Interpretability in the Era of Large Language Models: Opportunities and Challenges

Anonymous authors
Paper under double-blind review

Abstract

Interpretable machine learning has exploded as an area of interest over the last decade, sparked by the rise of increasingly large datasets and deep neural networks. Simultaneously, large language models (LLMs) have demonstrated remarkable capabilities across a wide array of tasks, offering a chance to rethink opportunities in interpretable machine learning. Notably, the capability to explain in natural language allows LLMs to expand the scale and complexity of patterns that can be given to a human. However, these new capabilities raise new challenges, such as hallucinated explanations and immense computational costs.

In this position paper, we start by reviewing existing methods to evaluate the emerging field of LLM interpretation (both interpreting LLMs and using LLMs for explanation). We contend that, despite their limitations, LLMs hold the opportunity to redefine interpretability with a more ambitious scope across many applications, including in auditing LLMs themselves. We highlight two emerging research priorities for LLM interpretation: using LLMs to directly analyze new datasets and to generate interactive explanations.

1 Introduction

Machine learning (ML) and natural language processing (NLP) have seen a rapid expansion in recent years, due to the availability of increasingly large datasets and powerful neural network models. In response, the field of interpretable ML* has grown to incorporate a diverse array of techniques and methods for understanding these models and datasets (Doshi-Velez & Kim, 2017; Murdoch et al., 2019; Molnar, 2019). One part of this expansion has focused on the development and use of inherently interpretable models (Rudin et al., 2021), such as sparse linear models, generalized additive models, and decision trees. Alongside these models, post-hoc interpretability techniques have become increasingly prominent, offering insights into predictions after a model has been trained. Notable examples include methods for assessing feature importance (Ribeiro et al., 2016; Lundberg & Lee, 2017), and broader post-hoc techniques, e.g., model visualizations (Yosinski et al., 2015; Bau et al., 2018), or interpretable distillation (Tan et al., 2018; Ha et al., 2021).

Meanwhile, pre-trained large language models (LLMs) have shown impressive proficiency in a range of complex NLP tasks, significantly advancing the field and opening new frontiers for applications (Brown et al., 2020; Touvron et al., 2023; OpenAI, 2023). However, the inability to effectively interpret these models has debilitated their use in high-stakes applications such as medicine, and raised issues related to regulatory pressure, safety, and alignment (Goodman & Flaxman, 2016; Amodei et al., 2016; Gabriel, 2020). Moreover, this lack of interpretability has limited the use of LLMs (and other neural-network models) in fields such as science and data analysis (Wang et al., 2023a; Kasneci et al., 2023; Ziems et al., 2023). In these settings, the end goal is often to elicit a trustworthy interpretation rather than to deploy an LLM. In other settings, interpretability may instead be used as a tool to audit, understand, or manipulate LLMs.

In this work, we contend that LLMs hold the opportunity to rethink interpretability with a more ambitious scope. LLMs can elicit more elaborate explanations than the previous generation of interpretable ML techniques. While previous methods have often relied on restricted interfaces such as saliency maps, LLMs

*We use the terms interpretable, explainable, and transparent interchangeably.

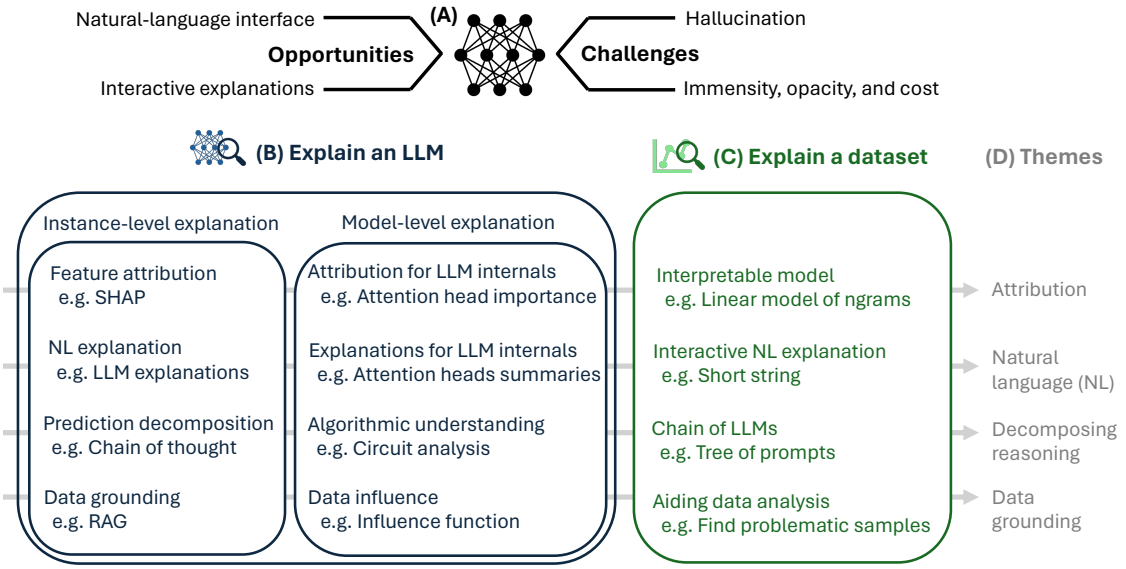

Figure 1: Categorization of LLM interpretation research. **(A)** LLMs raise unique opportunities and challenges for interpretation (Sec. 3). **(B)** Explaining an LLM can be categorized into methods that seek to explain a single generation from an LLM (i.e. instance-level explanation, Sec. 4.1) or the LLM in its entirety (i.e. model-level explanation, Sec. 4.2). Instance-level explanation methods build on many techniques that were originally developed for interpreting non-LLM models, such as feature attribution methods. More recent instance-level explanation techniques use LLMs themselves to yield interpretations, e.g., through post-hoc natural language (NL) explanations, asking an LLM to build explanations into its generation process, or through data grounding. Similar techniques have been developed and applied to model-level explanation, although it also includes unique types of explanations, e.g., analyzing individual attention heads or circuits inside an LLM. **(C)** Sec. 5 analyzes the emerging area that uses an LLM to aid in directly explaining a *dataset*. In this setting, an LLM is given a new dataset (which can consist of either text or tabular features) and is used to help analyze it. LLM-based techniques for dataset explanation are quite diverse, including helping to build interpretable models, generate NL explanations, generate chains of NL explanations, or construct data visualizations. **(D)** Common themes emerge among methods for instance-level explanation, model-level explanation, and dataset explanation.

can communicate directly in expressive natural language. This allows users to make targeted queries, such as *Can you explain your logic?*, *Why didn't you answer with (A)?*, or *Explain this data to me.*, and get immediate, relevant responses. We believe simple questions such as these, coupled with techniques for grounding and processing data, will allow LLMs to articulate previously incomprehensible model behaviors and data patterns directly to humans in understandable text. However, unlocking these opportunities requires tackling new challenges, including hallucinated (i.e. incorrect or baseless) explanations, along with the immense size, cost, and inherent opaqueness of modern LLMs.

Contributions and overview We evaluate LLM interpretation and highlight emerging research priorities, taking a broader scope than recent works, e.g., those focused on explaining LLM predictions (Zhao et al., 2023), mechanistic interpretability (Räuker et al., 2023), social science (Ziems et al., 2023), or science more generally (Wang et al., 2023a; Birhane et al., 2023; Pion-Tonachini et al., 2021). Rather than providing an exhaustive overview of methods, we highlight the aspects of interpretability that are unique to LLMs and showcase them with practically useful methods.

Specifically, we begin with a background and definitions (Sec. 2) before proceeding to analyze the unique opportunities and challenges that LLMs present for interpretation (Sec. 3). We then ground these oppor-

tunities in two complementary categories for LLM-based interpretation (see Fig. 1). The first is generating explanations for *an existing LLM* (Sec. 4), which is useful for auditing a model's performance, alignment, fairness, etc. The second is explaining *a dataset* (Sec. 5); in this setting, an LLM is used to help analyze a new dataset (which can consist of either text or tabular features). While we cover these areas separately, there is a great deal of interplay between them. For example, explaining a dataset relies on many tools originally developed to extract reliable explanations when explaining an LLM. Conversely, explaining an LLM often requires understanding the dataset on which it was trained, as well as what dataset characteristics elicit different LLM behaviors.

Throughout the paper, we highlight dataset explanation and interactive explanation as emerging research priorities. Together, these two areas have great potential real-world significance in domains from science to statistics, where they can facilitate the process of scientific discovery, data analysis, and model building. Throughout, we focus on pre-trained LLMs, mostly applied to text data, but also applied to tabular data.

2 Background: definitions and evaluation

2.1 Definitions

Interpretability, when used without context, is a poorly defined concept. Precisely defining interpretability requires understanding the problem and audience an interpretation is intended to serve. In light of this imprecision, interpretable ML has largely become associated with a narrow set of techniques, including feature attribution, saliency maps, and transparent models. However, LLM interpretation is broader in scope and more expressive than these methods. Here, we paraphrase the definition of interpretable ML from Murdoch et al. 2019 to define LLM interpretation as the *extraction of relevant knowledge from an LLM concerning relationships either contained in data or learned by the model.* We emphasize that this definition applies to both interpreting an LLM and to using an LLM to generate explanations. Moreover, the definition relies on the extraction of *relevant* knowledge, i.e., knowledge that is useful for a particular problem and audience. For example, in a code generation context, a relevant interpretation may help a user quickly integrate an LLM-generated code snippet. In contrast, a relevant interpretation in a medical diagnosis setting may inform a user whether or not a prediction is trustworthy.

Large language model is a term that is often used imprecisely. Here, we use it to refer to transformer-based neural language models that contain tens to hundreds of billions of parameters, and which are pre-trained on massive text data, e.g., PaLM (Chowdhery et al., 2023), LLaMA (Touvron et al., 2023), and GPT-4 (OpenAI, 2023). Compared to early pre-trained language models (such as BERT (Devlin et al., 2018)), LLMs are not only much larger, but also exhibit stronger language understanding, generation abilities, and explanation capabilities. After an initial computationally intensive pre-training stage, LLMs often undergo instruction finetuning and further alignment with human preferences to improve instruction following (Ouyang et al., 2022) or to improve interactive chat capabilities, e.g., the LLaMA-2 chat model (Touvron et al., 2023). They are sometimes also further adapted via supervised finetuning to improve performance in a specific domain, such as medicine (Singhal et al., 2023).

Prompting is the most common interface for applying LLMs (and our main focus in this paper), once they have been trained. In prompting, a text prompt is fed into an LLM and used to generate subsequent output text, e.g. an answer to a question. *Few-shot prompting* is a type of prompting that involves providing an LLM with a small number of examples to allow it to better understand the task it is being asked to perform.

Interpretability is intricately related to other research areas where LLMs have begun to play an expanding role. For example, in the field of causal inference, interpreting and querying LLMs may help extract and test hypotheses for causal relationships contained in data (Kıcıman et al., 2023). Additionally, interpretability is a major topic when considering the bias, fairness, privacy, and security of LLMs (Yao et al., 2024; Li et al., 2023d), where it can be used to expose, fix, or exploit issues with an LLM.

2.2 Evaluating LLM interpretations

Evaluating interpretability is difficult in general (Doshi-Velez & Kim, 2017), requiring that interpretations be faithful to an underlying process in an LLM/dataset while remaining understandable to a human. This process is made even more difficult by the use of LLMs, which broadens the space of explanations and complicates rigorous understanding. In this section, we briefly overview three key approaches to evaluating LLM interpretations through (i) human studies, (ii) automated metrics, and (iii) using the interpretation to improve model performance. Careful consideration is required to match the choice of metric to the context of a particular problem.

Evaluation with human studies Since different interpretations are relevant to different contexts, the ideal way to evaluate an interpretation is by studying whether its usage in a real-world setting with humans improves a desired outcome (Kim et al., 2017). This approach directly measures the utility of an interpretation in its appropriate context. However, these targeted human studies have some shortcomings: They are often laborious to conduct and are context-specific, i.e. an interpretation that helps humans in some contexts may not help them in others. Additionally, human studies that simply measure a proxy for usefulness (i.e. measuring human judgment of explanations) are often uninformative, as they may not translate into improvements in practice (Adebayo et al., 2018). A recent meta-analysis finds that introducing NLP explanations into settings with humans yields widely varying utilities, ranging from completely unhelpful to very useful (Chaleshtori et al., 2023). An important piece of this evaluation is the notion of complementarity (Bansal et al., 2021), i.e., that explanations should help LLMs complement human performance in a team setting, rather than improve their performance in isolation.

Evaluation with automated metrics While human studies provide the most realistic evaluation, automated metrics (that can be computed without involving humans) are desirable to ease and scale evaluation, especially in model-level explanation. An increasingly popular approach is to use LLMs themselves in evaluation, although great care must be taken to avoid introducing biases, e.g., an LLM systematically scoring its own outputs too positively (Zheng et al., 2023). One way to reduce bias is to use LLMs as part of a structured evaluation process tailored to a particular problem, rather than directly querying LLMs for evaluation scores. For example, one common setting is evaluating a natural-language interpretation of a given function (which may be any component of a pre-trained LLM). In this setting, one can evaluate an explanation's ability to simulate the function's behavior (Bills et al., 2023), the function's output on LLM-generated synthetic data (Singh et al., 2023b), or its ability to recover a groundtruth function (Schwettmann et al., 2023; Zhong et al., 2023). In a question-answering setting, many automated metrics have been proposed for measuring the faithfulness of a natural-language explanation for an individual answer to a question (Atanasova et al., 2023; Parcalabescu & Frank, 2023; Chen et al., 2022).

Evaluation through improving model performance A third avenue for evaluating interpretations is through their ability to control or improve model performance. This approach provides strong evidence for the utility of an explanation, although it does not encompass all critical use cases of interpretability (particularly those directly involving human interaction). Model improvements can take various forms, the simplest of which is simply improving accuracy at downstream tasks. For example, few-shot accuracy was seen to improve when aligning an LLM's rationales with explanations generated using post-hoc explanation methods (Krishna et al., 2023) or explanations distilled from large models (Mukherjee et al., 2023). Moreover, employing few-shot explanations during inference (not training) can significantly improve few-shot LLM accuracy, especially when these explanations are further optimized (Lampinen et al., 2022; Ye & Durrett, 2023). Beyond general performance, explanations can be used to overcome specific shortcomings of a model. For example, one line of work identifies and addresses shortcuts/spurious correlations learned by an LLM (Du et al., 2023; Kang & Choi, 2023; Bastings et al., 2021). Model editing, a related line of work, enables precise modifications to certain model behaviors, enhancing overall performance (Meng et al., 2022; Mitchell et al., 2022; Hernandez et al., 2023).

3 Unique opportunities and challenges of LLM interpretation

In this section, we discuss two unique opportunities and two challenges that arise in LLM interpretation (see Fig. 1A).

Unique opportunities of LLM interpretation First among LLM interpretation opportunities is the ability to provide *a natural-language interface* to explain complex patterns. This interface is familiar to humans, potentially ameliorating the difficulties that practitioners often face when using explainability techniques (Kaur et al., 2020; Weld & Bansal, 2019). Additionally, natural language can be used to build a bridge between humans and a range of other modalities, e.g., DNA, chemical compounds, or images (Taylor et al., 2022; Liu et al., 2023b; Radford et al., 2021), that may be difficult for humans to interpret on their own. In these cases, natural language allows for expressing complex concepts through explanations at different levels of granularity, potentially grounded in evidence or discussions of counterfactuals. For example, a natural language explanation for an LLM's answer to a question may highlight the LLM's coarse reasoning if the LLM is prompted for a high-level explanation or instead to specific words in the input if the LLM is asked for a more fine-grained explanation.

A second major opportunity is the ability for LLMs to generate *interactive explanations*. Interactivity allows users to tailor explanations to their unique needs, e.g., by asking follow-up questions and performing analysis on related examples. Interviews with decision-makers, including physicians and policymakers, indicate that they strongly prefer interactive explanations, particularly in the form of natural-language dialogues (Lakkaraju et al., 2022). Interactivity further allows LLM explanations to be decomposed into many different LLM calls, each of which can be audited independently. This can be enabled in different ways, e.g., having a user repeatedly chat with an LLM using prompting, or providing a user a sequence of LLM calls and evidence to analyze.

Unique challenges of LLM interpretation These opportunities bring new challenges. First and foremost is the issue of *hallucination*, i.e. incorrect or baseless LLM generations. In our setting, we focus on hallucinated explanations generated by an LLM; Flexible explanations provided in natural language can quickly become less grounded in evidence, whether the evidence is present in a given input or presumed to be present in the knowledge an LLM has learned from its training data. Hallucinated explanations are unhelpful or even misleading, and thus techniques for identifying and combating hallucination are critical to the success of LLM interpretation.

A second challenge is the *immensity and opaqueness* of LLMs. Models have grown to contain tens or hundreds of billions of parameters (Brown et al., 2020; Touvron et al., 2023), and continue to grow in size. This makes it infeasible for a human to inspect or even comprehend the units of an LLM. Moreover, it necessitates efficient algorithms for interpretation, as even generating a single token from an LLM often incurs a non-trivial computational cost. In fact, LLMs are often too large to be run locally or can be accessed only through a proprietary text API, necessitating the need for interpretation algorithms that do not have full access to the model (e.g., no access to the model weights or the model gradients).

4 Explaining an LLM

In this section, we study techniques for explaining an LLM, including explaining a single generation from an LLM (Sec. 4.1) or an LLM in its entirety (Sec. 4.2). Most of the methods we cover are specific to LLMs, but we also cover some general interpretable ML techniques (e.g. feature attributions) as well as some methods that apply to neural networks that are not LLMs (e.g. probing); see a full breakdown in Table 1.

Table 1: Breakdown of LLM interpretation methods. First categorizes methods based on the model class they apply to: LLMs, deep neural networks (DNNs), or general ML methods (ML). Additionally categorizes methods based on whether they fall under instance-level interpretation (I-L), model-level interpretation (M-L), and whether they us an LLM as a tool to generate natural-language explanations (N-L). All methods in Sec. 5 use LLMs as a tool generate different forms of interpretations at the dataset-level.

Methods	Example works	Applicability	Usage
Feature attributions	Lundberg & Lee 2017;Enguehard 2023	ML	I-L
Interpreting attention scores	Jain & Wallace 2019;Bibal et al. 2022	LLM	I-L
Natural-language explanations	Bhattacharjee et al. 2023;Chen et al. 2023b	LLM	I-L, N-L
Chain of thought	Wei et al. 2022	LLM	I-L, N-L
RAG	Guu et al. 2020;Worledge et al. 2023	LLM	I-L
Probing	Conneau et al. 2018;Zou et al. 2023	DNN	M-L
Understanding internal model components	Mu & Andreas 2020;Singh et al. 2023b	DNN	M-L
Toy LLM models	Elhage et al. 2021;Olsson et al. 2022	LLM	M-L, N-L
Investigating training data	Grosse et al. 2023;Kandpal et al. 2023	ML	M-L
Chat-based interactivity	Slack et al. 2022;Wang et al. 2024	LLM	M-L, N-L

4.1 Instance-level explanation

Instance-level explanation, i.e., explaining a single generation from an LLM, has been a major focus in the recent interpretability literature. It allows for understanding and using LLMs in high-stakes scenarios, e.g., healthcare.

Feature attributions The simplest approach for providing instance-level explanations in LLMs provides feature attributions for input tokens. These feature attributions assign a relevance score to each input feature, reflecting its impact on the model's generated output. Various attribution methods have been developed, including perturbation-based methods (Lundberg & Lee, 2017), gradient-based methods (Sundararajan et al., 2017; Montavon et al., 2017; Li et al., 2015), and linear approximations (Ribeiro et al., 2016). Recently, these methods have been specifically adapted for transformer models, addressing unique challenges such as discrete token embeddings (Sikdar et al., 2021; Enguehard, 2023) and computational costs (Chen et al., 2023a). Moreover, the conditional distribution learned by an LLM can be used to enhance existing attribution methods, e.g., by performing input marginalization (Kim et al., 2020). These works focus on attributions at the token-level, but may consider using a less granular breakdown of the input to improve understandability (Zafar et al., 2021).

Interpreting attention scores Besides feature attributions, attention mechanisms within an LLM offer another avenue for visualizing token contributions to an LLM generation (Wiegreffe & Pinter, 2019), though their faithfulness/effectiveness remains unclear (Jain & Wallace, 2019; Bibal et al., 2022). Interestingly, recent work suggests that LLMs themselves can generate post-hoc attributions of important features through prompting (Kroeger et al., 2023). This approach could be extended to enable eliciting different feature attributions that are relevant in different contexts.

Natural-language explanations Beyond token-level attributions, LLMs can also generate instance-level explanations directly in natural language. While the generation of natural-language explanations predates the current era of LLMs (e.g., in text classification (Camburu et al., 2018; Rajani et al., 2019) or image classification (Hendricks et al., 2016)), the advent of more powerful models has significantly enhanced their effectiveness. Natural-language explanations generated by LLMs have shown the ability to elucidate model predictions, even simulating counterfactual scenarios (Bhattacharjee et al., 2023), and expressing nuances like uncertainty (Xiong et al., 2023; Tanneru et al., 2023; Zhou et al., 2024). Despite their potential benefits, natural language explanations remain extremely susceptible to hallucination or inaccuracies, especially when generated post-hoc (Chen et al., 2023b; Ye & Durrett, 2022).

Chain of thought One starting point for combating these hallucinations is integrating an explanation within the answer-generation process itself. Chain-of-thought prompting exemplifies this approach (Wei et al., 2022), where an LLM is prompted to articulate its reasoning step-by-step before arriving at an answer. This reasoning chain generally results in more accurate and faithful outcomes, as the final answer is often more aligned with the preceding logical steps. The faithfulness of the produced step-by-step explanation can be tested by introducing perturbations in the reasoning process and observing the effects on the final output (Madaan & Yazdanbakhsh, 2022; Wang et al., 2022a; Lanham et al., 2023). Alternative methods for generating this reasoning chain exist, such as tree-of-thoughts (Yao et al., 2023), which extends chain-of-thought to instead generate a tree of thoughts used in conjunction with backtracking, graph-of-thoughts (Besta et al., 2023), and others (Nye et al., 2021; Press et al., 2022; Zhou et al., 2022). All of these methods not only help convey an LLM's intermediate reasoning to a user, but also help the LLM to follow the reasoning through prompting, often enhancing the reliability of the output. However, like all LLM-based generations, the fidelity of these explanations can vary (Lanham et al., 2023; Wei et al., 2023).

RAG An alternative path to reducing hallucinations during generation is to employ retrieval-augmented generation (RAG). In RAG, an LLM incorporates a retrieval step in its decision-making process, usually by searching a reference corpus or knowledge base using text embeddings (Guu et al., 2020; Peng et al., 2023); see review (Worledge et al., 2023). This allows the information that is used to generate an output to be specified and examined explicitly, making it easier to explain the evidence an LLM uses during decision-making.

4.2 Model-level explanation

Rather than studying individual generations, model-level explanation aims to understand an LLM as a whole, usually by analyzing its parameters. Recently, many model-level explanations of how model parameters algorithmically yield model behaviors have been labeled *mechanistic interpretability*, though the use/omission of this label is often imprecise, e.g. see (Räuker et al., 2023). Model-level explanations can help to audit a model for concerns beyond generalization, e.g., bias, privacy, and safety, helping to build LLMs that are more efficient / trustworthy. They can also yield mechanistic understanding about how LLMs function. Towards this end, researchers have focused on summarizing the behaviors and inner workings of LLMs through various lenses. Generally, these works require access to model weights and do not work for explaining models that are only accessible through a text API, e.g., GPT-4 (OpenAI, 2023).

Probing One popular method for understanding neural-network representations is probing. Probing techniques analyze a model's representation by decoding embedded information (Adi et al., 2016; Conneau et al., 2018; Giulianelli et al., 2018). In the context of LLMs, probing has evolved to include the analysis of attention heads (Clark et al., 2019), embeddings (Morris et al., 2023a), and different controllable aspects of representations (Zou et al., 2023). More recent methods directly decode an output token to understand what is represented at different positions and layers (Belrose et al., 2023; Ghandeharioun et al., 2024) or connecting multimodal embeddings with text embeddings to help make them more understandable (Oikarinen & Weng, 2022; Oikarinen et al., 2023; Bhalla et al., 2024). These methods can provide a deeper understanding of the nuanced ways in which LLMs process and represent information.

Understanding internal model components In addition to probing, many works study LLM representations at a more granular level. This includes categorizing or decoding concepts from individual neurons (Mu & Andreas, 2020; Gurnee et al., 2023; Dalvi et al., 2019; Lakretz et al., 2019) or directly explaining the function of attention heads in natural language (Bills et al., 2023; Hernandez et al., 2022b). Beyond individual neurons, there is growing interest in understanding how groups of neurons combine to perform specific tasks, e.g., finding a circuit for indirect object identification (Wang et al., 2022b), for entity binding (Feng & Steinhardt, 2023), or for multiple shared purposes (Merullo et al., 2023). More broadly, this type of analysis can be applied to localize functionalities rather than fully explain a circuit, e.g., localizing factual knowledge within an LLM (Meng et al., 2022; Dai et al., 2021). A persistent problem with these methods is that they are difficult to scale to immense LLMs, leading to research in (semi)-automated methods that can scale to today's largest LLMs (Lieberum et al., 2023; Wu et al., 2023).

Toy LLM models A complementary approach to mechanistic understanding uses miniature LLMs as a test bed for investigating complex phenomena. For example, examining a 2-layer transformer model reveals

information about what patterns are learned by attention heads as a function of input statistics (Elhage et al., 2021) or helps identify key components, such as induction heads or ngram heads that copy and utilize relevant tokens (Olsson et al., 2022; Akyürek et al., 2024). This line of mechanistic understanding places a particular focus on studying the important capability of in-context learning, i.e., given a few input-output examples in a prompt, an LLM can learn to correctly generate an output for a new input (Garg et al., 2022; Zhou et al., 2023).

Investigating training data A related area of research seeks to interpret an LLM by understanding the influence of its training data distribution. Unlike other methods we have discussed, this requires access to an LLM's training dataset, which is often unknown or inaccessible. In the case that the data is known, researchers can employ techniques such as influence functions to identify important elements in the training data (Grosse et al., 2023). They can also study how model behaviors arise from patterns in training data, such as hallucination in the presence of long-tail data (Kandpal et al., 2023), in the presence of repeated training data (Hernandez et al., 2022a), in-context learning (Chen et al., 2024b), statistical patterns that contradict proper reasoning (McKenna et al., 2023), and others (Swayamdipta et al., 2020).

Chat-based interactivity All these interpretation techniques can be improved via LLM-based interactivity, allowing a user to investigate different model components via follow-up queries and altered prompts from a user. For example, one recent work introduces an end-to-end framework for explanation-based debugging and improvement of text models, showing that it can quickly yield improvements in text-classification performance (Lee et al., 2022). Another work, Talk2Model, introduces a natural-language interface that allows users to interrogate a tabular prediction model through a dialog, implicitly calling many different model explainability tools, such as calculating feature importance (Slack et al., 2022).[†] More recent work extends Talk2Model to a setting interrogating an LLM about its behavior (Wang et al., 2024).

Model manipulation Finally, the insights gained from model-level understanding are beginning to inform practical applications, with current areas of focus including model editing (Meng et al., 2022), improving instruction following (Zhang et al., 2023b), and model compression (Sharma et al., 2023). These areas simultaneously serve as a sanity check on many model-level interpretations and as a useful path to enhancing the reliability of LLMs.

5 Explaining a dataset

As LLMs improve their context length and capabilities, they can be leveraged to explain an entire dataset, rather than explaining an LLM or its generations. This can aid with data analysis, knowledge discovery, and scientific applications. Fig. 2 shows an overview of dataset explanations at different levels of granularity, which we cover in detail below. We distinguish between tabular and text data, but note that most methods can be successfully applied to either, or both simultaneously in a multimodal setting.

5.1 Tabular dataset explanation

LLM-aided data analysis One way LLMs can aid in dataset explanation is by making it easier to interactively visualize and analyze tabular data. This is made possible by the fact that LLMs can simultaneously understand code, text, and numbers by treating them all as input tokens. Perhaps the most popular method in this category is ChatGPT Code Interpreter[‡], which enables uploading datasets and building visualizations on top of them through an interactive text interface. Underlying this interface, the LLM makes calls to tools such as python functions to create the visualizations. This capability is part of a broader trend of LLM-aided visualization, e.g., suggesting automatic visualizations for dataframes (Dibia, 2023), helping to automate data wrangling (Narayan et al., 2022), or even conducting full-fledged data analysis (Huang et al., 2023a) with accompanying write-ups (Ifargan et al., 2024). These capabilities benefit from a growing line of work that analyzes how to effectively represent and process tabular data with LLMs (Li et al., 2023b; Zhang et al., 2023a;c).

[†]Note that Talk2Model focuses on interpreting prediction models rather than LLMs.

[‡]`https://openai.com/blog/chatgpt-plugins#code-interpreter`

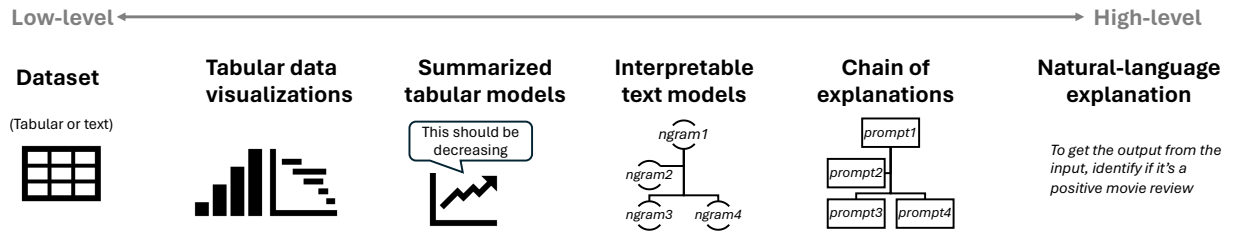

Figure 2: Dataset explanations at different levels of granularity. Dataset explanation involves understanding a new dataset (consisting of either text or tabular features) using a pre-trained LLM. Low-level explanations are more faithful to the dataset but involve more human effort to extract meaningful insights. Many dataset interpretations use prediction models (classification or regression) as a means to identify and explain patterns between features.

Explaining models fit to tabular datasets LLMs can also help explaining datasets by directly analyzing models that have been fit to tabular data. Unlike model-level explanation, where the goal is to understand the model, in dataset explanation, the goal is to understand patterns in the data through the model (although similar techniques can be used for both problems). For example, one recent work uses LLMs to analyze generalized additive models (GAMs) that are fit to tabular data (Lengerich et al., 2023). GAMs are interpretable models that can be represented as a set of curves, each representing the contribution of a feature to the output prediction as a function of the feature's value. An LLM can analyze the fitted model (and thereby the underlying dataset) by processing each curve as a set of numerical tokens and then detecting and describing patterns in each curve. Lengerich et al. find that LLMs can identify surprising characteristics in the curves and the underlying data, largely based on their prior knowledge of a domain. Rather than using an interpretable GAM model, another approach is to distill dataset insights by analyzing classifier predictions. For example, MaNtLE generates natural-language descriptions of a classifier's rationale based on the classifier's predictions, and these explanations are found to identify explainable subgroups that contain similar feature patterns (Menon et al., 2023).

5.2 Text dataset explanation

Fully interpretable models Text data poses different challenges for dataset explanation than tabular data because it is sparse, high-dimensional, and modeling it requires many high-order interactions. As a result, interpretable models that have been successful in the tabular domain (e.g., sparse linear models (Tibshirani, 1996; Ustun & Rudin, 2016), GAMs (Hastie & Tibshirani, 1986; Lou et al., 2013; Caruana et al., 2015), decision trees (Breiman et al., 1984; Quinlan, 1986; Agarwal et al., 2022), and others (Rudin, 2018)), have struggled to accurately model text. One recent line of work addresses this issue by using LLMs to help build fully interpretable text models, such as linear models or decision trees (Singh et al., 2023a); the resulting models are surprisingly accurate, often outperforming even much larger LLM models. These interpretable models can help explain a dataset by showing which features (i.e. words or ngrams) are important for predicting different outcomes. Similar methods, use LLMs to build interpretable representations for text classification (McInerney et al., 2023), text style (Patel et al., 2023), or text regression (Benara et al., 2024).

Partially interpretable models Going beyond fully interpretable models, LLMs also help in building partially interpretable text models. Partially interpretable text models often employ chains of prompts; these chains allow for decomposing an LLM's decision-making process to analyze which dataset patterns a model learns. Prompt chains are usually constructed by humans or by querying a model to generate a chain of calls on-the-fly (Grunde-McLaughlin et al., 2023). For dataset explanation, the most relevant chains are sequences of explanations that are generated by an LLM. For example, a model can generate a single tree of explanations that is shared across all examples in a dataset, a process that enables understanding hierarchical structures stored within a dataset (Morris et al., 2023b). Rather than a tree, a single chain of prompts can often help an LLM employ self-verification, i.e. the model itself checks its previous generations

using a chain of prompts, a popular technique that often improves reliability (Pan et al., 2023; Madaan et al., 2023; Gero et al., 2023). As in instance-level explanation, an LLM can incorporate a retrieval step in its decision-making process (Worledge et al., 2023), and access to different tools can help make different steps (e.g., arithmetic) more reliable and transparent (Mialon et al., 2023).

Natural-language explanations Natural-language explanations hold the potential to produce rich, concise descriptions of patterns present in a dataset, but are prone to hallucination. One method, iPrompt (Singh et al., 2023c), aims to avoid hallucination by searching for a dataset explanation in the form of a single prompt, and verifying that the prompt induces an LLM to accurately predict a pattern in the underlying dataset. Another work learns and verifies a library of natural-language rules that help improve question answering (Zhu et al., 2023). Related methods use LLMs to provide descriptions that differentiate between groups in a dataset, followed by an LLM that verifies the credibility of the description (Zhong et al., 2022; 2023; Zhu et al., 2022). In addition to a raw natural-language explanation, LLMs can aid in summarizing textual information, e.g., through explainable clustering of a text dataset (Wang et al., 2023b) or creating prompt-based topic models (Pham et al., 2023).

6 Future research priorities

We now highlight research priorities surrounding LLM interpretation in three areas: explanation reliability, knowledge discovery, and interactive explanations.

Explanation reliability All LLM-generated explanations are bottlenecked by reliability issues. This includes hallucinations (Tonmoy et al., 2024), but encompasses a broader set of issues. For example, LLMs continue to be very sensitive to the nuances of prompt phrasing; minor variations in prompts can completely change the substance of an LLM output (Sclar et al., 2023; Turpin et al., 2023). Additionally, LLMs may ignore parts of their context, e.g., the middle of long contexts (Liu et al., 2023a) or instructions that are difficult to parse (Zhang et al., 2023b).

These reliability issues are particularly critical in interpretation, which often uses explanations to mitigate risk in high-stakes settings. LLM interpretations rarely come with theoretical guarantees of reliability, and those that do are often for very constrained settings. One work analyzing explanation reliability finds that LLMs often generate seemingly correct explanations that are actually inconsistent with their own outputs on related questions (Chen et al., 2023b), preventing a human practitioner from trusting an LLM or understanding how its explanations apply to new scenarios. Another study finds that explanations generated by an LLM may not entail the model's predictions or be factually grounded in the input, even on simple tasks with extractive explanations (Ye & Durrett, 2022).

Future work is required to improve the grounding of explanations and develop stronger, computationally tractable methods to test their reliability, perhaps through methods such as RAG (Worledge et al., 2023; Patel et al., 2024), self-verification (Pan et al., 2023), iterative prompting/grounding (Singh et al., 2023c), or automatically improving model self-consistency (Chen et al., 2024a; Li et al., 2023c; Akyürek et al., 2024).

Knowledge discovery (through dataset explanation) Dataset explanation using LLMs (Sec. 5) holds the potential to help with the generation and discovery of new knowledge from data (Wang et al., 2023a; Birhane et al., 2023; Pion-Tonachini et al., 2021), rather than simply helping to speed up data analysis or visualization. Dataset explanation could initially help at the level of brainstorming scientific hypotheses that can then be screened or tested by human researchers (Yang et al., 2023). During and after this process, LLM explanations can help with using natural language to understand data from otherwise opaque domains, such as chemical compounds (Liu et al., 2023b) or DNA sequences (Taylor et al., 2022). In the algorithms domain, LLMs have been used to uncover new algorithms, translating them to humans as readable computer programs (Romera-Paredes et al., 2023). These approaches could be combined with data from experiments to help yield new data-driven insights.

LLM explanations can also be used to help humans better understand and perform a task. Explanations from transformers have already begun to be applied to domains such as Chess, where their explanations

can help improve even expert players (Schut et al., 2023). Additionally, LLMs can provide explanations of expert human behavior, e.g. "Why did the doctor prescribe this medication given this information about the patient?", that are helpful in understanding, auditing, and improving human behavior (Tu et al., 2024).

Interactive explanations Finally, advancements in LLMs are poised to allow for the development of more user-centric, interactive explanations. LLM explanations and follow-up questions are already being integrated into a variety of LLM applications, such as interactive task specification (Li et al., 2023a), recommendation (Huang et al., 2023b), and a wide set of tasks involving dialog. Furthermore, works like Talk2Model (Slack et al., 2022) enable users to interactively audit models in a conversational manner. This dialog interface could be used in conjunction with many of the methods covered in this work to help with new applications, e.g., interactive dataset explanation.

7 Conclusions

In this paper, we have explored the vast and dynamic landscape of interpretable ML, particularly focusing on the unique opportunities and challenges presented by LLMs. LLMs' advanced natural language generation capabilities have opened new avenues for generating more elaborate and nuanced explanations, allowing for a deeper and more accessible understanding of complex patterns in data and model behaviors. As we navigate this terrain, we assert that the integration of LLMs into interpretative processes is not merely an enhancement of existing methodologies but a transformative shift that promises to redefine the boundaries of machine learning interpretability.

Our position is anchored in the belief that the future of interpretable ML hinges on our ability to harness the full potential of LLMs. To this end, we outlined several key stances and directions for future research, such as enhancing explanation reliability and advancing dataset interpretation for knowledge discovery. As LLMs continue to improve rapidly, these explanations (and all the methods discussed in this work) will advance correspondingly to enable new applications and insights. In the near future, LLMs may be able to offer the holy grail of interpretability: explanations that can reliably aggregate and convey extremely complex information to us all.

Broader impact statement

This paper presents work whose goal is to advance the field of LLM interpretation, a crucial step toward addressing the challenges posed by these often opaque models. Although LLMs have gained widespread use, their lack of transparency can lead to significant harm, underscoring the importance of interpretable AI. There are many potential positive societal consequences of this form of interpretability, e.g., facilitating a better understanding of LLMs and how to use them safely, along with a better understanding of scientific data and models. Nevertheless, as is the case with most ML research, the interpretations could be used to interpret and potentially improve an LLM or dataset that is being used for nefarious purposes.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pp. 9505–9515, 2018. ↪4.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*, 2016. ↪7.

Abhineet Agarwal, Yan Shuo Tan, Omer Ronen, Chandan Singh, and Bin Yu. Hierarchical shrinkage: improving the accuracy and interpretability of tree-based methods. *arXiv:2202.00858*, 2 2022. arXiv: 2202.00858. ↪9.

Afra Feyza Akyürek, Ekin Akyürek, Leshem Choshen, Derry Wijaya, and Jacob Andreas. Deductive closure training of language models for coherence, accuracy, and updatability. *arXiv preprint arXiv:2401.08574*, 2024. ↪10.

Ekin Akyürek, Bailin Wang, Yoon Kim, and Jacob Andreas. In-context language learning: Arhitectures and algorithms, 2024. ↪8.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016. ↪1.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations. *arXiv preprint arXiv:2305.18029*, 2023. ↪4.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2021. ↪4.

Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. "Will you find these shortcuts?" A protocol for evaluating the faithfulness of input salience methods for text classification. *arXiv preprint arXiv:2111.07367*, 2021. ↪4.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. GAN dissection: Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1811.10597*, 2018. ↪1.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023. ↪7.

Vinamra Benara, Chandan Singh, John X Morris, Richard Antonello, Ion Stoica, Alexander G Huth, and Jianfeng Gao. Crafting interpretable embeddings by asking llms questions. *arXiv preprint arXiv:2405.16714*, 2024. ↪9.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of Thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*, 2023. ↪7.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). *arXiv preprint arXiv:2402.10376*, 2024. ↪7.

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. LLMs as counterfactual explanation modules: Can ChatGPT explain black-box text classifiers? *arXiv preprint arXiv:2309.13340*, 2023. ↪6.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3889–3900, 2022. ↪6.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models, 2023. URL `https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html`. ↪4 and 7

Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. Science in the age of large language models. *Nature Reviews Physics*, pp. 1–4, 2023. ↪2 and 10

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984. ↪9.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. ↪1 and 5

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-SNLI: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31, 2018. ↪6.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1721–1730, 2015. ↪9.

Fateme Hashemi Chaleshtori, Atreya Ghosal, and Ana Marasović. On evaluating explanation utility for Human-AI decision-making in NLP. In *XAI in Action: Past, Present, and Future Applications*, 2023. ↪4.

Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and Swabha Swayamdipta. REV: information-theoretic evaluation of free-text rationales. *arXiv preprint arXiv:2210.04982*, 2022. ↪4.

Hugh Chen, Ian C Covert, Scott M Lundberg, and Su-In Lee. Algorithms to estimate shapley value feature attributions. *Nature Machine Intelligence*, pp. 1–12, 2023a. ↪6.

Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen McKeown. Do models explain themselves? Counterfactual simulatability of natural language explanations. *arXiv preprint arXiv:2307.08678*, 2023b. ↪6 and 10

Yanda Chen, Chandan Singh, Xiaodong Liu, Simiao Zuo, Bin Yu, He He, and Jianfeng Gao. Towards consistent natural-language explanations via explanation-consistency finetuning. *arXiv preprint arXiv:2401.13986*, 2024a. ↪10.

Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, and He He. Parallel structures in pre-training data yield in-context learning, 2024b. URL `https://arxiv.org/abs/2402.12530`. ↪8.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. ↪3.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does BERT look at? An analysis of BERT's attention. *arXiv preprint arXiv:1906.04341*, 2019. ↪7.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*, 2018. ↪6 and 7

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*, 2021. ↪7.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6309–6317, 2019. ↪7.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. ↪3.

Victor Dibia. LIDA: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. *arXiv preprint arXiv:2303.02927*, 2023. ↪8.

Finale Doshi-Velez and Been Kim. A roadmap for a rigorous science of interpretability. *arXiv preprint arXiv:1702.08608*, 2017. ↪1 and 4

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. Shortcut learning of large language models in natural language understanding. *Communications of the ACM (CACM)*, 2023. ↪4.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1, 2021. ↪6 and 8

Joseph Enguehard. Sequential integrated gradients: a simple but effective method for explaining language models. *arXiv preprint arXiv:2305.15853*, 2023. ↪6.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? *arXiv preprint arXiv:2310.17191*, 2023. ↪7.

Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020. ↪1.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? A case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022. ↪8.

Zelalem Gero, Chandan Singh, Hao Cheng, Tristan Naumann, Michel Galley, Jianfeng Gao, and Hoifung Poon. Self-verification improves few-shot clinical information extraction. *arXiv preprint arXiv:2306.00024*, 2023. ↪10.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. PatchScope: A unifying framework for inspecting hidden representations of language models, 2024. ↪7.

Mario Giulianelli, Jacqueline Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. *arXiv preprint arXiv:1808.08079*, 2018. ↪7.

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *arXiv preprint arXiv:1606.08813*, 2016. ↪1.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*, 2023. ↪6 and 8

Madeleine Grunde-McLaughlin, Michelle S Lam, Ranjay Krishna, Daniel S Weld, and Jeffrey Heer. Designing LLM chains by adapting techniques from crowdsourcing workflows. *arXiv preprint arXiv:2312.11681*, 2023. ↪9.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*, 2023. ↪7.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-augmented language model pre-training. *ArXiv*, abs/2002.08909, 2020. ↪6 and 7

Wooseok Ha, Chandan Singh, Francois Lanusse, Srigokul Upadhyayula, and Bin Yu. Adaptive wavelet distillation from neural networks through interpretations. *Advances in Neural Information Processing Systems*, 34:20669–20682, 2021. ↪1.

Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–318, 1986. ↪9.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *European conference on computer vision*, pp. 3–19. Springer, 2016. ↪6.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, et al. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*, 2022a. ↪8.

Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep visual features. In *International Conference on Learning Representations*, 2022b. ↪7.

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. Inspecting and editing knowledge representations in language models, 2023. ↪4.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. Benchmarking large language models as AI research agents. *arXiv preprint arXiv:2310.03302*, 2023a. ↪8.

Xu Huang, Jianxun Lian, Yuxuan Lei, Jing Yao, Defu Lian, and Xing Xie. Recommender AI agent: Integrating large language models for interactive recommendations. *arXiv preprint arXiv:2308.16505*, 2023b. ↪11.

Tal Ifargan, Lukas Hafner, Maor Kern, Ori Alcalay, and Roy Kishony. Autonomous llm-driven research from data to human-verifiable research papers. *arXiv preprint arXiv:2404.17605*, 2024. ↪8.

Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019. ↪6.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pp. 15696–15707. PMLR, 2023. ↪6 and 8

Cheongwoong Kang and Jaesik Choi. Impact of co-occurrence on factual knowledge of large language models. *arXiv preprint arXiv:2310.08256*, 2023. ↪4.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. ChatGPT for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023. ↪1.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–14, 2020. ↪5.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023. ↪3.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *arXiv preprint arXiv:1711.11279*, 2017. ↪4.

Siwon Kim, Jihun Yi, Eunji Kim, and Sungroh Yoon. Interpretation of nlp models through input marginalization. *arXiv preprint arXiv:2010.13984*, 2020. ↪6.

Satyapriya Krishna, Jiaqi Ma, Dylan Slack, Asma Ghandeharioun, Sameer Singh, and Himabindu Lakkaraju. Post hoc explanations of language models can improve language models. *arXiv preprint arXiv:2305.11426*, 2023. ↪4.

Nicholas Kroeger, Dan Ley, Satyapriya Krishna, Chirag Agarwal, and Himabindu Lakkaraju. Are large language models post hoc explainers? *arXiv preprint arXiv:2310.05797*, 2023. ↪6.

Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. Rethinking explainability as a dialogue: A practitioner's perspective. *arXiv preprint arXiv:2202.01875*, 2022. ↪5.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. The emergence of number and syntax units in lstm language models. *arXiv preprint arXiv:1903.07435*, 2019. ↪7.

Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*, 2022. ↪4.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson E. Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, John Kernion, Kamil.e Lukovsiut.e, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Samuel McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, T. J. Henighan, Timothy D. Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Janina Brauner, Sam Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *ArXiv*, abs/2307.13702, 2023. ↪7.

Dong-Ho Lee, Akshen Kadakia, Brihi Joshi, Aaron Chan, Ziyi Liu, Kiran Narahari, Takashi Shibuya, Ryosuke Mitani, Toshiyuki Sekiya, Jay Pujara, et al. XMD: An end-to-end framework for interactive explanation-based debugging of nlp models. *arXiv preprint arXiv:2210.16978*, 2022. ↪8.

Benjamin J Lengerich, Sebastian Bordt, Harsha Nori, Mark E Nunnally, Yin Aphinyanaphongs, Manolis Kellis, and Rich Caruana. LLMs understand glass-box models, discover surprises, and suggest repairs. *arXiv preprint arXiv:2308.01157*, 2023. ↪9.

Belinda Z Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. Eliciting human preferences with language models. *arXiv preprint arXiv:2310.11589*, 2023a. ↪11.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066*, 2015. ↪6.

Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. Table-GPT: Table-tuned gpt for diverse table tasks. *arXiv preprint arXiv:2310.09263*, 2023b. ↪8.

Xiang Lisa Li, Vaishnavi Shrivastava, Siyan Li, Tatsunori Hashimoto, and Percy Liang. Benchmarking and improving generator-validator consistency of language models. *arXiv preprint arXiv:2310.01846*, 2023c. ↪10.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*, 2023d. ↪3.

Tom Lieberum, Matthew Rahtz, János Kramár, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. Does circuit analysis interpretability scale? Evidence from multiple choice capabilities in chinchilla. *arXiv preprint arXiv:2307.09458*, 2023. ↪7.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *ArXiv*, abs/2307.03172, 2023a. ↪10.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, 2023b. ↪5 and 10

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 623–631, 2013. ↪9.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4768–4777, 2017. ↪1 and 6

Aman Madaan and Amir Yazdanbakhsh. Text and patterns: For effective chain of thought, it takes two to tango. *arXiv preprint arXiv:2209.07686*, 2022. ↪7.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback, 2023. ↪10.

Denis Jered McInerney, Geoffrey Young, Jan-Willem van de Meent, and Byron C Wallace. Chill: Zero-shot custom interpretable feature extraction from clinical notes with large language models. *arXiv preprint arXiv:2302.12343*, 2023. ↪9.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. Sources of hallucination by large language models on inference tasks. *arXiv preprint arXiv:2305.14552*, 2023. ↪8.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual knowledge in GPT. *arXiv preprint arXiv:2202.05262*, 2022. ↪4, 7, and 8

Rakesh R Menon, Kerem Zaman, and Shashank Srivastava. MaNtLE: Model-agnostic natural language explainer. *arXiv preprint arXiv:2305.12995*, 2023. ↪9.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Circuit component reuse across tasks in transformer language models. *arXiv preprint arXiv:2310.08744*, 2023. ↪7.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*, 2023. ↪10.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale, 2022. ↪4.

Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2019. ↪1.

Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017. ↪6.

John X Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M Rush. Text embeddings reveal (almost) as much as text. *arXiv preprint arXiv:2310.06816*, 2023a. ↪7.

John X Morris, Chandan Singh, Alexander M Rush, Jianfeng Gao, and Yuntian Deng. Tree prompting: efficient task adaptation without fine-tuning. *arXiv preprint arXiv:2310.14034*, 2023b. ↪9.

Jesse Mu and Jacob Andreas. Compositional explanations of neurons. *Advances in Neural Information Processing Systems*, 33: 17153–17163, 2020. ↪6 and 7

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of GPT-4. *arXiv preprint arXiv:2306.02707*, 2023. ↪4.

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 116(44):22071–22080, 2019. doi: 10.1073/pnas.1900654116. ↪1 and 3

Avanika Narayan, Ines Chami, Laurel Orr, Simran Arora, and Christopher Ré. Can foundation models wrangle your data? *arXiv preprint arXiv:2205.09911*, 2022. ↪8.

Maxwell Nye, Anders Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models. *ArXiv*, abs/2112.00114, 2021. ↪7.

Tuomas Oikarinen and Tsui-Wei Weng. Clip-dissect: Automatic description of neuron representations in deep vision networks. *arXiv preprint arXiv:2204.10965*, 2022. ↪7.

Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023. ↪7.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022. ↪6 and 8

OpenAI. GPT-4 technical report, 2023. ↪1, 3, and 7

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. ↪3.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*, 2023. ↪10.

Letitia Parcalabescu and Anette Frank. On measuring faithfulness of natural language explanations. *arXiv preprint arXiv:2311.07466*, 2023. ↪4.

Ajay Patel, Delip Rao, Ansh Kothary, Kathleen McKeown, and Chris Callison-Burch. Learning interpretable style embeddings via prompting llms. *arXiv preprint arXiv:2305.12696*, 2023. ↪9.

Ravi Patel, Angus Brayne, Rogier Hintzen, Daniel Jaroslawicz, Georgiana Neculae, and Dane Corneil. Retrieve to explain: Evidence-driven predictions with language models. *arXiv preprint arXiv:2402.04068*, 2024. ↪10.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Lidén, Zhou Yu, Weizhu Chen, and Jianfeng Gao. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *ArXiv*, abs/2302.12813, 2023. ↪7.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, and Mohit Iyyer. TopicGPT: A prompt-based topic modeling framework. *arXiv preprint arXiv:2311.01449*, 2023. ↪10.

Luca Pion-Tonachini, Kristofer Bouchard, Hector Garcia Martin, Sean Peisert, W Bradley Holtz, Anil Aswani, Dipankar Dwivedi, Haruko Wainwright, Ghanshyam Pilania, Benjamin Nachman, et al. Learning from learning machines: a new generation of AI technology to meet the needs of science. *arXiv preprint arXiv:2111.13786*, 2021. ↪2 and 10

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models, 2022. ↪7.

J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986. ↪9.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021. ↪5.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! Leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*, 2019. ↪6.

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. Toward transparent AI: A survey on interpreting the inner structures of deep neural networks. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 464–483. IEEE, 2023. ↪2 and 7

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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*, pp. 1135–1144. ACM, 2016. ↪1 and 6

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 2023. ↪10.

Cynthia Rudin. Please stop explaining black box models for high stakes decisions. *arXiv preprint arXiv:1811.10154*, 2018. ↪9.

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *arXiv preprint arXiv:2103.11251*, 2021. ↪1.

Lisa Schut, Nenad Tomasev, Tom McGrath, Demis Hassabis, Ulrich Paquet, and Been Kim. Bridging the Human-AI knowledge gap: Concept discovery and transfer in alphazero. *arXiv preprint arXiv:2310.16410*, 2023. ↪11.

Sarah Schwettmann, Tamar Rott Shaham, Joanna Materzynska, Neil Chowdhury, Shuang Li, Jacob Andreas, David Bau, and Antonio Torralba. FIND: A function description benchmark for evaluating interpretability methods. *arXiv e-prints*, pp. arXiv–2309, 2023. ↪4.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*, 2023. ↪10.

Pratyusha Sharma, Jordan T Ash, and Dipendra Misra. The truth is in there: Improving reasoning in language models with layer-selective rank reduction. *arXiv preprint arXiv:2312.13558*, 2023. ↪8.

Sandipan Sikdar, Parantapa Bhattacharya, and Kieran Heese. 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)*, pp. 865–878, 2021. ↪6.

Chandan Singh, Armin Askari, Rich Caruana, and Jianfeng Gao. Augmenting interpretable models with large language models during training. *Nature Communications*, 14(1):7913, 2023a. ↪9.

Chandan Singh, Aliyah R Hsu, Richard Antonello, Shailee Jain, Alexander G Huth, Bin Yu, and Jianfeng Gao. Explaining black box text modules in natural language with language models. *arXiv preprint arXiv:2305.09863*, 2023b. ↪4 and 6

Chandan Singh, John X. Morris, Jyoti Aneja, Alexander M. Rush, and Jianfeng Gao. Explaining patterns in data with language models via interpretable autoprompting, 2023c. ↪10.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023. ↪3.

Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. Talktomodel: Understanding machine learning models with open ended dialogues. *arXiv preprint arXiv:2207.04154*, 2022. ↪6, 8, and 11

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *ICML*, 2017. ↪6.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. *arXiv preprint arXiv:2009.10795*, 2020. ↪8.

Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. Distill-and-Compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 303–310, 2018. ↪1.

Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. Quantifying uncertainty in natural language explanations of large language models. *arXiv preprint arXiv:2311.03533*, 2023. ↪6.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022. ↪5 and 10

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996. ↪9.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 2024. ↪10.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. ↪1, 3, and 5

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*, 2024. ↪11.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*, 2023. ↪10.

Berk Ustun and Cynthia Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102: 349–391, 2016. ↪9.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001*, 2022a. ↪7.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023a. ↪1, 2, and 10

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. *arXiv preprint arXiv:2211.00593*, 2022b. ↪7.

Qianli Wang, Tatiana Anikina, Nils Feldhus, Josef van Genabith, Leonhard Hennig, and Sebastian Möller. LLMCheckup: Conversational examination of large language models via interpretability tools. *arXiv preprint arXiv:2401.12576*, 2024. ↪6 and 8

Zihan Wang, Jingbo Shang, and Ruiqi Zhong. Goal-driven explainable clustering via language descriptions. *arXiv preprint arXiv:2305.13749*, 2023b. ↪10.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. ↪6 and 7

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023. ↪7.

Daniel S Weld and Gagan Bansal. The challenge of crafting intelligible intelligence. *Communications of the ACM*, 62(6):70–79, 2019. ↪5.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*, 2019. ↪6.

Theodora Worledge, Judy Hanwen Shen, Nicole Meister, Caleb Winston, and Carlos Guestrin. Unifying corroborative and contributive attributions in large language models. *arXiv preprint arXiv:2311.12233*, 2023. ↪6, 7, and 10

Zhengxuan Wu, Atticus Geiger, Christopher Potts, and Noah D. Goodman. Interpretability at scale: Identifying causal mechanisms in Alpaca. *ArXiv*, abs/2305.08809, 2023. ↪7.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. *arXiv preprint arXiv:2306.13063*, 2023. ↪6.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. Large language models for automated open-domain scientific hypotheses discovery. *arXiv preprint arXiv:2309.02726*, 2023. ↪10.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023. ↪7.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, pp. 100211, 2024. ↪3.

Xi Ye and Greg Durrett. The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*, 35:30378–30392, 2022. ↪6 and 10

Xi Ye and Greg Durrett. Explanation selection using unlabeled data for chain-of-thought prompting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 619–637, 2023. ↪4.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015. ↪1.

Muhammad Bilal Zafar, Philipp Schmidt, Michele Donini, Cédric Archambeau, Felix Biessmann, Sanjiv Ranjan Das, and Krishnaram Kenthapadi. More than words: Towards better quality interpretations of text classifiers. *arXiv preprint arXiv:2112.12444*, 2021. ↪6.

Han Zhang, Xumeng Wen, Shun Zheng, Wei Xu, and Jiang Bian. Towards foundation models for learning on tabular data. *arXiv preprint arXiv:2310.07338*, 2023a. ↪8.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. Tell your model where to attend: Post-hoc attention steering for LLMs. *arXiv preprint arXiv:2311.02262*, 2023b. ↪8 and 10

Tianping Zhang, Shaowen Wang, Shuicheng Yan, Jian Li, and Qian Liu. Generative table pre-training empowers models for tabular prediction. *arXiv preprint arXiv:2305.09696*, 2023c. ↪8.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *arXiv preprint arXiv:2309.01029*, 2023. ↪2.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. ↪4.

Ruiqi Zhong, Charlie Snell, Dan Klein, and Jacob Steinhardt. Describing differences between text distributions with natural language. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27099–27116. PMLR, 17–23 Jul 2022. ↪10.

Ruiqi Zhong, Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, and Jacob Steinhardt. Goal driven discovery of distributional differences via language descriptions. *ArXiv*, abs/2302.14233, 2023. ↪4 and 10

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022. ↪7.

Hattie Zhou, Arwen Bradley, Etai Littwin, Noam Razin, Omid Saremi, Josh Susskind, Samy Bengio, and Preetum Nakkiran. What algorithms can transformers learn? A study in length generalization. *arXiv preprint arXiv:2310.16028*, 2023. ↪8.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. Relying on the unreliable: The impact of language models' reluctance to express uncertainty, 2024. ↪6.

Zhaocheng Zhu, Yuan Xue, Xinyun Chen, Denny Zhou, Jian Tang, Dale Schuurmans, and Hanjun Dai. Large language models can learn rules. *arXiv preprint arXiv:2310.07064*, 2023. ↪10.

Zhiying Zhu, Weixin Liang, and James Zou. Gsclip: A framework for explaining distribution shifts in natural language. *arXiv preprint arXiv:2206.15007*, 2022. ↪10.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*, 2023. ↪1 and 2

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *ArXiv*, abs/2310.01405, 2023. ↪6 and 7