International Journal of Computing and Artificial Intelligence

E-ISSN: 2707-658X
P-ISSN: 2707-6571
IJCAI 2022; 3(2): 102-108
Received: 22-06-2022
Accepted: 25-07-2022
www.computersciencejournals.com/ijcai

Rahul Manche
CVR College of Engineering, Hyderabad, Telangana, India

Praveen Kumar Myakala
Independent Researcher, Texas, United States of America

Corresponding Author:
Rahul Manche
CVR College of Engineering, Hyderabad, Telangana, India

Explaining black-box behavior in large language models

Rahul Manche and Praveen Kumar Myakala

DOI: https://doi.org/10.33545/27076571.2022.v3.i2a.126

Abstract
Large language models (LLMs), such as GPT and BERT, have transformed natural language processing (NLP), achieving state-of-the-art performance across diverse tasks. However, their opaque decision-making processes raise critical concerns about transparency, trust, and accountability, particularly in high-stakes domains. This paper reviews key interpretability techniques, including attention visualization, layer wise relevance propagation, feature attribution, and counterfactual analysis, and proposes a comprehensive taxonomy for explaining LLM behavior. Additionally, we address the ethical implications of explainability, such as bias mitigation, privacy preservation, and regulatory compliance. By synthesizing current methods and identifying challenges, this work lays the foundation for future research into interpretable and ethically responsible LLMs.

Keywords: Explainable AI, large language models, transparency, transformers, attention mechanism, interpretability, neural networks

Introduction
Large Language Models (LLMs), such as GPT, BERT, and T5, have become foundational in natural language processing (NLP). These models, trained on vast datasets, achieve state-of-the-art performance in tasks such as translation, summarization, and question answering. Despite their impressive capabilities, the decision-making processes of LLMs remain largely opaque, raising critical concerns about their trustworthiness and reliability.

The Need for Explainability
The opacity of LLMs stems from their complex architectures, which involve billions of parameters [1] and layers of abstraction. As these models increasingly influence high stakes domains, explainability becomes crucial:

Transparency for Trust: Users need to trust AI outputs, particularly in applications like healthcare, where opaque models can lead to life-critical consequences [2].

Debugging and Optimization: Developers require interpretability tools to diagnose and rectify biases or errors in model outputs [3].

Ethical and Legal Compliance: Frameworks such as GDPR mandate transparency in algorithmic decisions, making explainability a regulatory requirement [4].

Scope of This Paper
This paper reviews existing methods for explaining LLMs, categorizes interpretability techniques, and identifies challenges. Emphasis is placed on techniques like attention visualization [5], counterfactual explanations [4], and feature attribution [6].
Additionally, we explore the ethical implications of explainability in AI.

Technical Foundations of Large Language Models
Large language models (LLMs) are mainly built on the transformer architecture, which revolutionized natural language processing (NLP) through its ability to handle sequential data efficiently [7]. This section outlines the key technical components that underpin LLM, including self-attention mechanisms, positional encoding, and embeddings.

Transformer Architecture

The transformer architecture, introduced by Vaswani *et al.* [7], relies on self-attention mechanisms to compute contextual representations of tokens in an input sequence. Unlike recurrent neural networks (RNNs), transformers enable parallel processing of sequences, addressing issues of scalability and long-term dependency modeling.

The core operation in a transformer layer is the self-attention mechanism, mathematically defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V,$$

Where

***Q*:** Query matrix, representing the input sequence projected into query space.

***K*:** Key matrix, representing the input sequence projected into key space.

***V*:** Value matrix, representing the input sequence projected into value space.

***d_k*:** Dimensionality of the key vectors, used for scaling.

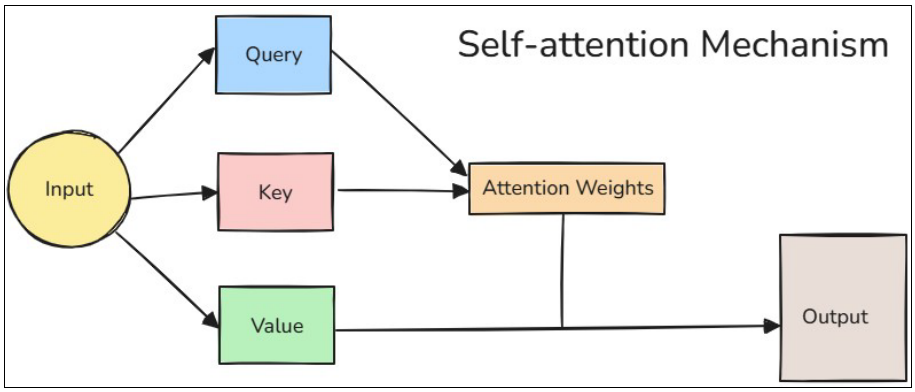

Fig 1: self-attention mechanism layer

Positional Encoding

Transformers lack inherent awareness of token order due to their nonsequential processing. To encode positional information, transformers add positional encodings to token embeddings. These encodings are defined as:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right), \quad PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right)$$

Where

***pos*:** The position of the token in the input sequence.

***i*:** The dimension index.

***d*:** The dimensionality of the model.

Embeddings and Representations

Each token in the input sequence is first mapped to a dense vector representation using an embedding layer. These token embeddings are combined with positional encodings to form the input representation:

$$x_i = E[w_i] + P[i],$$

Where

***E*[w_i]:** The embedding of the token w_i.

***P*[*i*]:** The positional encoding for the *i*-th token.

Pre-training and Fine-Tuning

LLMs are typically trained in two stages:

Pre-training: The model learns general purpose language representations on massive unlabeled datasets using self-supervised tasks such as masked language modeling (MLM) in BERT [8] or causal language modeling in GPT [1].

Fine-tuning: The pre-trained model is adapted to specific downstream tasks using labeled datasets [9].

The pre training objective for masked language modeling is defined as:

$$L_{\text{MLM}} = -\sum_{i=1}^{n} \log P(w_i | w_{\setminus i}),$$

Where

w_i represents the masked token, and $w_{\setminus i}$ represents the context tokens excluding w_i. The objective for causal language modeling, used in GPT, is defined as:

$$L_{\text{CLM}} = -\sum_{i=1}^{n} \log P(w_i | w_{<i}),$$

Where $w_{<i}$ represents all tokens preceding w_i.

The transformer architecture's ability to handle long sequences efficiently and learn context-rich representations has made it the foundation of state-of-the-art LLMs. The following sections build upon this foundation to explore methods for explaining LLM behavior.

Methods for Explaining Large Language Models

Interpreting the behavior of Large Language Models (LLMs) is a multifaceted challenge. Existing techniques aim to bridge the gap between the models high-dimensional decision-making processes and human understanding. This section categorizes key interpretability methods into attention visualization, layer-wise relevance propagation, counterfactual explanations, and feature attribution.

Attention Visualization

The self-attention mechanism in transformers provides a natural avenue for interpretability. Attention weights can reveal how the model focuses on different tokens when processing an input sequence. For example, in a machine translation task, attention maps can highlight which words in the source language are most relevant to a specific target word [5].

The attention score between tokens *i* and *j* is calculated as:

$$\alpha_{ij} = \frac{\exp\left(Q_i K_j^\top / \sqrt{d_k}\right)}{\sum_{j=1}^{n} \exp\left(Q_i K_j^\top / \sqrt{d_k}\right)}.$$

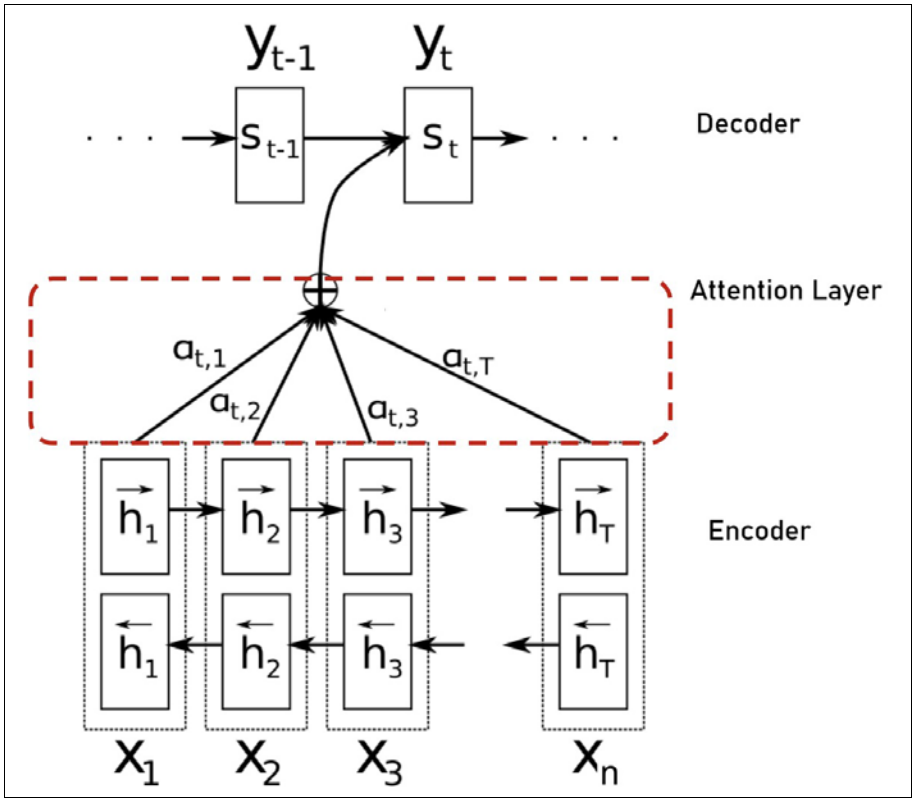

Fig 2: Attention map.

Here

***Qi* and *Kj*:** Query and key vectors for tokens *i* and *j*.

***dk*:** Dimensionality of the key vectors.

While attention maps provide insights into token-to-token dependencies, they have limitations. Studies have shown that attention weights do not always correlate with feature importance, leading to potential misinterpretations [10].

Layer-Wise Relevance Propagation (LRP)

Layer-Wise Relevance Propagation (LRP) assigns relevance scores to input features by back propagating the model's predictions through its layers [11]. This method traces how individual input tokens contribute to the final output.

The relevance score for a neuron *j* in layer *l* is computed as:

$$R_j^{(l)} = \sum_i \frac{z_{ij}^{(l)}}{\sum_k z_{ik}^{(l)}} R_i^{(l+1)}.$$

Here

$z_{ij}^{(l)}$**:** Activation of neuron *j* contributing to neuron *i* in the next layer.

$R_i^{(l+1)}$**:** Relevance score of neurons *i* in the subsequent layer.

LRP is particularly useful in tasks where detailed feature-level explanations are required, such as sentiment analysis or healthcare applications [2].

Counterfactual Explanations

Counterfactual explanations help users understand how small changes to inputs can alter the model's output [4]. These explanations are particularly intuitive, as they answer "what-if" questions.

For instance, consider a text classification model:

- Original input: *"The customer service was poor."*
- Counterfactual: *"The customer service was excellent."*

The optimization objective for generating counterfactuals is defined as:

$$L = \|f(x + \Delta x) - y_{\text{target}}\|^2 + \lambda \|\Delta x\|_1,$$

Where

***x*:** Original input.

Δ*x*: Perturbation applied to the input.

y_{target}**:** Target output.

λ: Regularization term to ensure minimal perturbation.

Feature Attribution Methods
Feature attribution methods estimate the contribution of each input token to the models output. Popular techniques include:

SHAP (Shapley Additive Explanations) [6]: Computes the marginal contribution of each feature using cooperative game theory.

Integrated Gradients [12]: Integrates gradients along a path from a baseline input to the actual input.
For example, SHAP values are computed as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$$

Where
ϕ_i**:** Contribution of feature *i*.
***S*:** Subset of features excluding *i*.
***f*(*S*):** Model output for subset *S*.

Challenges in Explaining Large Language Models
Despite the advancements in interpretability methods, explaining the behavior of Large Language Models (LLMs) remains a complex and unresolved challenge. This section discusses key obstacles, including high dimensionality, context dependency, computational overhead, and ethical concerns.

High Dimensionality
LLMs operate in high-dimensional parameter spaces, often involving billions of weights across multiple layers [13]. This complexity makes it difficult to distill meaningful insights from the model's internal processes.
For instance, interpreting the contribution of individual neurons or attention heads across multiple layers requires reducing high-dimensional information to a comprehensible format. Dimensionality reduction techniques, such as Principal Component Analysis (PCA), can provide partial insights, but they often oversimplify the underlying dynamics [14].

Context Dependency
The interpretation of tokens in LLMs is highly dependent on context. A single word can have vastly different embeddings based on the surrounding text [15]. This context sensitive nature complicates efforts to provide static explanations.
Consider the word *"bank"*:

- In the sentence *"She sat by the river bank"*, *"bank"* refers to a geographical feature.
- In *"He deposited money in the bank"*, it refers to a financial institution.

LLMs generate dynamic embeddings for such tokens, capturing these distinctions. However, explaining how these embeddings shift across layers remains a significant challenge.
The dynamic embedding $E(w|C)$ of a word *w* given its context *C* can be conceptualized as:

$$E(w|C) = \phi(w) + \psi(C),$$

Where
$\phi(w)$**:** The base embedding of the word *w*.
$\psi(C)$**:** The contextual adjustment based on *C*.

Computational Overhead
Interpreting LLMs often requires significant computational resources. Methods like SHAP [6] or Layer-Wise Relevance Propagation (LRP) [11] involve analyzing multiple permutations of the input, scaling poorly with model size.
The computational complexity of SHAP can be expressed as:

$$\mathcal{O}(2^n \cdot f),$$

Where
***n*:** Number of input features.
***f*:** Time to evaluate the model for a single input.
Strategies such as approximation techniques or model compression can help mitigate this overhead, but they often compromise the fidelity of explanations.

Ethical Considerations
Ethical concerns further complicate the development of interpretable LLMs. These include:

Bias Propagation: LLMs often inherit biases from their training data, and these biases may be amplified in explanations [16].

Privacy Risks: Providing detailed explanations may inadvertently expose sensitive data from the training corpus [17].

Manipulative Explanations: Simplified or selective explanations can be used to mislead users, raising concerns about accountability [18].
To address these concerns, fairness constraints can be incorporated into explanation mechanisms. For instance, ensuring fairness in decision-making can be modeled as:

$$\text{Fairness Loss} = \|P(y|A = a) - P(y|A = b)\|^2,$$

Where
***A*:** Sensitive attribute (e.g., race or gender).
***y*:** Model prediction.
The challenges outlined in this section underscore the need for scalable, context-aware, and ethically grounded methods for explaining LLMs. Addressing these issues will require advancements in both algorithmic techniques and interdisciplinary research.

Ethical Implications in Explainable AI for Large Language Models
As LLMs are increasingly deployed in sensitive and high-stakes domains, ensuring their ethical use is paramount. Explainable AI (XAI) introduces opportunities to enhance accountability and fairness but also raises unique ethical concerns. This section explores key implications, including bias mitigation, privacy preservation, and accountability in explanations.

Bias and Fairness

LLMs inherit biases from their training data, which can result in unfair or discriminatory outcomes [16]. Explanations should highlight these biases to prevent harmful downstream effects. For instance, if an LLM is used for hiring recommendations, gendered language in its training data may skew its outputs.

To mitigate bias, fairness-aware training or post-hoc explanation methods can be employed. Table 1 summarizes key approaches to address bias in LLMs.

Table 1 Fairness approaches in Explainable AI for LLMs

Approach	Description
Fair Representation Learning	Ensures representations are independent of sensitive attributes (e.g., gender, race) [19].
Counterfactual Explanations	Highlights feature changes that would lead to different decisions [4].
Post-hoc Bias Detection	Uses SHAP or similar methods to identify biased feature contributions [6].

Fairness in decision-making can be quantified using metrics like disparate impact, defined as:

$$DI = \frac{P(\text{Outcome}|A = a)}{P(\text{Outcome}|A = b)},$$

Where A is a sensitive attribute, and P(Outcome) is the probability of a favorable outcome.

Privacy Concerns

LLMs trained on sensitive or proprietary datasets can inadvertently memorize and reproduce confidential information [17]. For example, researchers demonstrated that GPT models could recall Social Security numbers embedded in training data [17, 20].

To address privacy risks, differential privacy techniques can be integrated into LLM training. The noise addition mechanism ensures that individual data points cannot be identified [21, 22].

$$P(M(D) = y) \leq e^{\epsilon} P(M(D') = y),$$

Where:

***M*(*D*):** Model output on dataset D.

***ϵ*:** Privacy budget, controlling the level of noise added.

***D*,*D*′:** Datasets differing by a single record.

Accountability in Explanations

Transparent explanations can enhance accountability but must be designed carefully to avoid manipulation. For instance, simplified explanations that omit important model limitations may mislead users [18]. Algorithm 1 outlines a pseudocode for generating accountable explanations in high-stakes domains.

Algorithm 1 Generating Accountable Explanations

Require: LLM f, input x, output y, explanation method g

Ensure: Accountable explanation e

```
Identify sensitive features S in x
Generate explanation e = g(f, x, y)
if S contributes significantly to y (e.g., φ(S) > threshold) then
    Adjust e to highlight sensitive feature impact
end if
Include model confidence and uncertainty in e return e
```

Ethical Guidelines and Standards

Regulations such as GDPR and the proposed EU AI Act emphasize the need for transparency and fairness in automated decision-making systems. Table 2 outlines ethical principles relevant to XAI.

By addressing the ethical implications discussed in this section, Explainable AI for LLMs can advance toward more transparent, fair, and responsible systems. The next section focuses on a human-centered framework to operationalize these goals.

Table 2: Ethical principles for Explainable AI

Principle	Description
Transparency	Provide users with clear and understandable explanations of AI decisions.
Fairness	Ensure equitable treatment of all groups, minimizing bias in outcomes.
Privacy Preservation	Avoid exposing sensitive data in explanations or outputs.
Accountability	Enable traceability and ensure stakeholders can identify the party responsible for AI decisions [4].

Human-Centered Framework for Explainable AI Design

Designing Explainable AI (XAI) systems that effectively cater to diverse user needs requires a human-centered approach. This section introduces a framework that integrates user profiling, dynamic explanation generation, and rigorous evaluation metrics to enhance usability, trust, and fairness.

User Profiling and Categorization

Users of LLM-based systems have varying levels of expertise, goals, and preferences. Tailored explanations are necessary to address these differences effectively. Table 3 summarizes common user categories and their needs.

Table 3: User profiles and explanation needs

User Category	Example	Explanation Preference
Non-Experts	General consumers, patients	Simplified, analogy-based explanations
Domain Experts	Doctors, financial analysts	Detailed, technical insights
Regulators/Policymakers	Legal pro-fessionals, compliance officers	High-level summaries with accountability metrics

User profiling can be formalized using a scoring mechanism, S_u, to represent user expertise and preference weights:

$$S_u = \sum_{i=1}^{n} w_i \cdot p_i$$

Where:
w_i: Weight assigned to each explanation attribute.
p_i: User preference for attribute i.

Dynamic Explanation Systems
Static, one-size-fits-all explanations often fail to meet diverse user requirements.
Dynamic systems tailor explanations based on user input, adapting in real time.
Algorithm 2 outlines a pseudocode for generating dynamic explanations tailored to user profiles.

Algorithm 2 Dynamic Explanation Generation

Require: LLM f, input x, user profile S_u
Ensure: Tailored explanation e

```
Analyze user profile S_u to determine explanation depth d
Generate initial prediction y = f(x)
Generate explanation e_0 = g(f, x, y)
if d > threshold then
    Enrich e_0 with technical details
else
    Simplify e_0 with analogy-based summaries
end if
Include interactivity features (e.g., "What if?" queries) return e_0
```

Evaluation Metrics for Explainability: Evaluating XAI systems involves both subjective and objective metrics. Table 4 lists commonly used metrics for assessing explanation effectiveness.
Subjective metrics, such as trust surveys, assess user perceptions, while objective metrics evaluate the practical impact of explanations on task performance.
Interactive Features for User Engagement
Interactive explanation systems enhance user understanding and engagement by allowing exploration of "what-if" scenarios and counterfactuals.

Table 4 Evaluation metrics for XAI systems

Metric	Description	Example Method
Trust	Measures user confidence in the AI system	Post-task surveys
Task Performance	Assesses decision accuracy with explanations	Comparing user outcomes with/without XAI
Cognitive Load	Gauges mental effort required to interpret explanations	Eye-tracking, response time analysis

Ethical and Fairness Considerations
Dynamic systems must incorporate fairness and ethics. For instance, user profiling systems should not discriminate based on sensitive attributes like gender or ethnicity.
Equation 16 formalizes a fairness constraint:

$$L_{\text{fair}} = \|e(x|A=a) - e(x|A=b)\|^2,$$

Where:
A: Sensitive attribute (e.g., gender).
$e(x)$: Explanation generated for input x.

The human-centered framework presented in this section emphasizes tailoring explanations to diverse user needs while ensuring fairness and accountability. In the next section, we summarize our findings and outline future directions for XAI in LLMs.

Conclusion and Future Directions: Explainable AI (XAI) for Large Language Models (LLMs) is a rapidly evolving field, driven by the need for transparency, accountability, and trust in artificial intelligence. This paper has explored the technical foundations of LLMs, key methods for interpreting their outputs, challenges in explanation design, ethical considerations, and a human-centered framework for building explainable systems.

Key Findings: Technical Foundations: LLMs leverage transformer architecture, which combines self-attention mechanisms, positional encoding, and embeddings to process text efficiently. Despite these advancements, the complexity of LLMs presents significant challenges for interpretability.

Interpretability Methods: Attention visualization, Layer-Wise Relevance Propagation (LRP),

counterfactual explanations, and feature attribution methods like SHAP provide tools for understanding LLM outputs. However, these methods often involve trade-offs between fidelity and simplicity.

Challenges: High dimensionality, context dependency, and computational overhead complicate the development of XAI systems for LLMs. Ethical concerns such as bias and privacy risks further exacerbate these challenges.

Human-Centered Design: Tailoring explanations to user profiles, employing interactive features, and incorporating fairness constraints are critical for building user-centric XAI systems.

Future Directions: Despite recent progress, there remain several open challenges and opportunities for advancing XAI in LLMs. Key areas for future research include:

Scalable Interpretability Techniques: Current interpretability methods often require significant computational resources, limiting their applicability to large-scale systems. Future research should focus on developing lightweight techniques that can explain LLM outputs without compromising accuracy or scalability.

Generalization across Domains: Most existing XAI methods are tailored to specific cases or datasets. A universal framework for LLM interpretability that generalizes across domains, such as healthcare, finance, and education, is needed to support diverse applications.

Ethics-Aware Design: Embedding ethical principles directly into XAI systems can prevent bias and ensure accountability. This includes integrating fairness-aware training, differential privacy, and robust mechanisms for detecting and mitigating ethical risks.

Interactive Explanation Systems: Future XAI systems should focus on interactivity, allowing users to engage dynamically with models through "what-if" scenarios, counterfactuals, and visualization tools.

Evaluation Frameworks for XAI Systems: Rigorous evaluation metrics, encompassing both subjective (e.g., user trust) and objective (e.g., task performance) measures, are essential for assessing the effectiveness of XAI systems. Developing standardized benchmarks for LLM explainability will facilitate comparative analysis and drive innovation.

Closing Remarks: As LLMs continue to permeate critical sectors, Explainable AI will play a pivotal role in ensuring their responsible deployment. By addressing the challenges outlined in this paper and embracing human-centered, ethics aware design principles, researchers and practitioners can create transparent, trustworthy, and equitable AI systems.
This work underscores the importance of interdisciplinary collaboration in advancing XAI. Bridging the gap between technical innovation, user needs, and ethical considerations will be key to unlocking the full potential of explainable LLMs.

References

1. Radford A, Wu J, Amodei D, *et al.* Language models are unsupervised multitask learners. OpenAI Blog. 2019;1(8):9.
2. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:1712.09923. 2017.
3. Lipton ZC. The mythos of model interpretability. Commun ACM. 2018;61(10):36-43.
4. Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv J Law Technol. 2017;31:841-887.
5. Clark K, Khandelwal U, Levy O, Manning CD. What does BERT look at? An analysis of BERT's attention. arXiv preprint arXiv:1906.04341. 2019.
6. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30.
7. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998-6008.
8. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018.
9. Raffel C, Shazeer N, Roberts A, *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(140):1-67.
10. Jain S, Wallace BC. Attention is not explanation. arXiv preprint arXiv:1902.10186. 2019.
11. Chefer H, Gur S, Wolf L. Transformer interpretability beyond attention visualization. Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR). 2021:782-791.
12. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. Proc 34th Int Conf Mach Learn (ICML). 2017:3319-3328.
13. Brown T, Mann B, Ryder N, *et al.* Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877-1901.
14. Olah C, Satyanarayan A, Johnson I, *et al.* The building blocks of interpretability. Distill. 2018;3(3):10.
15. Tenney I, Das D, Pavlick E. BERT rediscovers the classical NLP pipeline. Proc 57th Annu Meet Assoc Comput Linguist. 2019:4593-4601.
16. Binns R. Fairness in machine learning: Lessons from political philosophy. Proc 2018 Conf Fairness Accountability Transparency (FAT). 2018:149-159.
17. Carlini N, Tramer F, Wallace E, *et al.* Extracting training data from large language models. Proc 30th USENIX Secur Symp. 2021.
18. Danks D, London AJ. Algorithmic bias in autonomous systems. Proc 26th Int Joint Conf Artif Intell (IJCAI). 2017:4691-4697.
19. Zemel RS, Wu Y, Swersky K, Pitassi T, Dwork C. Learning fair representations. Proc 30th Int Conf Mach Learn (ICML). 2013:325-333.
20. Zhu D, Chen J, Shang W, Zhou X, Grossklags J, Hassan AE. DeepMemory: Model-based memorization analysis of deep neural language models. Proc 36th IEEE/ACM Int Conf Autom Softw Eng (ASE). 2021:1003-1015.
21. Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, *et al.* Deep learning with differential privacy. Proc 2016 ACM SIGSAC Conf Comput Commun Secur. 2016:308-318.
22. Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. J Priv Confid. 2006;7(3):17-51.