction heads*, the heads in late layers that attend to and copy false information from previous demonstrations, and whose ablation improves ICL performance. Furthermore, using causal tracing (Section 3.2.2), (Hendel et al., 2023; Todd et al., 2023) find that a small number attention heads transport a compact representation of the demonstrated task, which they call a *task vector* or *function vector* (FV). These FVs can be summed to create vectors that trigger new complex tasks and improve performance for few-shot prompting (Todd et al., 2023).

## 4.3 Controllable Generation

Though large language models have obtained superior performance in text generation, they sometimes fall short of producing factual content. Leveraging explainability provides opportunities for building inference-time and fast techniques to improve generation models' factuality, calibration, and controllability and align more with human preference.

### 4.3.1 Reducing Hallucination

Hallucinations in LLMs refer to generated content not based on training data or facts, various factors like imperfect learning and decoding contribute to this (Ji et al., 2023). To mitigate hallucinations, initial approaches used reinforcement learning from human feedback (Ouyang et al., 2022) and distillation into smaller models like Alpaca (Li et al., 2023d). Leveraging explainability provides a significantly less expensive way to reduce hallucination, enjoying the advantage of being adjustable and minimally invasive. For example, (Li et al., 2023b) use as few as 40 samples to locate and find "truthful" heads and directions through a trained probe (Section 3.2.1). They propose inference-time intervention (ITI), a computationally inexpensive strategy to intervene on the attention head to shift the activations in the "truthful" direction, which achieves comparable or better performance towards

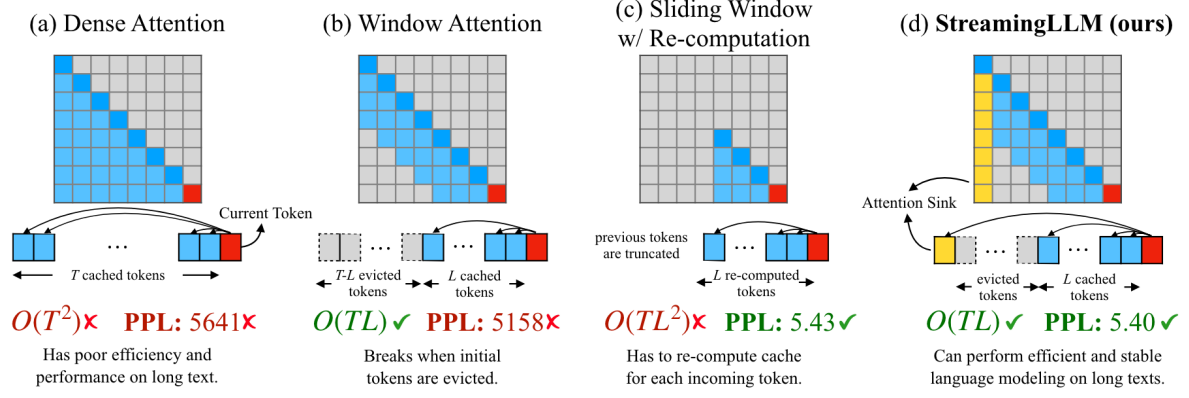


Figure 4: (a) Dense Attention (Vaswani et al., 2017) has  $O(T^2)$  time complexity and an increasing cache size. Its performance decreases when the text length exceeds the pre-training text length. (b) Window Attention () caches the most recent  $L$  tokens’ KV. While efficient in inference, performance declines sharply once the starting tokens’ keys and values are evicted. (c) Sliding Window (Pope et al., 2022) with Re-computation performs well on long texts, but its  $O(TL^2)$  complexity, stemming from quadratic attention in context re-computation, makes it considerably slow. (d) StreamingLLM keeps (Xiao et al., 2023) the attention sink (several initial tokens) for stable attention computation, combined with the recent tokens. It’s efficient and offers stable performance on extended texts.

instruction-finetuned model.

### 4.3.2 Ethical Alignment

As research on AI fairness gains increasing importance, there have been efforts to detect social bias (Fleisig et al., 2023; An and Rudinger, 2023) and suppress toxicity (Gehman et al., 2020; Schick et al., 2021) in LMs. Many previous debiasing methods (Qian et al., 2022) have focused on constructing anti-stereotypical datasets and then either retraining the LM from scratch or conducting fine-tuning. This line of debiasing approaches, although effective, comes with high costs for data construction and model retraining. Moreover, it faces the challenge of catastrophic forgetting if fine-tuning is performed (Zhao et al., 2023). While few work has focused on the interpretability of the fairness research, (dev, 2023) explore interpreting and mitigating social biases in LLMs by introducing the concept of *social bias neurons*. Inspired by the gradient-based attribution method IG (Section 3.1.1), they introduce an interpretable technique, denoted as integrated gap gradient ( $IG^2$ ), to pinpoint social bias neurons by back-propagating and integrating the gradients of the logits gap for a selected pair of demographics.<sup>3</sup> Derived from this interpretation, they suppress the pinpointed neuron activations to mitigate the bias. Extensive experiments have verified the effectiveness of this

<sup>3</sup>Demographic include properties like gender, sexuality, occupation, etc. 9 common demographics are collected and pairs of demographics are selected to reveal the fairness gap (dev, 2023).

method and have yielded the potential applicability of the explainability method for ethical alignment research in LLMs.

## 5 Evaluation

Recently, LLMs such as GPT-4 (OpenAI, 2023) have shown impressive abilities to generate natural language explanations for their predictions. However, it remains unclear whether these explanations actually assist humans in understanding the model’s reasoning (Zhao et al., 2023). Specifically designed evaluation methods are needed to better assess the performance of explainability methods, such as attribution. Furthermore, calibrated datasets and metrics are required to evaluate the application of explainability to downstream tasks, such as truthfulness evaluation<sup>4</sup>.

### 5.1 Evaluating Explanation Plausibility

One common technique to evaluate the plausibility of attribution explanation is to remove  $K\%$  of tokens with the highest or lowest estimated importance to observe its impact on the model’s output (Chen et al., 2020; Modarressi et al., 2023). Another approach to assessing explanation plausibility involves indirect methods, such as measuring the performance of model editing, particularly for “locate-then-edit” editing methods, which heavily rely on interpretation accuracy. Recent research (Yao et al., 2023; Zhao et al., 2023) suggests that

<sup>4</sup>Due to space limit, we only discuss the most commonly used evaluation approaches in explainability research



having evaluation datasets is crucial for evaluating factual editing in LLMs. Two commonly used datasets for this purpose are ZsRE (Levy et al., 2017), a Question Answering (QA) dataset that employs question rephrasings generated through back-translation, and CounterFact (Meng et al., 2023a), a more challenging dataset that includes counterfactuals starting with low scores in comparison to correct facts.

## 5.2 Evaluating Truthfulness

Model truthfulness is an important metric for measuring the trustworthiness of generative models. We expect model outputs to be both informative and factually correct and faithful. Ideally, human annotators would label model answers as true or false, given a gold standard answer, but this is often costly. (Lin et al., 2022) propose the use of two fine-tuned GPT-3-13B models (GPT-judge) to classify each answer as true or false and informative or not. Evaluation using GPT-judge is a standard practice on TruthfulQA benchmark, a widely used dataset adversarially constructed to measure whether a language model is truthful in generating answers (Askell et al., 2021; Li et al., 2023b; Chuang et al., 2023). The main metric of TruthfulQA is **true\*informative**, a product of scalar truthful and informative scores. This metric not only captures how many questions are answered truthfully but also prevents the model from indiscriminately replying with “I have no comment” by assessing the informativeness of each answer.

## 6 Conclusion

In this paper, we have presented a comprehensive overview on explainability for LLMs and their applications. We have summarized methods for local and global analysis based on the objectives of explanations. Additionally, we have discussed the use of explanations to enhance models and the evaluation of these methods. Major future research directions in understanding LLMs include developing explanation methods tailored to different language models and making LLMs more trustworthy and aligned with human values by leveraging explainability knowledge. As LLMs continue to advance, explainability will become incredibly vital to ensure that these models are transparent, fair, and beneficial. We hope that this literature review provides a useful overview of this emerging research area and highlights open problems and directions

for future research.

## References

2023. [The devil is in the neurons: Interpreting and mitigating social biases in language models](#). In *Openreview for International Conference on Learning Representations 2024*.
- Samira Abnar and Willem Zuidema. 2020. [Quantifying attention flow in transformers](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.
- Haozhe An and Rachel Rudinger. 2023. [Nichelle and nancy: The influence of demographic attributes and tokenization length on first name biases](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 388–401, Toronto, Canada. Association for Computational Linguistics.
- Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. [A general language assistant as a laboratory for alignment](#).
- Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. [A diagnostic study of explainability techniques for text classification](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.
- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. [Layer normalization](#).
- Oren Barkan, Edan Hauon, Avi Caciularu, Ori Katz, Itzik Malkiel, Omri Armstrong, and Noam Koenigstein. 2021. [Grad-sam: Explaining transformers via gradient self-attention maps](#). In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management, CIKM '21*, page 2882–2887, New York, NY, USA. Association for Computing Machinery.
- Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. 2022. ["will you find these shortcuts?" a protocol for evaluating the faithfulness of input salience methods for text classification](#).
- Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. [Eliciting latent predictions from transformers with the tuned lens](#).

- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. [Longformer: The long-document transformer](#).
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#).
- Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. 2020. [Generating hierarchical explanations on text classification via feature interaction detection](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5578–5593, Online. Association for Computational Linguistics.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. [Evaluating large language models trained on code](#).
- Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. [Dola: Decoding by contrasting layers improves factuality in large language models](#).
- Bilal Chughtai, Lawrence Chan, and Neel Nanda. 2023. [A toy model of universality: Reverse engineering how networks learn group operations](#).
- Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. [Knowledge neurons in pretrained transformers](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.
- Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. 2023. [Analyzing transformers in embedding space](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16124–16170, Toronto, Canada. Association for Computational Linguistics.
- Alexander Yom Din, Taelin Karidi, Leshem Choshen, and Mor Geva. 2023. [Jump to conclusions: Short-cutting transformers with linear transformations](#).
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. [A survey on in-context learning](#).
- Joseph Enguehard. 2023. [Sequential integrated gradients: a simple but effective method for explaining language models](#).
- Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. [Pathologies of neural models make interpretations difficult](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.
- Javier Ferrando, Gerard I. Gállego, and Marta R. Costajussà. 2022. [Measuring the mixing of contextual information in the transformer](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8698–8714, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Eve Fleisig, Aubrie Amstutz, Chad Atalla, Su Lin Blodgett, Hal Daumé III, Alexandra Olteanu, Emily Sheng, Dan Vann, and Hanna Wallach. 2023. [Fair-Prism: Evaluating fairness-related harms in text generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6231–6251, Toronto, Canada. Association for Computational Linguistics.
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. [RealToxicityPrompts: Evaluating neural toxic degeneration in language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.
- Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. [Dissecting recall of factual associations in auto-regressive language models](#).
- Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. [Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

- Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. 2023. [Overthinking the truth: Understanding how language models process false demonstrations](#).
- Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. [Self-attention attribution: Interpreting information interactions inside transformer](#).
- Roei Hendel, Mor Geva, and Amir Globerson. 2023. [In-context learning creates task vectors](#).
- Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2023. [Inspecting and editing knowledge representations in language models](#).
- John Hewitt and Christopher D. Manning. 2019. [A structural probe for finding syntax in word representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.
- Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. 2023. [Transformer-patcher: One mistake worth one neuron](#).
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Computing Surveys*, 55(12):1–38.
- Shahar Katz and Yonatan Belinkov. 2023. [Interpreting transformer’s attention dynamic memory and visualizing the semantic information flow of gpt](#).
- Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2017. [The \(un\)reliability of saliency methods](#).
- Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. 2016. [Investigating the influence of noise and distractors on the interpretation of neural networks](#).
- Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. [Attention is not only a weight: Analyzing transformers with vector norms](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.
- Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2023. [Analyzing feed-forward blocks in transformers through the lens of attention map](#).
- Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. [Zero-shot relation extraction via reading comprehension](#). In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.
- Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. [Emergent world representations: Exploring a sequence model trained on a synthetic task](#).
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. [Inference-time intervention: Eliciting truthful answers from a language model](#).
- Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. 2023c. [Pmet: Precise model editing in a transformer](#). *ArXiv*, abs/2308.08742.
- Zhihui Li, Max Gronke, and Charles Steidel. 2023d. [Alpaca: A new semi-analytic model for metal absorption lines emerging from clumpy galactic environments](#).
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [Truthfulqa: Measuring how models mimic human falsehoods](#).
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranajape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. [Lost in the middle: How language models use long contexts](#).
- Scott Lundberg and Su-In Lee. 2017. [A unified approach to interpreting model predictions](#).
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023a. [Locating and editing factual associations in gpt](#).
- Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023b. [Mass-editing memory in a transformer](#).
- Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022. [Memory-based model editing at scale](#).
- Ali Modarressi, Mohsen Fayyaz, Ehsan Aghazadeh, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2023. [DecompX: Explaining transformers decisions by propagating token decomposition](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2649–2664, Toronto, Canada. Association for Computational Linguistics.
- Ali Modarressi, Mohsen Fayyaz, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2022. [GlobEnc: Quantifying global token attribution by incorporating the whole encoder layer in transformers](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 258–271, Seattle, United States. Association for Computational Linguistics.
- OpenAI. 2023. [Gpt-4 technical report](#).



- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#).
- Judea Pearl et al. 2000. Models, reasoning and inference. *Cambridge, UK: Cambridge University Press*, 19(2):3.
- Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu, and Qun Liu. 2022. [Copen: Probing conceptual knowledge in pre-trained language models](#).
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. [Language models as knowledge bases?](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.
- Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Anselm Levskaya, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. 2022. [Efficiently scaling transformer inference](#).
- Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. 2022. [Perturbation augmentation for fairer NLP](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9496–9521, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Ori Ram, Liat Bezael, Adi Zicher, Yonatan Belinkov, Jonathan Berant, and Amir Globerson. 2023. [What are you token about? dense retrieval as distributions over the vocabulary](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2481–2498, Toronto, Canada. Association for Computational Linguistics.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier.
- Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. [Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp](#).
- Sandipan Sikdar, Parantapa Bhattacharya, and Kieran Heese. 2021. [Integrated directional gradients: Feature interaction attribution for neural NLP models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 865–878, Online. Association for Computational Linguistics.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. [Axiomatic attribution for deep networks](#). In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.
- Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2023. [Function vectors in large language models](#).
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).
- Transformer Circuits. 2022. [Mechanistic interpretations of transformer circuits](#). Accessed: [insert access date here].
- Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *NIPS*.
- Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. [Causal mediation analysis for interpreting neural nlp: The case of gender bias](#).
- 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](#).
- Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. [Label](#)



words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9840–9855, Singapore. Association for Computational Linguistics.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. [Ethical and social risks of harm from language models](#).

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. [Efficient streaming language models with attention sinks](#).

Sen Yang, Shujian Huang, Wei Zou, Jianbing Zhang, Xinyu Dai, and Jiajun Chen. 2023. [Local interpretation of transformer based on linear decomposition](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10270–10287, Toronto, Canada. Association for Computational Linguistics.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. [Editing large language models: Problems, methods, and opportunities](#).

Yordan Yordanov, Vid Kocijan, Thomas Lukasiewicz, and Oana-Maria Camburu. 2022. [Few-shot out-of-domain transfer learning of natural language explanations in a label-abundant setup](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3486–3501, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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](#).