On Behalf of the Stakeholders: Trends in NLP Model Interpretability in the Era of LLMs

Nitay Calderon and **Roi Reichart**
Faculty of Data and Decision Sciences, Technion
nitay@campus.technion.ac.il and roiri@technion.ac.il

Abstract

Recent advancements in NLP systems, particularly with the introduction of LLMs, have led to widespread adoption of these systems by a broad spectrum of users across various domains, impacting decision-making, the job market, society, and scientific research. This surge in usage has led to an explosion in NLP model interpretability and analysis research, accompanied by numerous technical surveys. Yet, these surveys often overlook the needs and perspectives of explanation stakeholders. In this paper, we address three fundamental questions: Why do we need interpretability, what are we interpreting, and how? By exploring these questions, we examine existing interpretability paradigms, their properties, and their relevance to different stakeholders. We further explore the practical implications of these paradigms by analyzing trends from the past decade across multiple research fields. To this end, we retrieved thousands of papers and employed an LLM to characterize them. Our analysis reveals significant disparities between NLP developers and non-developer users, as well as between research fields, underscoring the diverse needs of stakeholders. For example, explanations of internal model components are rarely used outside the NLP field. We hope this paper informs the future design, development, and application of methods that align with the objectives and requirements of various stakeholders.

1 Introduction

Recent advancements in Natural Language Processing (NLP), particularly with the introduction of Large Language Models (LLMs), have dramatically enhanced model performance. These models are now capable of executing a wide array of tasks and have been adopted across various domains and research fields (Aletras et al., 2016; Calderon et al., 2024; Yang et al., 2024). Their applications extend beyond the NLP community, and they are widely used by the general public (Choudhury and Shamszare, 2023; Kasneci et al., 2023; von Garrel and Mayer, 2023). However, these black-box models are complex and opaque (Wallace et al., 2019; Calderon et al., 2023; Luo et al., 2024). While performance has advanced, this comes at the cost of understanding their underlying mechanisms (Lyu et al., 2022; Madsen et al., 2023; Singh et al., 2024).

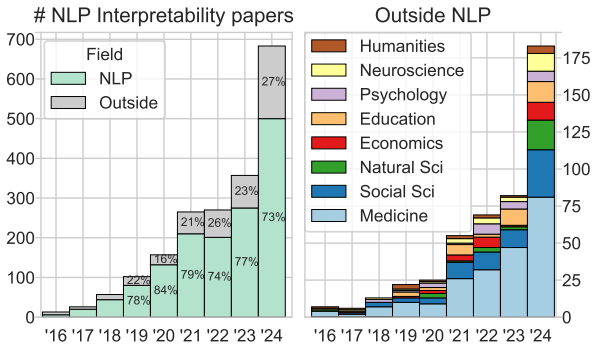

Figure 1: Number of *NLP Interpretability* papers published over time. Each year spans from June of the previous year to the following June. The left plot shows the distribution of papers across NLP and the other fields (*Outside*). The right plot shows trends in other fields besides NLP. Only papers that use, propose, or discuss interpretability methods applied to natural language are counted, following relevance filtering by an LLM.

The ability to explain decisions is particularly crucial, given that NLP models, especially LLMs, significantly influence individual decision-making (Tu et al., 2024; Yu et al., 2024), society (Samuel, 2023; Taubenfeld et al., 2024), the job market (Eloundou et al., 2023), and academic research (Editorials, 2023; Liang et al., 2024). Moreover, model interpretability and analysis are utilized for scientific insights and discoveries (Roscher et al., 2020; Allen et al., 2023; Badian et al., 2023; Birhane et al., 2023; Lissak et al., 2024b).

Unsurprisingly, research on model interpretability and analysis has become one of the most prolific areas within the NLP community and beyond, yielding thousands of publications in recent years, as illustrated in Figure 1. Consequently, many tech-

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 656–693

nical NLP model interpretability and analysis surveys have emerged, reviewing hundreds of methods (Belinkov and Glass, 2019; Danilevsky et al., 2020; Balkir et al., 2022; Sajjad et al., 2022a; Bereska and Gavves, 2024; Luo et al., 2024; Zhao et al., 2024; Mosbach et al., 2024). In this paper, we aim to bridge a gap in the existing literature and discuss model interpretability from the stakeholders' perspective. Our goals are to broaden the NLP community's point of view on the application of interpretability methods in various fields and to promote the design and development of methods that align with the objectives, expectations, and requirements of various stakeholders.

We will explore three key questions: why do we need interpretability (§2), what are we interpreting (§3), and how are we interpreting (§4)? This approach allows us to examine common interpretability paradigms (Table 1), their properties and their applications by different stakeholders.

We start by presenting four perspectives on interpretability and their relevant stakeholders in §2. Next, in §3, we address a pressing issue in the literature: the lack of consensus on the definition of interpretability. We examine various definitions within and outside the NLP community and propose a broad definition: *Extracting insights into a mechanism of the NLP system and communicating them to the stakeholders in understandable terms*.

In §4, we propose six properties of interpretability methods and discuss the relevance of each property to different stakeholders. For example, the *scope* property distinguishes between local and global explanations. If the stakeholder is a physician, a local explanation that clarifies the prediction for an individual patient is preferred. Conversely, a global explanation is more suitable for a scientist, as it facilitates understanding broader phenomena.

We survey in Appendix §B seven prevalent interpretability paradigms, explain which properties characterize each (see Table 1), and discuss their applications by different stakeholders. Throughout the survey, we review over 200 works.

Following that, in §5 we aim to understand how the paradigms and their properties are reflected in practice by analyzing trends over the years and across different research fields. To this end, we retrieved over 14,000 papers using the Semantic Scholar API and employed an LLM to select only relevant ones, resulting in 2,000 papers. Furthermore, we utilized the LLM to annotate papers with their interpretability paradigm and properties.[1] Importantly, the LLM annotation is in strong agreement (over 90%) with human expert annotation. To the best of our knowledge, this is the first successful application of an LLM for such a task.

Below, we summarise our main findings:

1. Within the NLP community, interpretability paradigm trends have remained stable over the decade. However, the introduction of LLMs in the past two years has prompted a notable shift.
2. Outside the NLP community, our main claim gains support: different stakeholders have varying needs, reflected in significant differences between research fields in terms of both the paradigms and their properties.
3. Comparing NLP developers to non-developers reveals that the latter group is less interested in understanding internal model components.
4. Non-developers opt for popular methods not originally developed within the NLP community, such as SHAP and LIME, likely due to their user-friendly and easy-to-apply software.
5. LLMs have shifted the trends in interpretability research: not only has the number of published papers doubled, but there has also been a substantial increase in the use of natural language explanations. These explanations are utilized in nearly half of the papers outside the NLP field.

We hope this first-of-its-kind paper, which reviews NLP interpretability through the stakeholders' perspective and rigorously analyzes trends within and outside the NLP field, will pave the way for improved design, development, and application of these essential methods. To further this aim, we outline in §6 practical steps that NLP researchers can undertake to promote the adoption of interpretability methods in other disciplines.

2 Why Do We Need Interpretability?

Understanding why interpretability is necessary provides a solid framework for discussing, assessing, and enhancing interpretability methods, ensuring they meet practical objectives and expectations. So, when and why do we need interpretability? We gather ideas from other surveys (Gade et al., 2020; Räuker et al., 2023; Saeed and Omlin, 2023) and propose a decomposition of the need for interpretability into four perspectives (see Figure 2): *algorithmic, business, scientific, and social.*

[1] Data: www.github.com/nitaytech/InterpreTrends

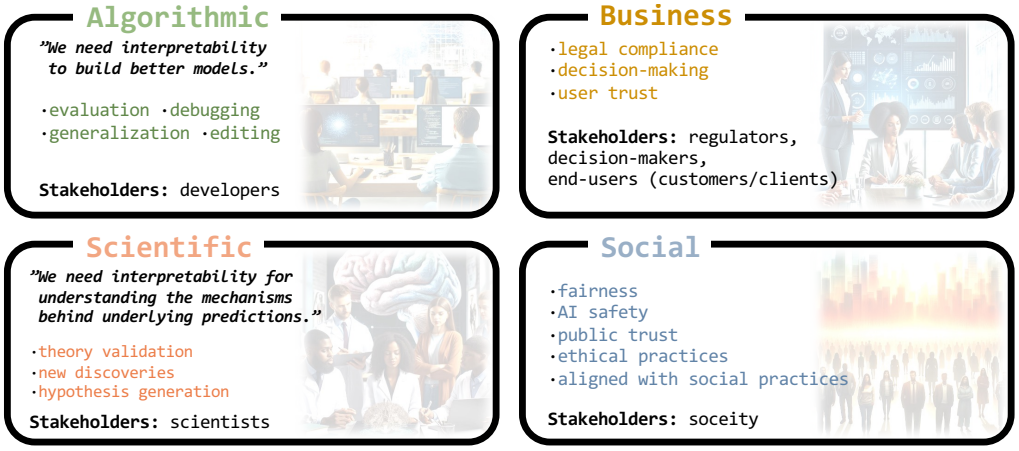

Figure 2: Overview of four perspectives on the need for interpretability proposed in this paper.

The four perspectives define the objective and use case of the interpretability method. Clearly, there can be overlaps between the different perspectives, particularly with the algorithmic one. For example, using interpretability to build a better model (algorithmic perspective) might coincide with making it fairer (social perspective) or one that promotes more informed business decisions (business perspective). Similarly, promoting social values through interpretability (social perspective) can build customer trust (business perspective).

Besides the objectives of the interpretability method, another key consideration is the *stakeholders*–the audience to whom the explanation is aimed and communicated. Accordingly, when designing the interpretability method, we should consider not only the objective (and the usage) of the explanation but also the stakeholders, their level of expertise, and their familiarity with NLP models. By identifying different stakeholders' specific requirements and concerns, we can foster practical interpretability methods that align with their expectations (Kaur et al., 2021). We next discuss the four perspectives and the main stakeholders (**in bold**):

1. The Algorithmic Perspective: emphasizes using interpretability for building better models. Thus, the stakeholders are **developers**. Interpretability allows for an open-ended, more rigorous evaluation beyond standard metrics (Ribeiro et al., 2018; Lertvittayakumjorn and Toni, 2021; Kabir et al., 2024). It helps uncover why the model fails, offering insights into identifying and rectifying mistakes (Yao et al., 2021) and improving its generalization. For instance, Ghaeini et al. (2019) use saliency maps, and Joshi et al. (2022) use counterfactual explanations for modifying the training objective and improving model robustness. Gekhman et al. (2024) study the source of hallucinations in LLMs by curating a diagnostic set that utilizes the model's pre-existing knowledge. Moreover, by understanding how the model works, we can intervene and modify it or design better models from the start (e.g., reverse engineering) (Meng et al., 2022; Arad et al., 2023). For example, Dai et al. (2022) use attributions to locate knowledge neurons, modify them, and edit factual knowledge of the NLP model. The algorithmic perspective underscores interpretability for debugging, refining model deployment, and forecasting progress.

2. The Business Perspective: focuses on leveraging interpretability across various sectors to enhance informed decision-making, legal compliance, and user trust. Models often aid decision-making at the **business** level (e.g., sentiment analysis for market research (Hartmann et al., 2022)) and at the **user** level (e.g., LLMs assisting **physicians** in **patient** diagnostics (Clusmann et al., 2023)). In both cases, interpretability aids ensure well-grounded and trustworthy decisions (Lai and Tan, 2019).

Legal compliance includes cases where interpretability is explicitly the **regulator's** requirement, such as the *GDPR's "right to explanation"* (Goodman and Flaxman, 2017) and the *Algorithmic Accountability Act* proposed in the US (MacCarthy, 2019), or cases where interpretability is instrumental in ensuring the **business** adheres to legal standards, thereby reducing the risk of legal penalties. For instance, when NLP models are used to process credit applications (Zhang et al., 2020; Yang et al., 2022; Sanz-Guerrero and Arroyo, 2024), they must comply with the *Equal Credit Opportunity Act (ECOA)*, which prohibits discrimination.

Finally, interpretability enhances transparency, fostering trust and goodwill. When **end-users** understand how decisions are made, they are more likely to trust AI systems, improving **business** reputation. For example, Facebook's 'Why am I seeing this ad' tool is designed specifically to provide more transparency and build trust (Pavón, 2023). The link between interpretability and trust is well-documented (Parasuraman and Riley, 1997; Miller et al., 2016; Buçinca et al., 2020).

3. The Scientific Perspective: Language is tightly connected to human behavior, cognition, and communication. **Researchers** and **scientists** from various disciplines, such as social science (Lazer et al., 2020; Ziems et al., 2024), psychology (Ophir et al., 2022), psychiatry (Rezaii et al., 2022), psycholinguistics (Wilcox et al., 2018), health (Singhal et al., 2023; Thirunavukarasu et al., 2023), neuroscience (Goldstein et al., 2022; Tikochinski et al., 2023), finance (El-Haj et al., 2019), behavioral economics (Shapira et al., 2023, 2024), political science (Gennaro and Ash, 2022), and beyond, are now turning to NLP to model scientific phenomena, decode complex patterns and derive meaningful insights about humanity. Science is all about gaining knowledge, and interpretability enables us to understand the underlying mechanisms and patterns the NLP model identifies, facilitating deeper comprehension and advancing scientific discoveries (Roscher et al., 2020). For example, by interpreting the representations of Facebook posts extracted by an NLP model, Lissak et al. (2024b) identify a new risk factor for suicide ideation: boredom.

4. The Social Perspective: addresses the broader impact of NLP systems on society, fairness, the ethical implications of its use and AI safety. Since NLP models are optimized using data that may contain human biases and prejudices (Blodgett et al., 2020; Dev et al., 2021), interpretability is crucial for understanding the rationale behind the models, ensuring they serve what they are designed for rather than reflecting their training data (Ruder et al., 2022). Interpretability can confirm the model predictions are just and equitable (Orgad et al., 2022; Attanasio et al., 2023; Santosh et al., 2024), foster public trust, promote ethical practices, and prevent misuse or other harmful consequences (Bereska and Gavves, 2024; Lissak et al., 2024a). Accordingly, interpretability helps **society** embrace the model or reject it, depending on how well it aligns with expected social values.

3 Definitions

3.1 What is an Interpretability Method?

In the AI literature, the terms *interpretability* and *explainability* are often subjects of debate, and there is no clear consensus on their definitions (Doshi-Velez and Kim, 2017; Lipton, 2018; Krishnan, 2019). While these terms are used interchangeably in much of the NLP literature (Jacovi and Goldberg, 2020; Lyu et al., 2022; Zhao et al., 2024), many papers in the XAI literature distinguish between the two (Rudin, 2018; Arrieta et al., 2020), see our note in §A.2.1. Moreover, within this broad umbrella of model interpretability, the NLP literature also discusses model analysis (Belinkov and Glass, 2019; Mosbach et al., 2024).

For the purposes of this paper, we embrace a broad perspective and define both *interpretability and explainability methods* as:

> **Interpretability Method**
> *Any approach that extracts insights into a mechanism of the NLP system.*

We justify this broad definition, which explicitly encompasses model analysis, because our paper focuses on the perspective of stakeholders for whom, to some extent, analysis alone may suffice to achieve their objectives. For instance, a regulator might only need to ensure that model performance does not significantly differ between two subpopulations. This does not necessarily demand that the interpretation elucidate the precise cause of each decision. Moreover, our broad definition does not restrict the interpretability method to explain the full system, but rather, only a mechanism within it. For example, developers might want to gain insights about specific components of the system to improve or modify their functionality.

3.2 What is an Explanation?

Miller (2017) and Lipton (2018) rightfully emphasize that interpretability should not be confused with an explanation. Miller (2017) distinguishes between (causal) attributions and (causal) explanations. Attribution involves extracting relationships and causes, but it is not necessarily an explanation, even if a person could use attributions to derive their own explanation. Explanation also involves selecting, contextualizing, and presenting causes and relationships to the stakeholders. Thus, explanations are about communicating insights in a

	Paradigm	Examples	Mechanism	Scope	Time	Access	Presentation
§B.1	Feature Attributions	Perturbations, Gradients, Propogations, Surrogate (**LIME/SHAP**), Attentions	input-output, concept-output	local	post-hoc	specific, agnostic	scores, visualiztion
§B.2	Probing	Probing and **Clustering**	input-internal	global	post-hoc	specific	scores
§B.3	Mechanistic Interpretability	Stimuli, Sparse Autoencoders, Patching, Scrubbing, Logits lens	internal-internal	global	post-hoc	specific	visualiztion, text
§B.4	Diagnostic Sets	Challenge/Probing sets, Test suites	input-output	global	post-hoc	agnostic	scores
§B.5	Counterfactuals	Contrastive examples, **Adversarial attacks**, Concept counterfactuals	input-output, concept-output	local, global	post-hoc	specific, agnostic	scores, examples
§B.6	Natural Lang. Explanations	Extractive, Abstractive, explain-then-predict, predict-and-explain, CoT	input-output	local	intrinsic	specific	text
§B.7	Self-explaining Models	**Classic ML**, Concept bottleneck, KNN-based, Neural module nets	input-output, input-concept-output	local, global	intrinsic	specific	scores, examples, text

Table 1: Overview of the interpretability paradigms discussed in this paper, categorised by their *what* and *how* properties (§4). A detailed survey of these paradigms is provided in §B. **In bold**, methods (SHAP, LIME, Clustering, Adversarial Attacks, Classic ML) that were analyzed separately of their paradigm in our trend analysis in §5.

way that aligns with human cognitive biases and social expectations. In some sense, the output of interpretability methods is an attribution.

Most existing work in the NLP literature is on *how we extract* insights and not about *communicating* them. Since this paper is directed at this community rather than the HCI or XAI communities, we mostly focus on interpretability methods. However, to begin the discussion about the *what* and *how* parts (see the paragraph below the following definition), we must first define an explanation. This is because the *what* and *how* are derived from the *why* – the stakeholders, and clearly, they are part of an explanation. To this end, we have gathered common (though not formal) definitions from seminal works in the literature (Doshi-Velez and Kim, 2017; Lipton, 2018; Murdoch et al., 2019; Arrieta et al., 2020; Lyu et al., 2022; Räuker et al., 2023), and propose the following definition:

> **Explanation (explaining):**
> *Extracting insights into a mechanism of the NLP system and communicating them to the stakeholders in understandable terms.*

We define and elaborate on the **mechanism** and **understandable terms** aspects of the above definition in Appendix §C. These two aspects are related to the *what* part: *what mechanism are we interpreting, what terms are we using to describe its states, and what is the scope of the explanation?*

Conversely, the **extracting** and **communicating** aspects are related to the *how* part: *how are we interpreting and extracting insights and how are we presenting and communicating insights?* Note that **extracting** is essentially the interpretability method defined in §3.1.

To summarize, an interpretability (or explainability or analysis) method extracts insights from a model, whereas an explanation involves communicating these insights to stakeholders. This process includes filtering and selecting relevant insights, processing them, and presenting them in an understandable terms. For example, computing SHAP values is an interpretability method, while visualizing these values using the SHAP Python package[2] and providing guidance on interpreting these visualizations constitute an explanation.

4 Properties and Categorization

In this section, we propose and briefly describe properties that answer the *what* and *how* questions derived from our interpretability definitions. In Table 1, we present a categorization of interpretability paradigms based on the properties. In Appendix §A, we thoroughly examine the properties and discuss their alignment with the objectives, requirements, and expectations of various stakeholders.

[*what*] **Explained mechanism §A.1.1:** Interpretability methods can explain different mechanisms of the NLP system. While most methods explain the whole system (an ***input-output*** mechanism), other methods explain input representations (an ***input-internal*** mechanism) or internal components such as neurons, attention heads, circuits, and more (an ***internal-internal*** mechanism). In addition, this property covers any abstraction of the

[2] https://shap.readthedocs.io

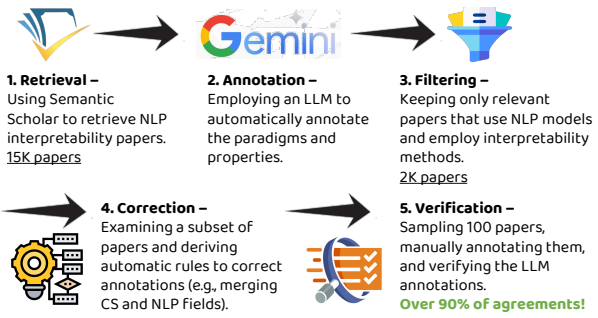

Figure 3: An illustration of our five-stage procedure for annotating NLP interpretability papers, with the stages fully detailed in Appendix §D.

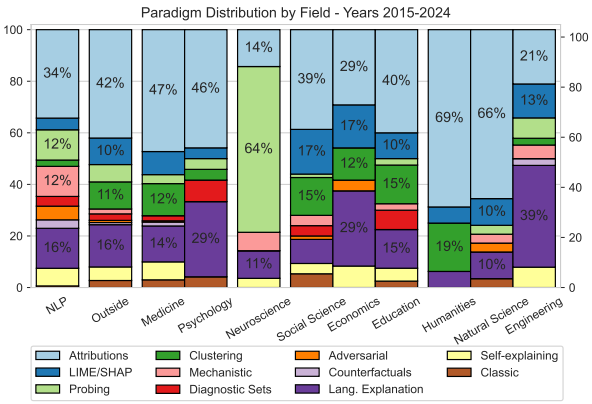

Figure 4: Distribution of NLP interpretability paradigms by research field, including papers in years 2015-24.

mechanism states (see §C.2), for example, explaining the impact of concepts conveyed in the text instead of explaining long and complex raw input. In this case, which is thoroughly discussed in §A.1.2, the explained mechanism is ***concept-output***.

[*what*] **Scope §A.1.3:** Determined by whether the explanation is ***local*** – describes the mechanism for an individual input instance, or ***global*** – describes the mechanism for the entire data distribution.

[*how*] **Time §A.2.1:** Determined by the time the explanation is formed. ***Post-hoc*** methods produce explanations after the prediction, while ***intrinsic*** methods are built-in: the explanation is generated during the prediction, and the model relies on it.

[*how*] **Access §A.2.2:** Determined by accessibility requirement to the explained model. ***Model-agnostic*** methods can only access its inputs and outputs, while ***model-specific*** methods require access to the explained model during the training time of the interpretability method and can access its internal components or representations.

[*how*] **Presentation §A.2.3:** Determined by how insights extracted by the interpretability method are presented to the stakeholder. This includes ***scores***, such as importance scores or metrics, and ***visualization***, such as heatmaps and graphs. Other explanations present similar or contrastive ***examples*** to stakeholders or communicate insights through ***texts*** written in natural language.

[*how*] **Causal-based §A.2.4:** Providing faithful explanations might involve incorporating techniques from the causality literature. This property determines whether the method is ***causal-based*** or ***not***.

5 Trends in Model Interpretability

In this section, we analyze trends over the last decade in papers that propose or employ an interpretability method in the NLP field or fields outside of NLP. The analysis covers trends in interpretability method paradigms and their properties.

5.1 Data

Our data collection process consists of five stages and is illustrated in Figure 3. In the first stage, we utilized the Python client[3] of the Semantic Scholar API[4] to retrieve 14,676 NLP interpretability papers by searching queries such as `NLP interpretability` (a full list of queries is provided in Box D.1). Subsequently, we employed an LLM (`gemini-1.5-pro-preview-0514`)[5] to determine the relevance of each paper based on its title and abstract. A paper is considered relevant if it relates to NLP research, employs NLP methods or models with text input, and proposes, utilizes, or discusses an interpretability method. After relevancy filtering, 2,009 papers remained (see Figure 1 for their distribution across fields).

In addition, we used the LLM to annotate various attributes, including the research field, whether an LLM is employed, the paradigm of the interpretability method and its mechanism, scope, accessibility and whether it is causal-based or not. The zero-shot prompt is provided in Box D.4. See Appendix §D for additional details about our retrieval and annotation processes.

To verify the LLM annotations, we randomly sampled 100 papers, which one of the authors manually annotated. The agreement statistics are presented in Table 4. Notably, 96% of the papers the LLM annotated as relevant were indeed relevant. Furthermore, over 90% of the annotations across

[3] www.github.com/danielnsilva/semanticscholar
[4] www.semanticscholar.org/product/api
[5] www.ai.google.dev/#gemini-api

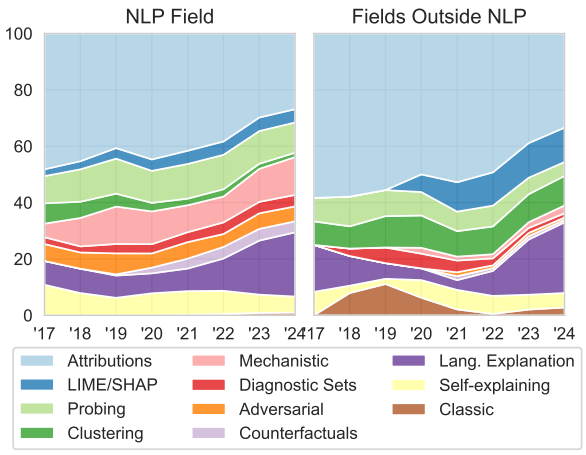

Figure 5: Trends in NLP interpretability paradigms over time in the NLP field (left plot) and in fields outside of NLP (right plot). The plots show the percentages of papers for each paradigm, as predicted by an LLM. The data smoothed using a one-year moving average.

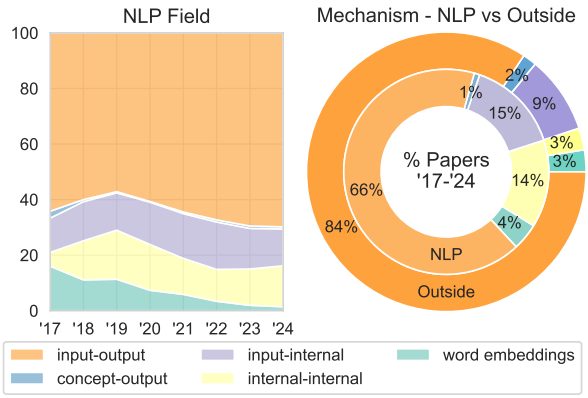

Figure 6: Trends in the explained mechanism. The left plot presents trends over time in the NLP field, showing the percentages of papers for each mechanism, as predicted by an LLM. The right plot presents pie charts with the percentage distribution of the mechanisms: the inner pie shows the distribution within the NLP field, and the outer pie shows for fields outside of NLP.

each property were correct. When excluding annotations labeled as '*unknown*' (e.g., where the LLM indicated the method scope was unknown, but sufficient domain knowledge could infer it), over 95% of the annotations were correct. To the best of our knowledge, this is the first paper to utilize an LLM successfully for such a task.

5.2 Results

We present the results in the following figures and tables, all illustrating trends in the NLP field and external fields, thereby emphasizing differences between developers and non-developer stakeholders.[6]

(1) Figure 1 in §1 presents the number of interpretability papers by research field and year.[7] (2) Figure 4 displays the distribution of interpretability method paradigms across each field, while (3) Figure 5 illustrates trends over the last decade. (4) Figure 6 presents the distribution of the explained mechanisms, and (5) Table 2 reports statistics on method properties. (6) Table 3 emphasizes trends between papers that employ LLMs and those that do not. Finally, (7) Table 5 in the appendix provides the absolute number of papers and average citations for each paradigm.

Below we discuss our key findings:

Inside: Stable trends in the NLP community. Figure 5 shows that paradigm trends within the NLP community are generally stable over time. However, two leading paradigms, Feature Attributions and Natural Language Explanations, demonstrate contrasting trends: the proportion of Feature Attribution papers has gradually decreased (from ~45% in 2017 to ~30% in 2024) while papers on Natural Language Explanations have increased (from ~10% in 2017 to ~25% in 2024). The latter rise is likely attributed to advancements in text generation capabilities, which will be discussed later. The next two most common paradigms—Probing and Mechanistic Interpretability, each account for about 12% (see Figure 4).

Regarding the trends in mechanisms illustrated in Figure 6, the explanation of Word Embedding, which was very popular a decade ago, has diminished over the years. Currently, two-thirds of the papers explain the input-output mechanism.

Inside vs Outside: Non-developers care less about model internals. We observe notable differences when comparing paradigm distributions between the NLP field and outside of NLP. While Feature Attribution is the dominant paradigm in both, Mechanistic Interpretability and Adversarial Attacks hold a large share within NLP but are rarely seen outside of it. Conversely, Clustering and Surrogate Models (such as LIME and SHAP) are common outside of NLP but not frequently encountered in general NLP papers.

We attribute these distinctions to two main reasons. The first reason is that non-developers care less about model internals and are more concerned with input-output mechanisms. This is evident in

[6]While developers may be stakeholders in fields outside of NLP, and vice versa, the primary distinction remains applicable. Most stakeholders in NLP are developers, while those in other fields are typically non-developers.

[7]Note that each year spans from June of the previous year to the following June.

	Scope		Accessibility		Causal-based	
	local	global	specific	agnostic	causal	not
NLP	57.3	42.7	84.6	15.4	5.2	94.8
Outside	61.7	38.3	80.1	19.9	1.9	98.1
↪ Healthcare	66.5	33.5	76.7	23.3	2.0	98.0
↪ Neuroscience	25.0	75.0	92.6	7.4	0.0	100
↪ Social	57.7	42.3	76.9	23.1	2.0	98.0

Table 2: Percentage of papers by properties (§4) across fields. *Outside* encompasses all fields outside NLP and CS. *Healthcare* includes Medicine and Psychology, while *Social* includes Social Sciences, Economics and Education fields.

the right plot of Figure 6, where there are five times more internal-internal mechanism papers in the NLP field. Moreover, although 9% of the papers outside of NLP explain an input-internal mechanism (representations), most involve field-specific techniques. For example, Probing is the most common paradigm in the neuroscience field (64% of the papers, see Figure 4), where researchers try to align model representations with brain activities (Goldstein et al., 2022; Tikochinski et al., 2023).

The second reason is the ease of application and the level of support for these methods in popular code packages. These aspects are particularly important for non-developers. For instance, LIME and SHAP packages are widely used across many domains beyond NLP (Kaur et al., 2020), and clustering or classic ML methods are readily available in popular data science packages like Scikit-learn.

Outside NLP: Different fields, different needs. The choice of interpretability method depends on the stakeholder's objectives and needs. Different research fields have distinct requirements, as clearly shown in Figure 4, where paradigm distributions vary across the fields. These differing needs are also reflected in method properties in Table 2. For instance, in healthcare fields, local explanations are much more prominent. This makes sense considering that the main stakeholders, patients and therapists, are interested in understanding individual decisions. Conversely, in neuroscience and social science, scientists aim to understand cognitive mechanisms or social phenomena, thus preferring global explanations.

LLMs dramatically change the trends. The introduction of LLMs in the last two years has drastically improved the capabilities of NLP models. These models have been widely adopted not only by NLP researchers but also by practitioners in various fields. This is evident in Table 3, where

		NLP		Outside	
		No LLMs	*LLMs*	*No LLMs*	*LLMs*
Year	2022	97.5	2.5	100.0	0.0
	2023	72.4	27.6	81.8	18.2
	2024	33.3	66.7	49.3	50.7
Paradigms ('23 + '24)	Attributions	37.4	**19.4**	41.9	**24.3**
	LIME/SHAP	6.3	3.3	17.5	**4.3**
	Probing	11.4	10.6	6.9	3.5
	Clustering	3.4	0.5	16.9	**2.6**
	Mechanistic	10.6	**15.2**	2.5	2.6
	Diagnostic	3.7	4.8	1.2	1.7
	Adversarial	4.6	5.6	1.2	0.0
	Counterfactuals	3.4	4.3	0.0	1.7
	Lang. Expl.	10.9	**30.8**	6.2	**48.7**
	Self-explain	7.1	4.8	4.4	6.1
	Classic	1.1	0.8	1.2	4.3

Table 3: Percentage of '23-'24 interpretability papers by field (**NLP** and fields **Outside** NLP) and by whether the paper employs an *LLM*. The top three rows present the distribution for each field and year (*LLMs* +*No LLMs*=100%). The 11 bottom rows present the distribution by paradigms, each column summing to 100%.

LLM papers have become prominent both within the NLP field (66.7% of the papers in 2024) and outside of it (from 18.2% in 2023 to 50.7% in 2024).

The widespread adoption of LLMs has shifted interpretability paradigms. Although paradigm trends in NLP were stable, the introduction of LLMs tripled the portion of Natural Language Explanation papers (30.8%), likely due to the strong generation capabilities of LLMs. Outside NLP, this paradigm accounts for nearly half of the papers that employ LLMs (48.7% compared to 6.2% in non-LLM papers). This is another indication that non-developers favor methods that do not require advanced technical skills, as generating textual explanations can be done through simple prompting.

We anticipate more trend shifts in the LLM era, particularly toward methods that leverage strong generation capabilities, such as generating Counterfactuals and dedicated Diagnostic Sets, which is already evident in a 30% increase in these paradigms.

6 Conclusions and Recommendations

In this half-position-half-survey paper, we reviewed hundreds of works on NLP model interpretability and analysis from the past decade. Unlike other surveys, we examined interpretability methods, paradigms, and properties from the stakeholders' perspective. Additionally, we conducted a first-of-its-kind large-scale trend analysis by exploring the usage of interpretability methods within the NLP community and in research fields outside of it.

Our analysis reveals substantial diversity between research fields, particularly between NLP developers and non-developer stakeholders. To bridge these gaps and promote the adoption of NLP interpretability methods in other fields, we recommend the following steps for NLP researchers:

Clearly define the stakeholders and applications of your work. Researchers should explicitly state in the introduction who the stakeholders of their method are, the needs it addresses, its core properties, and its potential applications within and outside the NLP community. Articulating these aspects helps position the research within a broader context and ensures relevant audiences can effectively engage with the method. Additionally, demonstrating applications of interpretability methods in other fields can enhance their visibility and adoption. Publishing NLP research in interdisciplinary venues (Ophir et al., 2020; Badian et al., 2023) fosters cross-domain collaboration and broadens the impact beyond NLP.

Develop user-friendly code and write detailed guides for non-technical users. Researchers outside the NLP community sometimes utilize specific methods due to specific needs (e.g., probing in neuroscience is used for aligning representations with brain activity). Yet, many utilize methods for the wrong reason: extensive familiarity with popular methods in non-NLP domains and with well-documented code in common data science libraries (e.g., SHAP, LIME, and Scikit-learn).

To encourage the adoption of NLP interpretability methods beyond our community, researchers should prioritize developing user-friendly code accompanied by detailed guides for non-technical users. Additionally, the code should generate attractive and easy-to-understand visualizations. Making the methods more accessible can help integrate them into other scientific and industrial domains.

Expand the reach of prevalent NLP interpretability paradigms. Two paradigms have gained traction in NLP, particularly with the rise of LLMs: Natural Language Explanations and Mechanistic Interpretability. We found that natural language explanation methods are also extremely prevalent in non-NLP fields. We believe this rapid adoption is concerning, as their reliability remains a topic of ongoing debate in research. Our community should investigate the faithfulness of these methods (Lanham et al., 2023; Parcalabescu and Frank, 2023; Bao et al., 2024; Wu et al., 2024) and determine whether they can replace traditional, extensively researched methods.

Conversely, while Mechanistic interpretability research is trending within the NLP community, explanations of internal model components are rarely used in other fields. Our community should explore whether and how mechanistic interpretability can be adapted more broadly (Sharkey et al., 2025).

We need more concept-level, self-explaining, and causal-based methods. In Appendix §A.1.2, we highlight the potential of high-level concept explanations, particularly for non-expert stakeholders such as end-users, given the challenges of explaining lengthy raw textual inputs. Even though they can improve the accessibility of model insights (Poursabzi-Sangdeh et al., 2021), concept-level methods remain largely underutilized, accounting for only 2% of the papers, as shown in Figure 6.

Stakeholders using NLP models for decision-making require faithful explanations (Feder et al., 2022). In Appendix §A.2.4, we highlight the important role of causality in fostering faithfulness. Yet, Table 2 indicates that causal-based methods are rarely used (5.2% in NLP and 1.9% outside).

Finally, building on the seminal calls of XAI researchers (Rudin, 2018; Arrieta et al., 2020), we believe in self-explaining methods as a promising path toward the "holy grail" of NLP: achieving intrinsic interpretability while minimizing performance degradation. Yet, as Table 3 indicates, only about 7% of papers focus on self-explaining models, leaving them largely underexplored.

The LLM era presents new research opportunities. Despite the expectation that non-developers would benefit from concept-level, self-explaining, and causal-based methods, their adoption remains limited. We believe this is mainly due to the lack of research and development within the NLP community. This gap restricts the broader applicability of NLP models, particularly in domains where transparency and interpretability are essential.

The increasing capabilities of LLMs provide an unprecedented opportunity to develop novel concept-level, self-explaining, and causal-based interpretability methods. Indeed, many of the works discussed in this paper demonstrate such potential (e.g., Gat et al. (2023) and Stacey et al. (2024)). By expanding research in these directions, the NLP community can contribute to developing models that are more reliable, explainable, and accessible to a broader range of stakeholders.

7 Limitations

Other Modalities. The focus of our paper, while broad, centers on NLP and does not address other input modalities beyond text, such as visual or audio. These modalities, especially when considering the recent advancement of large multimodal models, could be vital for certain stakeholders, and it can be believed that the conclusions from our analysis would not be generalized to interpretability methods of vision and audio systems.

LLM Annotations. Even though we manually verified the LLM annotations on a subset of 100 papers and observed high agreement rates with human annotations (over 95%), it is possible that the LLM introduced potential biases. The statistics might have differed slightly if all 2000+ papers had been manually annotated. However, the manual annotation process is extremely time-consuming and requires high-domain expertise. This process involved reading full abstracts and assessing the nine annotation properties (900 annotations). Therefore, while our findings benefit from high agreement rates between LLM and human annotations, they also emphasize the need for continuous human oversight and validation in studies that use automated tools for literature analysis (Calderon et al., 2025).

Acknowledgments

NC is funded by the Clore Foundation's PhD fellowship. We gratefully acknowledge the support provided by Google's Gemma Academic Program, which has significantly contributed to advancing our research. We would also like to thank the members of the DDS@Technion NLP group for their valuable feedback and advice.

References

Eldar David Abraham, Karel D'Oosterlinck, Amir Feder, Yair Ori Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. 2022. Cebab: Estimating the causal effects of real-world concepts on NLP model behavior. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7747–7763. Association for Computational Linguistics.

Firoj Alam, Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Abdul Rafae Khan, and Jia Xu. 2023. Conceptx: A framework for latent concept analysis. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 16395–16397. AAAI Press.

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preotiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: a natural language processing perspective. *PeerJ Comput. Sci.*, 2:e93.

Genevera I. Allen, Luqin Gan, and Lili Zheng. 2023. Interpretable machine learning for discovery: Statistical challenges \& opportunities. *CoRR*, abs/2308.01475.

Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks: a user study. In *IUI '20: 25th International Conference on Intelligent User Interfaces, Cagliari, Italy, March 17-20, 2020*, pages 275–285. ACM.

Afra Amini, Tiago Pimentel, Clara Meister, and Ryan Cotterell. 2023. Naturalistic causal probing for morpho-syntax. *Transactions of the Association for Computational Linguistics*, 11:384–403.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 39–48. IEEE Computer Society.

Omer Antverg and Yonatan Belinkov. 2022. On the pitfalls of analyzing individual neurons in language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Dana Arad, Hadas Orgad, and Yonatan Belinkov. 2023. Refact: Updating text-to-image models by editing the text encoder. *CoRR*, abs/2306.00738.

Alejandro Barredo Arrieta, Natalia Díaz Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and

Francisco Herrera. 2020. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, 58:82–115.

Giuseppe Attanasio, Flor Miriam Plaza del Arco, Debora Nozza, and Anne Lauscher. 2023. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3996–4014. Association for Computational Linguistics.

Yael Badian, Yaakov Ophir, Refael Tikochinski, Nitay Calderon, Anat Brunstein Klomek, Eyal Fruchter, and Roi Reichart. 2023. Social media images can predict suicide risk using interpretable large language-vision models. *The Journal of clinical psychiatry*, 85 1.

Ananth Balashankar, Xuezhi Wang, Yao Qin, Ben Packer, Nithum Thain, Ed H. Chi, Jilin Chen, and Alex Beutel. 2023. Improving classifier robustness through active generative counterfactual data augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 127–139. Association for Computational Linguistics.

Esma Balkir, Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser. 2022. Challenges in applying explainability methods to improve the fairness of NLP models. *CoRR*, abs/2206.03945.

Guangsheng Bao, Hongbo Zhang, Linyi Yang, Cunxiang Wang, and Yue Zhang. 2024. Llms with chain-of-thought are non-causal reasoners. *CoRR*, abs/2402.16048.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2963–2977. Association for Computational Linguistics.

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James R. Glass. 2019. Identifying and controlling important neurons in neural machine translation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Comput. Linguistics*, 48(1):207–219.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Yonatan Belinkov and James R. Glass. 2019. Analysis methods in neural language processing: A survey. *Trans. Assoc. Comput. Linguistics*, 7:49–72.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *CoRR*, abs/2303.08112.

Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for AI safety - A review. *CoRR*, abs/2404.14082.

Milan Bhan, Jean-Noel Vittaut, Nina Achache, Victor Legrand, Nicolas Chesneau, Annabelle Blangero, Juliette Murris, and Marie-Jeanne Lesot. 2024. Mitigating text toxicity with counterfactual generation.

Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. 2023. Science in the age of large language models. *Nature Reviews Physics*, 5:277–280.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna M. Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5454–5476. Association for Computational Linguistics.

Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2022. Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. In *IUI 2022: 27th International Conference on Intelligent User Interfaces, Helsinki, Finland, March 22 - 25, 2022*, pages 807–819. ACM.

Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *IUI '20: 25th International Conference on Intelligent User Interfaces, Cagliari, Italy, March 17-20, 2020*, pages 454–464. ACM.

Aljoscha Burchardt, Vivien Macketanz, Jon Dehdari, Georg Heigold, Jan-Thorsten Peter, and Philip Williams. 2017. A linguistic evaluation of rule-based, phrase-based, and neural MT engines. *Prague Bull. Math. Linguistics*, 108:159–170.

Franck Burlot and François Yvon. 2017. Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 43–55. Association for Computational Linguistics.

Nitay Calderon, Eyal Ben-David, Amir Feder, and Roi Reichart. 2022. Docogen: Domain counterfactual generation for low resource domain adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 7727–7746. Association for Computational Linguistics.

Nitay Calderon, Subhabrata Mukherjee, Roi Reichart, and Amir Kantor. 2023. A systematic study of knowledge distillation for natural language generation with pseudo-target training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14632–14659. Association for Computational Linguistics.

Nitay Calderon, Naveh Porat, Eyal Ben-David, Alexander Chapanin, Zorik Gekhman, Nadav Oved, Vitaly Shalumov, and Roi Reichart. 2024. Measuring the robustness of nlp models to domain shifts. *arXiv preprint arXiv:2306.00168*.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for llm-as-a-judge: How to statistically justify replacing human annotators with llms.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9560–9572.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. Make up your mind! adversarial generation of inconsistent natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4157–4165. Association for Computational Linguistics.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 782–791. Computer Vision Foundation / IEEE.

Saneem A. Chemmengath, Amar Prakash Azad, Ronny Luss, and Amit Dhurandhar. 2022. Let the CAT out of the bag: Contrastive attributed explanations for text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7190–7206. Association for Computational Linguistics.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2021. An empirical survey of data augmentation for limited data learning in nlp. *Transactions of the Association for Computational Linguistics*, 11:191–211.

Yu Ying Chiu, Liwei Jiang, Maria Antoniak, Chan Young Park, Shuyue Stella Li, Mehar Bhatia, Sahithya Ravi, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. Culturalteaming: Ai-assisted interactive red-teaming for challenging llms' (lack of) multicultural knowledge. *CoRR*, abs/2404.06664.

Avishek Choudhury and Hamid Shamszare. 2023. Investigating the impact of user trust on the adoption and use of chatgpt: Survey analysis. *Journal of Medical Internet Research*, 25.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future. *CoRR*, abs/2309.15402.

Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, et al. 2023. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single \$&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2126–2136. Association for Computational Linguistics.

Róbert Csordás, Sjoerd van Steenkiste, and Jürgen Schmidhuber. 2021. Are neural nets modular? inspecting functional modularity through differentiable weight masks. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *CoRR*, abs/2309.08600.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference*

of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020, pages 447–459. Association for Computational Linguistics.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. 2023. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16124–16170. Association for Computational Linguistics.

Anubrata Das, Chitrank Gupta, Venelin Kovatchev, Matthew Lease, and Junyi Jessy Li. 2022. Prototex: Explaining model decisions with prototype tensors. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2986–2997. Association for Computational Linguistics.

Maria De-Arteaga, Alexey Romanov, Hanna M. Wallach, Jennifer T. Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 120–128. ACM.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M. Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1968–1994. Association for Computational Linguistics.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: dataset and metrics for measuring biases in open-ended language generation. In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 862–872. ACM.

Tanay Dixit, Bhargavi Paranjape, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. CORE: A retrieve-then-edit framework for counterfactual data generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2964–2984. Association for Computational Linguistics.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Jad Doughman and Wael Khreich. 2022. Gender bias in text: Labeled datasets and lexicons. *CoRR*, abs/2201.08675.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. Analyzing individual neurons in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4865–4880. Association for Computational Linguistics.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 31–36. Association for Computational Linguistics.

Nature Editorials. 2023. Tools such as chatgpt threaten transparent science; here are our ground rules for their use. *Nature*, 613:612.

Mahmoud El-Haj, Paul Rayson, Martin Walker, Steven Young, and Vasiliki Simaki. 2019. In search of meaning: Lessons, resources and next steps for computational analysis of financial discourse. *Journal of Business Finance & Accounting*, 46(3-4):265–306.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Amir Feder, Abhilasha Ravichander, Marius Mosbach, Yonatan Belinkov, Hinrich Schütze, and Yoav Goldberg. 2022. Measuring causal effects of data statistics on language model's 'factual' predictions. *CoRR*, abs/2207.14251.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Trans. Assoc. Comput. Linguistics*, 9:160–175.

Yanai Elazar, Jiayao Zhang, David Wadden, Bo Zhang, and Noah A. Smith. 2024. Estimating the causal effect of early arxiving on paper acceptance. In *Causal Learning and Reasoning, 1-3 April 2024, Los Angeles, California, USA*, volume 236 of *Proceedings of Machine Learning Research*, pages 913–933. PMLR.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. Gpts are gpts: An early look at the labor market impact potential of large language models. *CoRR*, abs/2303.10130.

Joseph Enguehard. 2023. Sequential integrated gradients: a simple but effective method for explaining language models. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 7555–7565. Association for Computational Linguistics.

Biaoyan Fang, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2023. It's not only what you say, it's also who it's said to: Counterfactual analysis of interactive behavior in the courtroom. In *Proceedings of the 13th International Joint Conference on Natural*

Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 - Volume 2: Short Papers, Nusa Dua, Bali, November 1-4, 2023, pages 197–207. Association for Computational Linguistics.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Trans. Assoc. Comput. Linguistics*, 10:1138–1158.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. Causalm: Causal model explanation through counterfactual language models. *Comput. Linguistics*, 47(2):333–386.

Amir Feder, Yoav Wald, Claudia Shi, Suchi Saria, and David M. Blei. 2023. Causal-structure driven augmentations for text OOD generalization. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

George Filandrianos, Edmund Dervakos, Orfeas Menis-Mastromichalakis, Chrysoula Zerva, and Giorgos Stamou. 2023. Counterfactuals of counterfactuals: a back-translation-inspired approach to analyse counterfactual editors. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9507–9525. Association for Computational Linguistics.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart M. Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1828–1843. Association for Computational Linguistics.

Krishna Gade, Sahin Cem Geyik, Krishnaram Kenthapadi, Varun Mithal, and Ankur Taly. 2020. Explainable AI in industry: practical challenges and lessons learned: implications tutorial. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, page 699. ACM.

Albert Garde, Esben Kran, and Fazl Barez. 2023. Deepdecipher: Accessing and investigating neuron activation in large language models. *CoRR*, abs/2310.01870.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, and et al. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1307–1323. Association for Computational Linguistics.

Itai Gat, Nitay Calderon, Roi Reichart, and Tamir Hazan. 2022. A functional information perspective on model interpretation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 7266–7278. PMLR.

Yair Ori Gat, Nitay Calderon, Amir Feder, Alexander Chapanin, Amit Sharma, and Roi Reichart. 2023. Faithful explanations of black-box nlp models using llm-generated counterfactuals. In *The Twelfth International Conference on Learning Representations*.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 9574–9586.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah D. Goodman, and Christopher Potts. 2022. Inducing causal structure for interpretable neural networks. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 7324–7338. PMLR.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? *CoRR*, abs/2405.05904.

Gloria Gennaro and Elliott Ash. 2022. Emotion and reason in political language. *The Economic Journal*, 132(643):1037–1059.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 30–45. Association for Computational Linguistics.

Reza Ghaeini, Xiaoli Z. Fern, Hamed Shahbazi, and Prasad Tadepalli. 2019. Saliency learning: Teaching the model where to pay attention. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4016–4025. Association for Computational Linguistics.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A

unifying framework for inspecting hidden representations of language models. *CoRR*, abs/2401.06102.

Amirata Ghorbani and James Y. Zou. 2020. Neuron shapley: Discovering the responsible neurons. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem H. Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 240–248. Association for Computational Linguistics.

Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal, and Christopher Ré. 2021. Robustness gym: Unifying the NLP evaluation landscape. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 42–55. Association for Computational Linguistics.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. 2022. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380.

Bryce Goodman and Seth R. Flaxman. 2017. European union regulations on algorithmic decision-making and a "right to explanation". *AI Mag.*, 38(3):50–57.

Shreya Goyal, Sumanth Doddapaneni, Mitesh M. Khapra, and Balaraman Ravindran. 2023. A survey of adversarial defenses and robustness in NLP. *ACM Comput. Surv.*, 55(14s):332:1–332:39.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1195–1205. Association for Computational Linguistics.

Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. 2021. Gradient-based adversarial attacks against text transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5747–5757. Association for Computational Linguistics.

Nitish Gupta, Kevin Lin, Dan Roth, Sameer Singh, and Matt Gardner. 2020. Neural module networks for reasoning over text. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Vikram Gupta, Haoyue Shi, Kevin Gimpel, and Mrinmaya Sachan. 2022. Deep clustering of text representations for supervision-free probing of syntax. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10720–10728. AAAI Press.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. 2024. Universal neurons in GPT2 language models. *CoRR*, abs/2401.12181.

Pantea Haghighatkhah, Antske Fokkens, Pia Sommerauer, Bettina Speckmann, and Kevin Verbeek. 2022. Better hit the nail on the head than beat around the bush: Removing protected attributes with a single projection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8395–8416. Association for Computational Linguistics.

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2022. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Majd Hawasly, Fahim Dalvi, and Nadir Durrani. 2024. Scaling up discovery of latent concepts in deep NLP models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 793–806. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2733–2743. Association for Computational Linguistics.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (gen-

uine) similarity estimation. *Comput. Linguistics*, 41(4):665–695.

Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven Mark Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 579. ACM.

Pengfei Hong, Rishabh Bhardwaj, Navonil Majumder, Somak Aditya, and Soujanya Poria. 2023. A robust information-masking approach for domain counterfactual generation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3756–3769. Association for Computational Linguistics.

Phillip Howard, Gadi Singer, Vasudev Lal, Yejin Choi, and Swabha Swayamdipta. 2022. Neurocounterfactuals: Beyond minimal-edit counterfactuals for richer data augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5056–5072. Association for Computational Linguistics.

Gloria Hristova and Nikolay Netov. 2022. Media coverage and public perception of distance learning during the COVID-19 pandemic: A topic modeling approach based on bertopic. In *IEEE International Conference on Big Data, Big Data 2022, Osaka, Japan, December 17-20, 2022*, pages 2259–2264. IEEE.

Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. 2017. Learning to reason: End-to-end module networks for visual question answering. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 804–813. IEEE Computer Society.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4198–4205. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3543–3556. Association for Computational Linguistics.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4459–4473. Association for Computational Linguistics.

Jae-young Jo and Sung-Hyon Myaeng. 2020. Roles and utilization of attention heads in transformer-based neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3404–3417. Association for Computational Linguistics.

Brihi Joshi, Aaron Chan, Ziyi Liu, Shaoliang Nie, Maziar Sanjabi, Hamed Firooz, and Xiang Ren. 2022. Er-test: Evaluating explanation regularization methods for language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3315–3336. Association for Computational Linguistics.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? on the ability of lstms to understand negative polarity items. In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 222–231. Association for Computational Linguistics.

Samia Kabir, Lixiang Li, and Tianyi Zhang. 2024. STILE: exploring and debugging social biases in pretrained text representations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 293:1–293:20. ACM.

Alexander John Karran, Théophile Demazure, Antoine Hudon, Sylvain Sénécal, and Pierre-Majorique Léger. 2022. Designing for confidence: The impact of visualizing artificial intelligence decisions. *Frontiers in Neuroscience*, 16.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, and et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*.

Shahar Katz and Yonatan Belinkov. 2023. VISIT: visualizing and interpreting the semantic information flow of transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 14094–14113. Association for Computational Linguistics.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna M. Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–14. ACM.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna M. Wallach, and Jennifer Wortman Vaughan. 2021. Interpreting interpretability: Understanding data scientists' use of interpretability tools

for machine learning. In *3rd Workshop on Data Science with Human in the Loop, DaSH@KDD, Virtual Conference, August 15, 2021*.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. 2020. Learning the difference that makes A difference with counterfactually-augmented data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Divyansh Kaushik, Amrith Setlur, Eduard H. Hovy, and Zachary Chase Lipton. 2021. Explaining the efficacy of counterfactually augmented data. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Daphna Keidar, Andreas Opedal, Zhijing Jin, and Mrinmaya Sachan. 2022. Slangvolution: A causal analysis of semantic change and frequency dynamics in slang. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1422–1442. Association for Computational Linguistics.

Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12685–12708. Association for Computational Linguistics.

Margaret King and Kirsten Falkedal. 1990. Using test suites in evaluation of machine translation systems. In *13th International Conference on Computational Linguistics, COLING 1990, University of Helsinki, Finland, August 20-25, 1990*, pages 211–216.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. *CoRR*, abs/2007.04612.

Enja Kokalj, Blaz Skrlj, Nada Lavrac, Senja Pollak, and Marko Robnik-Sikonja. 2021. BERT meets shapley: Extending SHAP explanations to transformer-based classifiers. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, EACL 2021, Online, April 19, 2021*, pages 16–21. Association for Computational Linguistics.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4364–4373. Association for Computational Linguistics.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. 2024. Atp*: An efficient and scalable method for localizing LLM behaviour to components. *CoRR*, abs/2403.00745.

Maya Krishnan. 2019. Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, 33:487 – 502.

Abhinav Kumar, Chenhao Tan, and Amit Sharma. 2022. Probing classifiers are unreliable for concept removal and detection. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Sawan Kumar and Partha P. Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8730–8742. Association for Computational Linguistics.

Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 29–38. ACM.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, and et al. 2023. Measuring faithfulness in chain-of-thought reasoning. *CoRR*, abs/2307.13702.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, and et al. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations.

David MJ Lazer, Alex Pentland, Duncan J Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra Gonzalez-Bailon, Gary King, Helen Margetts, et al. 2020. Computational social science: Obstacles and opportunities. *Science*, 369(6507):1060–1062.

Sabine Lehmann, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, Eva Dauphin, Herve Compagnion, Judith Baur, Lorna Balkan, and Doug Arnold. 1996. TSNLP - test suites for natural language processing. In *16th International Conference on Computational Linguistics, Proceedings of the Conference, COLING 1996, Center for Sprogteknologi, Copenhagen, Denmark, August 5-9, 1996*, pages 711–716.

Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 107–117. The Association for Computational Linguistics.

Michael A. Lepori and R. Thomas McCoy. 2020. Picking bert's brain: Probing for linguistic dependencies in contextualized embeddings using representational similarity analysis. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3637–3651. International Committee on Computational Linguistics.

Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-based human debugging of NLP models: A survey. *Trans. Assoc. Comput. Linguistics*, 9:1508–1528.

Ira Leviant and Roi Reichart. 2015. Judgment language matters: Multilingual vector space models for judgment language aware lexical semantics. *CoRR*, abs/1508.00106.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2019. Textbugger: Generating adversarial text against real-world applications. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220.

Shaobo Li, Xiaoguang Li, Lifeng Shang, Zhenhua Dong, Chengjie Sun, Bingquan Liu, Zhenzhou Ji, Xin Jiang, and Qun Liu. 2022. How pre-trained language models capture factual knowledge? A causal-inspired analysis. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1720–1732. Association for Computational Linguistics.

Yongqi Li, Mayi Xu, Xin Miao, Shen Zhou, and Tieyun Qian. 2024. Prompting large language models for counterfactual generation: An empirical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 13201–13221. ELRA and ICCL.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. 2024. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on AI conference peer reviews. *CoRR*, abs/2403.07183.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 158–167. Association for Computational Linguistics.

Zachary C. Lipton. 2018. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43.

Shir Lissak, Nitay Calderon, Geva Shenkman, Yaakov Ophir, Eyal Fruchter, Anat Brunstein Klomek, and Roi Reichart. 2024a. The colorful future of llms: Evaluating and improving llms as emotional supporters for queer youth. *CoRR*, abs/2402.11886.

Shir Lissak, Yaakov Ophir, Refael Tikochinski, Anat Brunstein Klomek, Itay Sisso, Eyal Fruchter, and Roi Reichart. 2024b. Bored to death: Artificial intelligence research reveals the role of boredom in suicide behavior. *Frontiers in Psychiatry*, 15.

Haoyang Liu, Maheep Chaudhary, and Haohan Wang. 2023. Towards trustworthy and aligned machine learning: A data-centric survey with causality perspectives. *CoRR*, abs/2307.16851.

Josh Magnus Ludan, Qing Lyu, Yue Yang, Liam Dugan, Mark Yatskar, and Chris Callison-Burch. 2023. Interpretable-by-design text classification with iteratively generated concept bottleneck. *CoRR*, abs/2310.19660.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774.

Siwen Luo, Hamish Ivison, Soyeon Caren Han, and Josiah Poon. 2024. Local interpretations for explainable natural language processing: A survey. *ACM Comput. Surv.*, 56(9):232:1–232:36.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2022. Towards faithful model explanation in NLP: A survey. *CoRR*, abs/2209.11326.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 - 4, 2023*, pages 305–329. Association for Computational Linguistics.

Mark MacCarthy. 2019. An examination of the algorithmic accountability act of 2019. *Available at SSRN 3615731*.

Aman Madaan, Katherine Hermann, and Amir Yazdanbakhsh. 2023. What makes chain-of-thought prompting effective? A counterfactual study. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1448–1535. Association for Computational Linguistics.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2023. Post-hoc interpretability for neural NLP: A survey. *ACM Comput. Surv.*, 55(8):155:1–155:42.

Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew E. Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 410–424. Association for Computational Linguistics.

Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, and et al. 2023. Inverse scaling: When bigger isn't better. *CoRR*, abs/2306.09479.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Rakesh Menon, Kerem Zaman, and Shashank Srivastava. 2023. MaNtLE: Model-agnostic natural language explainer. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13493–13511, Singapore. Association for Computational Linguistics.

Julian Michael, Jan A. Botha, and Ian Tenney. 2020. Asking without telling: Exploring latent ontologies in contextual representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6792–6812. Association for Computational Linguistics.

David Miller, Mishel Johns, Brian Mok, Nikhil Gowda, David Sirkin, Key Lee, and Wendy Ju. 2016. Behavioral measurement of trust in automation: the trust fall. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 60, pages 1849–1853. SAGE Publications Sage CA: Los Angeles, CA.

Tim Miller. 2017. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. Layer-wise relevance propagation: An overview. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*, pages 193–209. Springer.

John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 119–126. Association for Computational Linguistics.

Marius Mosbach, Vagrant Gautam, Tomás Vergara Browne, Dietrich Klakow, and Mor Geva. 2024. From insights to actions: The impact of interpretability and analysis research on NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 3078–3105. Association for Computational Linguistics.

Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. 2022. Shap-based explanation methods: A review for NLP interpretability. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 4593–4603. International Committee on Computational Linguistics.

Basel Mousi, Nadir Durrani, and Fahim Dalvi. 2023. Can llms facilitate interpretation of pre-trained language models? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3248–3268. Association for Computational Linguistics.

Romy Müller. 2024. How explainable AI affects human performance: A systematic review of the behavioural consequences of saliency maps. *CoRR*, abs/2404.16042.

W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *CoRR*, abs/2004.14546.

Benjamin Newman, Kai-Siang Ang, Julia Gong, and John Hewitt. 2021. Refining targeted syntactic evaluation of language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3710–3723. Association for Computational Linguistics.

Van Bach Nguyen, Paul Youssef, Jörg Schlötterer, and Christin Seifert. 2024. Llms for generating and evaluating counterfactuals: A comprehensive study. *CoRR*, abs/2405.00722.

Yaakov Ophir, Refael Tikochinski, Christa S. C. Asterhan, Itay Sisso, and Roi Reichart. 2020. Deep neural networks detect suicide risk from textual facebook posts. *Scientific Reports*, 10.

Yaakov Ophir, Refael Tikochinski, Anat Brunstein Klomek, and Roi Reichart. 2022. The hitchhiker's guide to computational linguistics in suicide prevention. *Clinical Psychological Science*, 10(2):212–235.

Juri Opitz and Anette Frank. 2022. SBERT studies meaning representations: Decomposing sentence embeddings into explainable semantic features. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 625–638, Online only. Association for Computational Linguistics.

Hadas Orgad, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2022. How gender debiasing affects internal model representations, and why it matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2602–2628. Association for Computational Linguistics.

Koyena Pal, Jiuding Sun, Andrew Yuan, Byron C. Wallace, and David Bau. 2023. Future lens: Anticipating subsequent tokens from a single hidden state. In *Proceedings of the 27th Conference on Computational Natural Language Learning, CoNLL 2023, Singapore, December 6-7, 2023*, pages 548–560. Association for Computational Linguistics.

Nicolas Papernot and Patrick D. McDaniel. 2018. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *CoRR*, abs/1803.04765.

Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253.

Letitia Parcalabescu and Anette Frank. 2023. On measuring faithfulness of natural language explanations. *CoRR*, abs/2311.07466.

Anselm Paulus, Arman Zharmagambetov, Chuan Guo, Brandon Amos, and Yuandong Tian. 2024. Advprompter: Fast adaptive adversarial prompting for llms. *CoRR*, abs/2404.16873.

Pedro Pavón. 2023. Increasing our ads transparency.

Ethan Perez, Saffron Huang, H. Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3419–3448. Association for Computational Linguistics.

Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2021. Manipulating and measuring model interpretability. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, pages 237:1–237:52. ACM.

Dheeraj Rajagopal, Vidhisha Balachandran, Eduard H. Hovy, and Yulia Tsvetkov. 2021. SELFEXPLAIN: A self-explaining architecture for neural text classifiers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 836–850. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4932–4942. Association for Computational Linguistics.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. NORMAD: A benchmark for measuring the cultural adaptability of large language models. *CoRR*, abs/2404.12464.

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. 2023. 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 2023, Raleigh, NC, USA, February 8-10, 2023*, pages 464–483. IEEE.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7237–7256. Association for Computational Linguistics.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard H. Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 3363–3377. Association for Computational Linguistics.

Abhilasha Ravichander, Eduard H. Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. On the systematicity of probing contextualized word representations: The case of hypernymy in BERT. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics, *SEM@COLING 2020, Barcelona, Spain (Online), December 12-13, 2020*, pages 88–102. Association for Computational Linguistics.

Neguine Rezaii, Phillip Wolff, and Bruce H Price. 2022. Natural language processing in psychiatry:

the promises and perils of a transformative approach. *The British Journal of Psychiatry*, 220(5):251–253.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1527–1535. AAAI Press.

Marco Túlio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2021. Beyond accuracy: Behavioral testing of NLP models with checklist (extended abstract). In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4824–4828. ijcai.org.

Margaret E. Roberts, Brandon M Stewart, and Richard A. Nielsen. 2020. Adjusting for confounding with text matching. *American Journal of Political Science*, 64:887–903.

Elias Abad Rocamora, Yongtao Wu, Fanghui Liu, Grigorios G. Chrysos, and Volkan Cevher. 2024. Revisiting character-level adversarial attacks.

Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. 2020. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216.

Alexis Ross, Ana Marasovic, and Matthew E. Peters. 2021. Explaining NLP models via minimal contrastive editing (mice). In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3840–3852. Association for Computational Linguistics.

Alexis Ross, Tongshuang Wu, Hao Peng, Matthew E. Peters, and Matt Gardner. 2022. Tailor: Generating and perturbing text with semantic controls. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3194–3213. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulic, and Anders Søgaard. 2022. Square one bias in NLP: towards a multi-dimensional exploration of the research manifold. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2340–2354. Association for Computational Linguistics.

Cynthia Rudin. 2018. Please stop explaining black box models for high stakes decisions. *CoRR*, abs/1811.10154.

Rachneet Sachdeva, Martin Tutek, and Iryna Gurevych. 2024. Catfood: Counterfactual augmented training for improving out-of-domain performance and calibration. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 1876–1898. Association for Computational Linguistics.

Waddah Saeed and Christian W. Omlin. 2023. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowl. Based Syst.*, 263:110273.

Hassan Sajjad, Nadir Durrani, and Fahim Dalvi. 2022a. Neuron-level interpretation of deep NLP models: A survey. *Transactions of the Association for Computational Linguistics*, 10:1285–1303.

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Firoj Alam, Abdul Rafae Khan, and Jia Xu. 2022b. Analyzing encoded concepts in transformer language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3082–3101. Association for Computational Linguistics.

Mansi Sakarvadia, Arham Khan, Aswathy Ajith, Daniel Grzenda, Nathaniel Hudson, André Bauer, Kyle Chard, and Ian T. Foster. 2023. Attention lens: A tool for mechanistically interpreting the attention head information retrieval mechanism. *CoRR*, abs/2310.16270.

Jim Samuel. 2023. Response to the march 2023 'pause giant ai experiments: An open letter' by yoshua bengio, signed by stuart russell, elon musk, steve wozniak, yuval noah harari and others.... *SSRN Electronic Journal*.

Mikayel Samvelyan, Sharath Chandra Raparthy, Andrei Lupu, Eric Hambro, Aram H. Markosyan, Manish Bhatt, Yuning Mao, Minqi Jiang, Jack Parker-Holder, Jakob N. Foerster, Tim Rocktäschel, and Roberta Raileanu. 2024. Rainbow teaming: Open-ended generation of diverse adversarial prompts. *CoRR*, abs/2402.16822.

Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter W. Battaglia, and Tim Lillicrap. 2017. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4967–4976.

T. Y. S. S. Santosh, Nina Baumgartner, Matthias Stürmer, Matthias Grabmair, and Joel Niklaus.

2024. Towards explainability and fairness in swiss judgement prediction: Benchmarking on a multilingual dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 16500–16513. ELRA and ICCL.

Mario Sanz-Guerrero and Javier Arroyo. 2024. Credit risk meets large language models: Building a risk indicator from loan descriptions in P2P lending. *CoRR*, abs/2401.16458.

Sheikh Muhammad Sarwar, Dimitrina Zlatkova, Momchil Hardalov, Yoan Dinkov, Isabelle Augenstein, and Preslav Nakov. 2022. A neighborhood framework for resource-lean content flagging. *Trans. Assoc. Comput. Linguistics*, 10:484–502.

Sophia Schulze-Weddige and Thorsten Zylowski. 2021. User study on the effects explainable AI visualizations on non-experts. In *ArtsIT, Interactivity and Game Creation - Creative Heritage. New Perspectives from Media Arts and Artificial Intelligence. 10th EAI International Conference, ArtsIT 2021, Virtual Event, December 2-3, 2021, Proceedings*, volume 422 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 457–467. Springer.

Indira Sen, Dennis Assenmacher, Mattia Samory, Isabelle Augenstein, Wil M. P. van der Aalst, and Claudia Wagner. 2023. People make better edits: Measuring the efficacy of llm-generated counterfactually augmented data for harmful language detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 10480–10504. Association for Computational Linguistics.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 376–382. Association for Computational Linguistics.

Eilam Shapira, Reut Apel, Moshe Tennenholtz, and Roi Reichart. 2023. Human choice prediction in language-based non-cooperative games: Simulation-based off-policy evaluation. *CoRR*, abs/2305.10361.

Eilam Shapira, Omer Madmon, Roi Reichart, and Moshe Tennenholtz. 2024. Can large language models replace economic choice prediction labs? *CoRR*, abs/2401.17435.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adrià Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Thomas McGrath. 2025. Open problems in mechanistic interpretability.

Sandipan Sikdar, Parantapa Bhattacharya, and Kieran Heese. 2021. Integrated directional gradients: Feature interaction attribution for neural NLP models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 865–878. Association for Computational Linguistics.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking interpretability in the era of large language models. *CoRR*, abs/2402.01761.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9180–9211. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Oana-Maria Camburu, and Marek Rei. 2024. Atomic inference for NLI with generated facts as atoms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10188–10204, Miami, Florida, USA. Association for Computational Linguistics.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, and Marek Rei. 2022. Logical reasoning with span-level predictions for interpretable and robust NLI models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3809–3823. Association for Computational Linguistics.

Yiheng Su, Junyi Jessy Li, and Matthew Lease. 2023. Interpretable by design: Wrapper boxes combine neural performance with faithful explanations. *CoRR*, abs/2311.08644.

Michael Sullivan. 2024. It is not true that transformers are inductive learners: Probing NLI models with external negation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 1924–1945. Association for Computational Linguistics.

Jiao Sun, Swabha Swayamdipta, Jonathan May, and Xuezhe Ma. 2022. Investigating the benefits of free-form rationales. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5867–5882. Association for Computational Linguistics.

Aaquib Syed, Can Rager, and Arthur Conmy. 2023. Attribution patching outperforms automated circuit discovery. *CoRR*, abs/2310.10348.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 175–185. The Association for Computer Linguistics.

Zhen Tan, Lu Cheng, Song Wang, Bo Yuan, Jundong Li, and Huan Liu. 2024. Interpreting pretrained language models via concept bottlenecks. In *Advances in Knowledge Discovery and Data Mining - 28th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2024, Taipei, Taiwan, May 7-10, 2024, Proceedings, Part III*, volume 14647 of *Lecture Notes in Computer Science*, pages 56–74. Springer.

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic biases in LLM simulations of debates. *CoRR*, abs/2402.04049.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Laure Thompson and David Mimno. 2020. Topic modeling with contextualized word representation clusters. *CoRR*, abs/2010.12626.

Refael Tikochinski, Ariel Goldstein, Yoav Meiri, Uri Hasson, and Roi Reichart. 2024. Incremental accumulation of linguistic context in artificial and biological neural networks. *bioRxiv*.

Refael Tikochinski, Ariel Goldstein, Yaara Yeshurun, Uri Hasson, and Roi Reichart. 2023. Perspective changes in human listeners are aligned with the contextual transformation of the word embedding space. *Cerebral Cortex*, 33(12):7830–7842.

Michael Toker, Hadas Orgad, Mor Ventura, Dana Arad, and Yonatan Belinkov. 2024. Diffusion lens: Interpreting text encoders in text-to-image pipelines. *CoRR*, abs/2403.05846.

Marcos V. Treviso, Alexis Ross, Nuno Miguel Guerreiro, and André F. T. Martins. 2023. CREST: A joint framework for rationalization and counterfactual text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15109–15126. Association for Computational Linguistics.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle K. Barral, Katherine Chou, Gregory S. Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. 2024. Towards conversational diagnostic AI. *CoRR*, abs/2401.05654.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Mor Ventura, Eyal Ben-David, Anna Korhonen, and Roi Reichart. 2023. Navigating cultural chasms: Exploring and unlocking the cultural POV of text-to-image models. *CoRR*, abs/2310.01929.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 3: System Demonstrations*, pages 37–42. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Analyzing the source and target contributions to predictions in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual*

Event, August 1-6, 2021, pages 1126–1140. Association for Computational Linguistics.

Jörg von Garrel and Jana Mayer. 2023. Artificial intelligence in studies—use of chatgpt and ai-based tools among students in germany. *Humanities and Social Sciences Communications*, 10:1–9.

Ivan Vulic, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020. Multi-simlex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Comput. Linguistics*, 46(4):847–897.

Ivan Vulic, Goran Glavas, Fangyu Liu, Nigel Collier, Edoardo Maria Ponti, and Anna Korhonen. 2023. Probing cross-lingual lexical knowledge from multilingual sentence encoders. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2081–2097. Association for Computational Linguistics.

Eric Wallace, Shi Feng, and Jordan L. Boyd-Graber. 2018. Interpreting neural networks with nearest neighbors. In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 136–144. Association for Computational Linguistics.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. Allennlp interpret: A framework for explaining predictions of NLP models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019 - System Demonstrations*, pages 7–12. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou, and Muhao Chen. 2023a. A causal view of entity bias in (large) language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 15173–15184. Association for Computational Linguistics.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023b. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zihan Wang, Jingbo Shang, and Ruiqi Zhong. 2023c. Goal-driven explainable clustering via language descriptions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 10626–10649. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the Student Research Workshop, SRW@HLT-NAACL 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 88–93. The Association for Computational Linguistics.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Trans. Assoc. Comput. Linguistics*, 6:605–617.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 996–1005. Asian Federation of Natural Language Processing.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 11–20. Association for Computational Linguistics.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler-gap dependencies? In *Proceedings of the*

Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018, pages 211–221. Association for Computational Linguistics.

Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. 2018. Challenges of using text classifiers for causal inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4586–4598. Association for Computational Linguistics.

Tongshuang Wu, Marco Túlio Ribeiro, Jeffrey Heer, and Daniel S. Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6707–6723. Association for Computational Linguistics.

Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du, and Ninghao Liu. 2024. Usable XAI: 10 strategies towards exploiting explainability in the LLM era. *CoRR*, abs/2403.08946.

Zhengxuan Wu, Karel D'Oosterlinck, Atticus Geiger, Amir Zur, and Christopher Potts. 2023a. Causal proxy models for concept-based model explanations. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 37313–37334. PMLR.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah D. Goodman. 2023b. Interpretability at scale: Identifying causal mechanisms in alpaca. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4166–4176. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *CoRR*, abs/2306.13063.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Ben Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*, 18(6):160:1–160:32.

Kai Yang, Hui Yuan, and Raymond Y. K. Lau. 2022. Psycredit: An interpretable deep learning-based credit assessment approach facilitated by psychometric natural language processing. *Expert Syst. Appl.*, 198:116847.

Puyudi Yang, Jianbo Chen, Cho-Jui Hsieh, Jane-Ling Wang, and Michael I. Jordan. 2020. Greedy attack and gumbel attack: Generating adversarial examples for discrete data. *J. Mach. Learn. Res.*, 21:43:1–43:36.

Huihan Yao, Ying Chen, Qinyuan Ye, Xisen Jin, and Xiang Ren. 2021. Refining language models with compositional explanations. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8954–8967.

Chih-Kuan Yeh, Been Kim, Sercan Ömer Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. 2020. On completeness-aware concept-based explanations in deep neural networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Kun-Hsing Yu, Elizabeth Healey, Tze-Yun Leong, Isaac S. Kohane, and Arjun Kumar Manrai. 2024. Medical artificial intelligence and human values. *The New England journal of medicine*, 390 20:1895–1904.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9924–9959. Association for Computational Linguistics.

Omar Zaidan, Jason Eisner, and Christine D. Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, April 22-27, 2007, Rochester, New York, USA*, pages 260–267. The Association for Computational Linguistics.

Congzhi Zhang, Linhai Zhang, Deyu Zhou, and Guoqiang Xu. 2024a. Causal prompting: Debiasing large language model prompting based on front-door adjustment. *CoRR*, abs/2403.02738.

Huajie Zhang, Yuxin Ying, Fuzhen Zhuang, Haiqin Weng, Sun Ying, Zhao Zhang, Yiqi Tong, and Yan Liu. 2024b. Multi-round counterfactual generation: Interpreting and improving models of text classification. In *Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024*, pages 774–777. ACM.

Raymond Zhang, Neha Nayak Kennard, Daniel Scott Smith, Daniel A. McFarland, Andrew McCallum, and Katherine Keith. 2023. Causal matching with text embeddings: A case study in estimating the causal effects of peer review policies. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1284–1297. Association for Computational Linguistics.

Weiguo Zhang, Chao Wang, Yue Zhang, and Junbo Wang. 2020. Credit risk evaluation model with textual features from loan descriptions for P2P lending. *Electron. Commer. Res. Appl.*, 42:100989.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad-Reza Namazi-Rad. 2022. Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3886–3893. Association for Computational Linguistics.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2):20:1–20:38.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 15–20. Association for Computational Linguistics.

Mengjie Zhao, Tao Lin, Fei Mi, Martin Jaggi, and Hinrich Schütze. 2020. Masking as an efficient alternative to finetuning for pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2226–2241. Association for Computational Linguistics.

Ruochen Zhao, Shafiq R. Joty, Yongjie Wang, and Tan Wang. 2023a. Explaining language models' predictions with high-impact concepts. *CoRR*, abs/2305.02160.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. 2023b. On evaluating adversarial robustness of large vision-language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Carolina Zheng, Claudia Shi, Keyon Vafa, Amir Feder, and David M. Blei. 2023. An invariant learning characterization of controlled text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3186–3206. Association for Computational Linguistics.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. *CoRR*, abs/2401.06730.

Yuxiang Zhou and Yulan He. 2023. Causal inference from text: Unveiling interactions between variables. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10559–10571. Association for Computational Linguistics.

Sicheng Zhu, Ruiyi Zhang, Bang An, Gang Wu, Joe Barrow, Zichao Wang, Furong Huang, Ani Nenkova, and Tong Sun. 2023. Autodan: Automatic and interpretable adversarial attacks on large language models. *CoRR*, abs/2310.15140.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Comput. Linguistics*, 50(1):237–291.

Alexandra Zytek, Dongyu Liu, Rhema Vaithianathan, and Kalyan Veeramachaneni. 2022. Sibyl: Understanding and addressing the usability challenges of machine learning in high-stakes decision making. *IEEE Trans. Vis. Comput. Graph.*, 28(1):1161–1171.

Appendix

A Properties: Discussion

In this section, we thoroughly discuss the properties and categorization of interpretability methods presented in §4. We aim to provide the stakeholders' perspective, deepening our understanding of how these properties align with their objectives and requirements. We begin by discussing the *what* aspect properties in §A.1, followed by the *how* aspect properties in §A.2.

A.1 *What* Properties

A.1.1 The Explained Mechanism

In Appendix §C.1, we formally define what a *mechanism* is. Broadly, a mechanism can refer to the entire NLP system or a specific process or component within it. To better categorize interpretability methods, we distinguish four types of mechanisms. While most methods explain the whole system (an ***input-output*** mechanism), other methods explain input representations or hidden states (an ***input-internal*** mechanism). Another mechanism type focuses on explaining the functionality of internal components such as neurons, attention heads, circuits, and more (an ***internal-internal*** mechanism).

In addition, the mechanism property covers any abstraction of the mechanism states (see §C.2), for example, explaining the impact of concepts conveyed in the text input instead of explaining long and complex raw input. In this case, which is thoroughly discussed in the next subsection §A.1.2, the explained mechanism is ***concept-output***.

The choice of which mechanism to explain depends on the why: the objective of the explanation and the stakeholder's needs. Stakeholders mostly utilize methods that explain the full system (an *input-output* mechanism). However, many are interested in other mechanisms. For example, developers aim to understand the functionality of internal components such as neurons or layers to modify and edit factual knowledge encoded by them (Hase et al., 2023). Scientists might explore the representational space, for example, neuroscientists examine the brain by aligning model representations with brain activity (Tikochinski et al., 2024), and social scientists cluster representations to monitor opinions, such as attitudes towards COVID-19 vaccines (Hristova and Netov, 2022).

A.1.2 Raw Input or Abstracted Input

A common interpretability paradigm is feature attributions, where each input feature is assigned an importance score reflecting its relevance to the model prediction. In computer vision, the raw inputs consist of pixels, and feature attributions effectively highlight relevant areas that can be immediately and intuitively grasped (Alqaraawi et al., 2020; Müller, 2024). In contrast, explaining the raw input in NLP, often a lengthy and complex text, presents distinct challenges. For end-users, assigning scores to each token can be overwhelming as the cognitive load increases with the text length.

Instead, simplifying the system by abstracting the input to concepts or a summary, thus reducing the number of features explained, could lead to a better mental model of the system (Poursabzi-Sangdeh et al., 2021). For example, concept counterfactual methods (see §B.5, Feder et al. (2021) and Gat et al. (2023)) change a specific concept conveyed in the text. By contrasting the counterfactual predictions with the original prediction, we can gain digestible insights into how the concept impacts the prediction (a ***concept-output*** mechanism). Moreover, due to the vast space of textual data, providing global explanations by explaining the raw input is challenging. In contrast, concept-level explanations naturally support global explanations.

A.1.3 Scope: Local or Global

This categorization is based on the scope of the explanation: ***local*** or ***global***. A local explanation describes the mechanism for an individual instance. For example, feature attributions and attention visualizations (§B.1). Conversely, global explanations describe the mechanism for the entire data distribution, for example, probing (§B.2) and mechanistic interpretability (§B.3). Many local explanations can be generalized into global ones. For example, concept counterfactuals (§B.5) measure the causal effect of a concept on the prediction of an individual instance. A global average causal effect estimation can be derived by iterating the entire dataset and applying adjustments (Gat et al., 2023).

The choice of scope, local or global, depends on the objectives of the explanation and its stakeholders. For instance, developers debugging edge cases may prefer local explanations. Conversely, when aiming to improve the functionality of model components, developers might lean towards global explanations offered by mechanistic interpretability. End-users, such as clients and customers, require local explanations since they are concerned with decisions directly affecting them; this local need is also reinforced by the "right to explanation" (Goodman and Flaxman, 2017). Similarly, physicians using NLP systems must rely on local explanations. On the other hand, business decision-makers and scientists generally favor global explanations, which help identify broader trends and underlying patterns. From a social perspective, global explanations hold more significance. However, accumulating local evidence can progressively provide insights into global tendencies.

A.2 *How* Properties

A.2.1 Time: Post-hoc or Intrinsic

This property distinguishes between methods based on the time the explanation is formed. ***Post-hoc*** methods produce explanations after the prediction and are typically external to the explained model. Conversely, ***intrinsic*** methods are built-in; the explanation is generated during the prediction, and the model relies on it. Intrinsic methods include, for example, natural language explanations (§B.6) or self-explaining models (§A.1.2) such as concept bottleneck models, which train a deep neural network to extract human-interpretable features, which are then used in a classic transparent model (e.g., logistic regression).

In the XAI literature, this distinction also defines the difference between explainable AI (post-hoc) and interpretable AI (intrinsic) (Rudin, 2018; Arrieta et al., 2020). However, interpretable AI generally refers to transparent models (see (Lipton, 2018)), while in our categorization, intrinsic models can be opaque to some extent: in self-explaining methods, an opaque neural network extracts human-interpretable features; similarly, in natural language explanations, the explanation is generated by an opaque neural network. Intrinsic methods aim to produce more faithful and understandable insights and could better serve all stakeholders. However, they may also limit model architecture and thus could potentially degrade system performance, although this is not always the case (see Badian et al. (2023) for an example).

A.2.2 Access: Model Specific or Agnostic

This property distinguishes interpretability methods based on their access to the explained model. ***Model-agnostic*** methods do not assume any specific knowledge about the model and can only access its inputs and outputs. For example, diagnostic sets (§B.4), perturbation-based attributions (§B.1), or some counterfactual methods (§B.5). The latter two modify only the input and measure its impact on model prediction. On the other hand, ***model-specific*** methods require access to the explained model during the training time of the interpretability method. They can also access its internal components and representations. Hence, while a model-specific method can be applied only to one explained model, the same model-agnostic method can be applied to any model simultaneously.

Unlike model-specific methods, model-agnostic methods can not explain internal mechanisms. However, they can still be extremely valuable for some stakeholders. From an algorithmic perspective, they are useful during model selection and deployment. For example, developers juggling multiple models can easily rank them based on their vulnerability to confounding biases, such as gender bias. Moreover, regulators would prefer model-agnostic methods, utilizing a dedicated diagnostic set or a pool of counterfactuals to verify whether the model meets the required standards.

A.2.3 Presenting Insights

The presentation of insights extracted by the interpretability method falls under the *communicating* aspect of the explanation definition in §3.2. There

is extensive research in the XAI field that explores this aspect and examines how the presentation affects different stakeholders (Hohman et al., 2019; Schulze-Weddige and Zylowski, 2021; Bove et al., 2022; Karran et al., 2022; Zytek et al., 2022). Even though we do not delve into the stakeholder perspective, we still discuss this property since not all methods support every form of presentation. The design of interpretability methods and the choice of which to use depend on it.

The most common form of presentation is ***scores***, such as importance scores (§B.1), causal effects (§B.5) or metrics (§B.2 and §B.4). Scores are typically visualized using colors (Gat et al., 2022) or bar plots (Kokalj et al., 2021). Another form is ***visualization***, which includes means such as heatmaps (Jo and Myaeng, 2020), graphs (Vig, 2019), and diagrams (Katz and Belinkov, 2023). Others present similar or contrastive ***examples*** to stakeholders, along with their prediction, aiding in speculating on *why P and not Q?*. Such example presentations are found in counterfactual methods (§B.5) and KNN-based nets (§B.7). Finally, insights can also be conveyed through ***texts*** written in natural language (e.g., Menon et al. (2023) and §B.6).

A.2.4 Faithfulness and Causality

Note that some applications of interpretability methods are satisfied by correlational insights (*what knowledge the model encodes*), e.g., in a case when scientists explore new hypotheses which will then be validated in a controlled experiment (see (Lissak et al., 2024b)). However, most applications seek to understand the reasons behind specific predictions. In this context, faithfulness becomes a crucial principle, demanding that explanations accurately reflect the system's decision-making process (Jacovi and Goldberg, 2020). Unfaithful explanations, particularly those that seem plausible, can be misleading and dangerous and lead to potentially harmful decisions. As such, faithfulness is crucial in scenarios involving decision-makers and end-users. To ensure that explanations are faithful, establishing causality is essential (Feder et al., 2022). Indeed, Gat et al. (2023) theoretically demonstrated that non-causal methods often fail to provide faithful explanations.

A key approach to providing faithful explanations involves incorporating techniques from the causal inference literature, such as counterfactuals (Feder et al., 2021), interventions (Wu et al., 2023b), adjustment (Wood-Doughty et al., 2018), and matching (Zhang et al., 2023). Therefore, an important property of an interpretability method is whether it is ***causal-based*** or ***not***. We note that this categorization is not included in Table 1 as it pertains more to specific methods rather than to a paradigm. For a comprehensive survey on faithfulness in NLP interpretability, see Lyu et al. (2022).

B Common Interpretability Paradigms

This section aims to establish a clear link between the properties introduced in §4 and interpretability methods. To this end, we comprehensively review common interpretability paradigms, detailing relevant methods and works within each and explaining the paradigm's properties. Note that some methods may fall under multiple paradigms.

Our classification of methods into paradigms is inspired by previous surveys on model analysis (Belinkov and Glass, 2019), local methods (Luo et al., 2024), post-hoc methods (Madsen et al., 2023), faithful methods (Lyu et al., 2022), mechanistic interpretability (Räuker et al., 2023; Bereska and Gavves, 2024), LLMs (Singh et al., 2024; Zhao et al., 2024), and others (Danilevsky et al., 2020; Balkir et al., 2022; Sajjad et al., 2022a). Furthermore, while the categorization of the properties captures the standard characterization each paradigm, there may be exceptions with some methods.

B.1 Feature Attributions

Categorization

What: input-output or concept-output, local
How: post-hoc, specific or agnostic, scores

Feature attribution methods measure the relevance (sometimes referred to as importance) of each input feature, primarily tokens or words, and are a widely used ***local*** interpretability paradigm. Each input feature is assigned a ***score*** reflecting its relevance to a specific prediction, thus describing an ***input-output*** mechanism. Various attribution methods have been developed, which can be mainly categorized into four types.

Perturbation-based methods work by perturbing input examples, such as removing, masking, or altering input features at various levels, including tokens, embedding vectors, or hidden states (Wu et al., 2020; Li et al., 2016). Those are ***model-agnostic*** methods since the perturbations are applied to the input. In contrast, the following methods are ***model-specific***: *Gradient-based*

methods measure relevance via a regular backward pass (backpropagation) from the output through the model (Smilkov et al., 2017; Sikdar et al., 2021; Gat et al., 2022; Enguehard, 2023). *Propagation-based* methods define custom rules for different layer types (Montavon et al., 2019; Voita et al., 2021; Chefer et al., 2021). Other methods involve *surrogate models*, such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), which locally approximate a black-box model with a white-box surrogate model (Kokalj et al., 2021; Mosca et al., 2022). Rarely, the features are mapped into concepts (Yeh et al., 2020), describing a ***concept-output*** mechanism.

We also include here *attention-based* explanations, which aim to capture meaningful correlations between intermediate states of the instance (Jain and Wallace, 2019; Kovaleva et al., 2019; Wiegreffe and Pinter, 2019), because typically the intermediate state is represented by its corresponding token. Usually, attention-based explanations are presented with visualizations such as heatmaps.

B.2 Probing and Clustering

Categorization

What: input-internal, global
How: post-hoc, specific, scores or text

Probing typically involves training a classifier that takes the representations of the explained model and predicts some property (Belinkov, 2022), making it a ***post-hoc model-specific*** method. Typically, the predicted concept is a syntactic or semantic property (Adi et al., 2017; Conneau et al., 2018; Hewitt and Liang, 2019; Lepori and McCoy, 2020; Ravichander et al., 2021; Antverg and Belinkov, 2022; Amini et al., 2023; Vulic et al., 2023). Probing methods usually answer questions of how extractable a property is from a representation or what knowledge a model encodes. Thus, it can ***globaly*** describe the ***input-internal*** mechanism. However, even though the model encodes some property, it does not mean it uses it for prediction (Belinkov, 2022). Therefore, how we communicate probing insights to the stakeholders is important.

In the scope of probing, we also include *clustering* methods. While most clustering methods are used to discover patterns in data, here, clustering is employed to explore the model's learned space and gain insights about what it has encoded. Clustering is considered the unsupervised counterpart of probing (Michael et al., 2020; Gupta et al., 2022), and they share the same categorization: both methods explore the *input-internal* mechanisms of the system and are characterized as *global*, *post-hoc*, and *model-specific*. After representations are clustered, explanations are provided through cluster descriptions defined by gold labels, top keywords, concepts, topic modeling, ontologies, or LLM-generated text (Aharoni and Goldberg, 2020; Zhang et al., 2022; Thompson and Mimno, 2020; Gupta et al., 2022; Sajjad et al., 2022b; Alam et al., 2023; Mousi et al., 2023; Wang et al., 2023c; Hawasly et al., 2024; Lissak et al., 2024b). Finally, we also include works that explain representation-based similarity using concepts and semantic aspects (Opitz and Frank, 2022).

B.3 Mechanistic Interpretability

Categorization

What: internal-internal, global
How: post-hoc, specific, visualiztion or text

In contrast to probing, which is a top-down approach (i.e., we know in advance what we are looking for), mechanistic interpretability is a bottom-up approach that studies neural networks through analysis of the functionality of internal components of the NLP systems such as neurons, layers, and connections (Sajjad et al., 2022a; Räuker et al., 2023; Bereska and Gavves, 2024). The goal of such methods is to ***globaly*** explain one ***internal-internal*** mechanism of a ***specific*** model. Many mechanistic interpretability methods study how neurons respond to stimuli (real or synthetic examples) and ***visualize*** or ***describe*** the sensitivity of the neuron's activations (Finlayson et al., 2021; Vig et al., 2020; Geiger et al., 2021, 2022; Dai et al., 2022; Conmy et al., 2023; Garde et al., 2023; Gurnee et al., 2024).

Other works perturb or intervene in neurons to study their functionality (Bau et al., 2019; Ghorbani and Zou, 2020; Wang et al., 2023b), or mask network weights (Zhao et al., 2020; Csordás et al., 2021). Some works focus on gradients instead of activations (Durrani et al., 2020; Syed et al., 2023; Kramár et al., 2024) or train sparse autoencoders in an attempt to disentangle features, which are then described (Cunningham et al., 2023; Yu et al., 2023). Another line of work explores which information the internal states encode by projecting them into the vocabulary (Geva et al., 2022; Dar et al., 2023; Belrose et al., 2023; Pal et al., 2023;

Sakarvadia et al., 2023; Ghandeharioun et al., 2024) or even by generating images (Toker et al., 2024).

B.4 Diagnostic Sets

Categorization

What: input-output, global
How: post-hoc, agnostic, scores

Diagnostic sets, also known as challenge sets, probing sets, or test suites, are specialized collections of data designed to analyze specific properties of the NLP system or challenging cases. These sets are typically curated manually to target specific aspects of system behavior within a predefined NLP task, enabling the identification of strengths, weaknesses, and biases (Belinkov and Glass, 2019). Diagnostic sets are ***model-agnostic*** since they are curated independently from the analyzed model. They support ***scoring*** the model's predictive capabilities (***input-output*** mechanism) on subpopulations of interest, providing ***global*** insights on how it works within them. As one of the oldest techniques for analyzing NLP systems (King and Falkedal, 1990; Lehmann et al., 1996), diagnostic sets have been reintroduced as essential tools for understanding NLP models (Hill et al., 2015; Leviant and Reichart, 2015; Wang et al., 2019b; Vulic et al., 2020; Wang et al., 2019a; Gardner et al., 2020) and LLMs (Srivastava et al., 2022; McKenzie et al., 2023; Laskar et al., 2024). Rarely, diagnostic sets can be ***model-specific***. For example, the diagnostic dataset curated by Gekhman et al. (2024) involves examples not included in a specific LLM's pre-existing knowledge. Fine-tuning the same LLM using these examples increases hallucinations.

Many diagnostic sets are employed to examine linguistic phenomena (Burchardt et al., 2017; Burlot and Yvon, 2017; Sennrich, 2017; White et al., 2017; Giulianelli et al., 2018; Gulordava et al., 2018; Jumelet and Hupkes, 2018; Ravichander et al., 2020; Newman et al., 2021; Sullivan, 2024), while others evaluate biases such as gender bias (Waseem and Hovy, 2016; Webster et al., 2018; Zhao et al., 2018; De-Arteaga et al., 2019; Dhamala et al., 2021; Doughman and Khreich, 2022), cultural bias (Ventura et al., 2023; Chiu et al., 2024; Rao et al., 2024), and political bias (Smith et al., 2022; Taubenfeld et al., 2024). Beyond manually collecting diagnostic datasets or using simple rule-based programs, generative models are also being applied (Goel et al., 2021; Ribeiro et al., 2021; Ross et al., 2022). Importantly, these sets are crucial not only for evaluating the performance of NLP systems on specific examples or subpopulations but also serve as foundational elements in many probing and mechanistic interpretability methods.

B.5 Counterfactuals and Adversarial Attacks

Categorization

What: input-output or concept-output, local or global
How: post-hoc, agnostic or specific, scores or examples

The term *counterfactual (CF)* is frequently used in the NLP literature, often referring to various concepts. In this subsection, we aim to align the community's understanding of this term and clearly distinguish between CF-based methods. In the context of NLP, we adopt the following definition, which captures the fundamental characteristic common to all CF-based methods: *"a counterfactual for a given textual example is a result of a targeted intervening on the text while holding everything else equal."* (Calderon et al., 2022; Gat et al., 2023). The primary distinction among CF-based methods lies in the type of question the CFs aim to answer.

From a philosophical perspective, CFs answer *what-if* questions: *'If X had been different, then Y would be...'*. Presenting an ***alternation (CF)*** of the input example to stakeholders allows for speculation on the ***input-output*** mechanism: *'Why prediction A and not B?'* (Miller, 2017; Wu et al., 2021).

From a causal inference perspective, CFs answer questions such as *'How does C impact Y?'*, which can then help derive a ***score*** quantifying the causal effect of some concept *C* on the prediction: a ***concept-output*** mechanism (Abraham et al., 2022; Feder et al., 2022; Wu et al., 2023a).

Contrastive Examples. These methods address *what-if* questions and can explain a ***local*** prediction by ***presenting CFs*** to stakeholders. They typically focus on minimally editing the text to change the model prediction. The edited texts are commonly known as *contrastive examples*. Most approaches for generating contrastive examples are ***model-agnostic***. For instance, asking annotators to write them manually (Gardner et al., 2020; Kaushik et al., 2020; Sen et al., 2023), utilizing a generative model and applying edit operations (Wu et al., 2021; Ross et al., 2022; Li et al., 2024; Nguyen et al., 2024), or generating text until a proxy predictor indicates the

label has changed (Ross et al., 2021; Chemmengath et al., 2022; Filandrianos et al., 2023; Treviso et al., 2023; Bhan et al., 2024).

Adversarial Attacks. A prominent ***model-specific*** approach for generating contrastive examples is known as *adversarial attacks*, in which carefully crafted modifications barely noticeable to humans (e.g., a typo, extra space, or punctuation, etc...) are applied to the input and change the system predictions (Morris et al., 2020; Goyal et al., 2023). These attacks are typically generated through gradient-based token replacement (Ebrahimi et al., 2018; Li et al., 2019; Guo et al., 2021), and character-level perturbations (Belinkov and Bisk, 2018; Yang et al., 2020; Rocamora et al., 2024). With LLMs, the focus is on adversarial prompts that break model alignment (Perez et al., 2022; Zhu et al., 2023; Samvelyan et al., 2024; Paulus et al., 2024). Note that most applications of contrastive examples in the NLP literature, particularly adversarial attacks, are for data augmentation to improve model generalization or red teaming (Chen et al., 2021; Kaushik et al., 2021; Dixit et al., 2022; Balashankar et al., 2023; Zhao et al., 2023b; Sachdeva et al., 2024; Zhang et al., 2024b).

Concept Counterfactuals. The second group of CF-based methods, which address *How does C impact Y?* questions, is more theoretically grounded in the causal inference literature, making them more faithful (Lyu et al., 2022; Gat et al., 2023). Besides ***presenting*** stakeholders with explanations similar to contrastive examples, which allows for speculation on what would have happened if a concept *C* were different (e.g., a different gender of the writer), concept CFs can also be used to estimate the causal effect of high-level concepts on model predictions (Abraham et al., 2022; Feder et al., 2022). This is typically done by calculating the difference between the model's predictions for the original text and the counterfactual (CF) input.

In addition to providing a ***local score*** for an individual instance, concept CFs can deliver a ***global average causal effect*** estimation by iterating through the entire dataset and applying certain adjustments (Gat et al., 2023). The objective of the global score, similar to diagnostic sets, is to examine model behavior on subgroups. However, the score derived from CFs offers greater fidelity by relying on causation rather than correlation (Elazar et al., 2022; Keidar et al., 2022; Li et al., 2022; Liu et al., 2023; Wang et al., 2023a; Madaan et al., 2023; Zhou and He, 2023; Elazar et al., 2024).

Typically, a causal graph describing the input and output data-generating processes is provided, and an approximated counterfactual (CF) is generated by intervening on the concept of interest and adjusting for confounders (Feder et al., 2021; Gat et al., 2023). ***Model-agnostic*** methods focus on generating coherent, human-like CFs, either through controlled text generation (Calderon et al., 2022; Fang et al., 2023; Hong et al., 2023; Howard et al., 2022; Zheng et al., 2023) or by prompting LLMs (Gat et al., 2023; Feder et al., 2023; Zhang et al., 2024a). An alternative to the computationally intensive generation process is *causal matching*, where the example is paired with a similar control example that has a different concept value (Roberts et al., 2020; Zhang et al., 2023; Gat et al., 2023). In contrast, ***model-specific*** methods typically intervene on the latent space of the explained model (Ravfogel et al., 2020; Feder et al., 2021; Elazar et al., 2021; Haghighatkhah et al., 2022; Kumar et al., 2022; Wu et al., 2023a; Zhao et al., 2023a), or train a proxy model that mimics the CF behavior of the explained model (Wu et al., 2023a).

B.6 Natural Language Explanations

Categorization

What: input-output, local
How: intrinsic, specific, text

We define *Natural Language Explanations (NLE)* as any ***textual explanation*** extracted or generated by an NLP system that is used for justifying its own prediction. We do not consider generative models used to explain other model predictions as an NLE method. Thus, all NLE methods are ***model-specific***, ***intrinsic***, and ***local*** as they explain a single prediction. Usually, human-written explanations are used as an additional training signal for supervision (Wiegreffe and Marasovic, 2021; Sun et al., 2022; Kim et al., 2023).

NLE can be *abstractive* (by generating free-text) or *extractive* (by highlighting spans of relevant text in the input). The term *rationale* is often used in the extractive context to describe short and sufficient input spans for making a correct prediction (Zaidan et al., 2007). In addition, and following Camburu et al. (2018); Kumar and Talukdar (2020); Lyu et al. (2022), we divide NLE into *explain-then-predict* and *predict-and-explain* methods.

The *explain-then-predict* category comprises

methods that extract or generate an explanation and then independently predict the output by conditioning solely on the explanation, typically by training explainer and predictor components separately (Lei et al., 2016; Bastings et al., 2019; Camburu et al., 2020; Jain et al., 2020). The *predict-and-explain* category includes methods that explain and predict simultaneously (i.e., the output is predicted based on both the input and the explanation, such as chain-of-thoughts (CoT)) or first predict and then provide an explanation (Ling et al., 2017; Rajani et al., 2019; Narang et al., 2020; Marasovic et al., 2022), including explanations that reflect uncertainty (Xiong et al., 2023; Zhou et al., 2024). This category covers all the recent and commonly used CoT methods (Chu et al., 2023; Lyu et al., 2023).

In the era of LLMs, which are used daily by numerous end-users, NLE (either through CoT or explicitly asking the LLM to explain its output) has become the de facto method for explaining LLM outputs, despite being considered unfaithful (Lanham et al., 2023; Turpin et al., 2023). Moreover, NLE helps address challenges in explaining generative models since many interpretability methods were designed to explain a single decision rather than a sequence of decisions (a generated text).

B.7 Self-explaining Models

Categorization

What: input-output or input-concept-output, local or global

How: intrinsic, specific, scores or examples or text

Classic machine learning models, such as linear models, decision trees, Hidden Markov Models (HMMs), and Topic Models are often called transparent or whitebox models due to their simple structure and well-studied nature. These models represent the highest degree of self-explanation because explaining their decision-making process is relatively straightforward. Drawing inspiration from them, researchers attempt to design neural models with more structural transparency while maintaining their performance (Rajagopal et al., 2021; Das et al., 2022; Su et al., 2023).

An example is concept bottleneck models, which train a deep neural network to extract human-interpretable features and then apply a classic transparent that takes these features as an input, sometimes simultaneously. Concept bottleneck models describe relations of ***input-concepts and concepts-output***. The interpretable features used for training the network can be manually annotated (Koh et al., 2020; Rezaii et al., 2022; Tan et al., 2024), defined by domain experts and automatically extracted using an LLM (Badian et al., 2023), or automatically discovered and annotated (Yeh et al., 2020; Ludan et al., 2023). In concept bottleneck models, explanations can be ***global***, such as the ***linear regression weights*** of concepts, or local. In the case of ***local*** explanations, they are provided with respect to the predicted concepts of a specific instance. KNN-based networks, for example, replace the final softmax classifier head with a KNN classifier at test time (Papernot and McDaniel, 2018; Wallace et al., 2018; Sarwar et al., 2022). The ***local*** explanations in KNN-based networks are ***example-based***.

Another prominent line of works focuses on neural module networks, which decompose the task into small interpretable steps, which are then presented to the stakeholder (Andreas et al., 2016; Hu et al., 2017; Santoro et al., 2017; Gupta et al., 2020). Similarly, other methods break down the input into "atoms" and then combine the atom-level solutions to reach a final decision (Stacey et al., 2022, 2024). Presenting such decompositions helps in understanding the decision-making process.

Note that models that ***extract or generate explanations*** during their predictions are self-explaining models and are covered in §B.6.

C Mechanism and Understandable Terms

C.1 What is the Explained Mechanism?

> **Mechanism:**
> *A process that constitutes a relation between two states of the NLP system.*

To complete the definition, A *state of an NLP system* refers to any form of data at any stage within the data analysis process of the system. This includes the initial state, encompassing the raw input received, all intermediate states comprising various levels of transformed data, and the final state, the system's output or decision. For example, the raw input, tokenized input, embeddings, hidden states (of a specific layer), activations, attention scores, logits, output, decision. Accordingly, the mechanism we explain is defined by two system states. For instance, the mechanism between a sentence and the final output is the whole NLP model; the mechanism between the representations of the third layer and those of the fourth layer is the fourth

layer; the mechanism between the raw input and the tokenized input is the tokenizer.

Notably, the explained mechanism does not need to encompass the entire NLP system. It is acceptable for the mechanism to be only a subsystem or a component. Furthermore, it is acceptable for an explanation to be partial with respect to the mechanism. In other words, the explanation may provide specific insight into the mechanism without fully explaining every aspect and functionality. For example, a scientist who wishes to validate a hypothesis might only be interested in the impact of one concept (e.g., how tone impacts the popularity of social media content (Tan et al., 2014)). The idea of not providing a complete explanation is also grounded in the philosophy, psychology, and cognitive science literature. For instance, Miller (2017) advocates that explanations can be selective (humans select a few salient causes instead of a complete causal chain when explaining) and contrastive (Explanations should answer *Why P instead of Q?* rather than *Why P?*).

C.2 What are Understandable Terms?

> **Understandable terms:**
> *The level of abstraction of the states in the mechanism we explain.*

Note that in our description states can be either fully specified or abstracted to some extent. For example, if the input state is the text, then the interpretability method may consider the entire text, but it may also consider abstractions of the text, such as its summary or a list of concepts conveyed in the text. This also holds for the output state. For example, in probing methods (see §B.2), a classifier is trained to predict a property (often a linguistic property) from the representations of a particular layer of the model to provide insights into the knowledge encoded in model representations (Belinkov, 2022). Accordingly, the *input-representations* mechanism we explain is the part of the model that transforms input data into the probed layer's representations, and the output state of the mechanism (the representations) is abstracted to a property. For our convenience, we henceforth use the terminology of a *state* for describing a *fully specified state* or an *abstracted state*, remembering that a state may have several different possible abstractions.

The degree of *"understandable terms"*, the level of abstraction, or the form of cognitive chunks (Doshi-Velez and Kim (2017) define them to be the basic unit of an explanation) depends on the stakeholder and their specific needs, as they are the ones who utilize the explanation. This involves considering their level of expertise and familiarity with NLP models. For example, mechanistic interpretability methods (see §B.3) aim to explain states of internal components like neurons, targeting developers (Bereska and Gavves, 2024). While these terms are unsuitable for end-users, they can meet the *"understandable"* criterion for developers, even without abstractions.

	Para.	Mech.	Scope	Access.
Agreements	92%	93%	81%	92%
Disagreements with unknowns	12%	29%	69%	62%
Agreements without unknowns	93%	95%	92%	97%

Table 4: Agreement statistics between human and LLM annotations of different characteristics: *Paradigm*, *Mechanism*, *Scope* and *Accessibility*. The first row presents the portion (in percentages) of agreements. The second row presents the portion of disagreements that involve an *'unknown'* annotation (e.g., the LLM annotated the method scope as unknown, but sufficient domain knowledge could infer it.) within the disagreements. The third row presents the portion of agreements, excluding disagreements involving unknowns. **Additional statistics:** 96% of the papers annotated as relevant by the LLM were indeed relevant. 98% of the *Field* annotations were correct. 100% of the *Causal-based* property and of the *LLM* field (whether the paper employs an LLM, see Table 3) annotations were correct.

D Additional Analysis Details

Retrieval: We retrieved tens of thousands of NLP interpretability papers using the Semantic Scholar API and by searching queries such as `NLP interpretability` (a full list of queries is provided in Box D.1). We kept only papers whose titles or abstracts contained at least one NLP keyword (e.g., `NLP`, `LLM`, `BERT`; see Box D.2) and one interpretability keyword (e.g., `interpretability`, `XAI`, `explanation`; see Box D.3). This search and selection process yielded 14,676 papers.

Annotation and Filtering: For determining the relevancy of the papers and annotating them, we employed an LLM (`gemini-1.5-pro-preview-0514`) and used the zero-shot

prompt provided in Box D.4. We asked the LLM to determine the relevance of the paper, its field, the paradigm of the interpretability method, the mechanism being explained, the scope and accessibility of the method, and whether it is causal-based. Additionally, we asked the LLM to write a one-sentence summary of the paper and explain its paradigm annotation. The LLM was also instructed to explicitly extract the names of the interpretability methods employed in the paper. We generated three responses (in a JSON format with LLM annotations) for each paper and determined the final annotation of each question by the majority vote. After relevancy filtering, 2,009 papers remained.

Correction: We then sampled and examined a subset of 20 annotated papers. Following this, we decided to apply some automatic rules to fix the annotations: (1) We merged the 'computer science' field with the 'NLP' field; (2) For the mechanism annotation, we replaced internal components with 'internal-internal', and representations with 'input-internal'; (3) Many of the scope annotations were 'unknown'. In these cases, we replaced 'unknown' with 'local' for feature attributions and natural language explanation paradigms and with 'global' for probing, diagnostic sets, and mechanistic interpretability paradigms; (4) We replaced 'unknown' values of the accessibility annotations with 'model-specific' for the SHAP/LIME, probing and mechanistic interpretability paradigms, and with 'model-agnostic' for the diagnostic sets paradigm. (5) Initially, we instructed the LLM to determine whether an LLM was employed in the paper. However, it frequently misclassified models such as BERT as LLMs. To improve accuracy, we instead searched the abstracts for specific keywords such as `LLM, GPT4, ChatGPT, Gemini, Llama`; (6) Since 2024 is not over, we adjusted the publication year of the papers such that each year spans from June of the previous year to the following June.

Verification: To verify the accuracy of the LLM annotations, we randomly sampled another 100 papers, which one of the authors manually annotated. The agreement statistics are presented in Table 4. Note that many disagreements between human and LLM annotations involved an 'unknown' LLM annotation (the second row in Table 4 shows the proportion of such disagreements among all disagreements). For example, the LLM annotated the method scope as unknown, but sufficient domain knowledge could infer it. When excluding unknown disagreements, over 92% of the annotations for each question were correct. Excluding unknown disagreements when computing the agreement statistics is reasonable since we exclude 'unknown' annotations in our analysis in §5.

Paradigm	NLP			Outside		
	#	%	C	#	%	C
Attributions	491	32.8	20.6	200	39.0	9.7
LIME/SHAP	65	4.3	7.4	49	9.6	4.8
Probing	168	11.2	17.9	32	6.2	19.6
Clustering	35	2.3	10.5	50	9.7	6.0
Mechanistic	167	11.2	27.3	9	1.8	8.6
Diagnostic	54	3.6	17.5	12	2.3	4.9
Adversarial	76	5.1	53.1	4	0.8	6.8
Counterfactuals	47	3.1	24.1	4	0.8	0.5
Lang. Expl.	222	14.8	13.2	77	15.0	4.7
Self-explain	98	6.6	15.7	25	4.9	3.6
Classic	9	0.6	0.6	13	2.5	1.5
Unknown	64	4.3	32.3	38	7.4	6.7
Total	1495	100%	20.9	514	100%	7.8

Table 5: Absolute numbers (#), proportions (%), and average number of citations (C) of interpretability paradigm papers by field (**NLP** and fields **Outside** NLP) including all papers from 2015 to 2024.

Box D.1: Queries for semanticscholar search

NLP interpretability, NLP model interpretability, LLM interpretability, LLMs interpretability, language models interpretability, interpretability for NLP models, interpretability for NLP, interpretability for LLMs, interpretability for language models, NLP explainability, NLP model explainability, LLM explainability, language models explainability, explainability for NLP models, explainability for NLP, explainability for LLMs, explainability for language models, explaining NLP models, explaining LLMs, explaining language models, interpreting NLP models, interpreting LLMs, interpreting language models, NLP explanation, NLP model explanation, LLM explanation, LLMs explanation, language models explanation, explanation for NLP models, explanation for NLP, explanation for LLMs, explanation for language models, explanations for NLP models, explanations for NLP, explanations for LLMs, explanations for language models, NLP interpretation, NLP model interpretation, LLM interpretation, LLMs interpretation, language models interpretation, interpretation of NLP models, interpretation of LLMs, interpretation fo language models, black box NLP, black box NLP model, black box NLP models, black box LLM, black box LLMs, black box language models, black-box NLP, black-box NLP model, black-box NLP models, black-box LLM, black-box LLMs, black-box language models, white box NLP, white box NLP model, white box NLP models, white box LLM, white box LLMs, white box language models, white-box NLP, white-box NLP model, white-box NLP models, white-box LLM, white-box LLMs, white-box language models, NLP XAI, NLP model XAI, NLP models XAI, LLM XAI, LLMs XAI, language models XAI, XAI for NLP models, XAI for LLM, XAI for NLP, XAI for LLMs, XAI for language models, NLP explainable AI, LLM explainable AI, LLMs explainable AI, language models explainable AI, explainable AI for NLP models, explainable AI for LLM, explainable AI for NLP, explainable AI for LLMs, explainable AI for language models, explainable NLP models, explainable LLM, explainable NLP, explainable LLMs, explainable language models, interpretable AI for NLP models, interpretable AI for LLM, interpretable AI for NLP, interpretable AI for LLMs, interpretable AI for language models, interpretable NLP models, interpretable LLM, interpretable NLP, interpretable LLMs, interpretable language models, NLP user trust, user trust in NLP, user trust in NLP models, user trust in LLM, user trust in LLMs, user trust in language models, NLP transparency, NLP model transparency, LLM transparency, LLMs transparency, language models transparency, transparency of NLP models, transparency of LLMs, transparency of language models, transparent NLP, transparent NLP models, transparent LLMs, transparent LLM, transparent language models, trustworthy NLP models, trustworthy LLM, trustworthy NLP, trustworthy LLMs, trustworthy language models, NLP understanding, NLP model understanding, LLM understanding, LLMs understanding, language models understanding, accountability for NLP models, accountability for LLM, accountability for NLP, accountability for LLMs, accountability for language models, responsible AI for NLP models, responsible AI for LLM, responsible AI for NLP, responsible AI for LLMs, responsible AI for language models, responsible NLP models, responsible LLM, responsible NLP, responsible LLMs, responsible language models

Box D.2: NLP Keywords

nlp, language model, computatinal linguistics, language processing, llm, gpt, bert, llama

Box D.3: Interpretability Keywords

interpretability, explainability, explanation, interpretation, black box, blackbox, black-box, white box, whitebox, white-box, xai, explainable, user trust, interpretable, transparency, trustworthy, transparent, understanding, accountability

Box D.4: LLM prompt for annotating abstracts

You will be provided with the title and abstract of a paper focused on NLP model interpretability.
Carefully read both the title and the abstract. Your task is to extract key information regarding *only* the interpretability methods discussed in the paper.
Respond *only* in the JSON format below.
Please address the following questions and extract the specified information:

* "relevant" * - (bool) Determine if the paper is relevant if and only if an interpretability method is used, presented or proposed in the paper. If the paper does not discuss interpretability methods or uses one to explain results, the paper is not relevant. Answer true or false.

* "NLP research" * - (bool) Determine if the paper is related to NLP research, it can be that the paper is about domains other than NLP (e.g., medicine, social science, natural science, etc...), but uses NLP models with text input. Answer true or false.

* "LLM" * - (bool) Determine if an LLM is employed in the paper.

* "TL;DR interpretability method" * - (str) One sentence summarizing only the interpretability method used in the paper.

* "field" * - (str) Identify the research field of the paper, select from these options:
- "general NLP", "computer science", "medicine", "psychology", "neuroscience", "education", "engineering", "economics", "natural science", "humanities", "social science"
* "paradigm explanation" * - (str) One sentence explaining the interpretability paradigm used in the paper and justify your answer to the next question.

* "paradigm" * - (str) Select the paradigm of the interpretability method from the options below:
- "feature attributions": Measuring relevance or importance of each input feature (e.g., tokens or words), including methods like perturbations, gradients, propagations, attention scores and attention visualizations.
- "LIME/SHAP": Training and applying a surrogate model such as LIME or SHAP.
- "probing": Training a classifier from model representations that predict properties or concepts, or aligning model representations with signals (like brain activity).
- "clustering": Clustering the data with model representations or other clustering techniques such as Topic Modeling.
- "mechanistic": Explaining the functionality of internal components like weights, neurons, layers, attention heads, and circuits, using stimuli, activations, patching, scrubbing, logit lens, projections, etc.
- "diagnostic sets": Analyzing and evaluating the model using diagnostic sets, challenge sets, test suites, or subsets of examples with a common property (e.g., gender, culture).
- "adversarial attacks": Generating adversarial attacks or writing adversarial prompts that break alignment.
- "counterfactuals": Generating counterfactuals, contrastive examples, concept counterfactuals, causal matching and other causal-based methods.
- "natural language explanations": Providing natural language explanations, extractive or abstractive, including rationales and chain-of-thoughts.
- "classic": Classic and traditional ML models like Logistic Regression, Linear Regression, Decision Trees, Random Forest, XGBoost, SVM, HMM, KNN.
- "whitebox": Special model architectures, inherently explainable, that provide intrinsic explanations, such as Concept Bottleneck, Neural Module Networks, Knowledge Graphs, KNN-based.
- "unknown": If it cannot be inferred from the title and abstract.

* "methods" * - (list) List the interpretability methods mentioned in the paper. Note that there

might be more than one method.

* "explaining what" * - (str) Specify what the interpretability method explains in the model, does it epxlain the whole model (input-output), input concepts (concept-output), representations, or internal components. Select from the following options:
  - "input-output", "concept-output", "representations", "word embeddings", "neurons", "layers", "attention heads", "MLPs", "unknown"

* "causal" * - (bool) Determine if the abstract mentions the interpretability method is causal-based. Answer true or false.

* "local or global" - (str) Determine if the explanation is global (general insights about the model or the whole data) or local (explaining an individual example). Select from the following options:
  - "global", "local", "both", "unknown"

* "specific or agnostic" - (str) Determine if the explanation is model-specific (requires access to the model internals, or the interpretability method is trained using the explained model) or model-agnostic (does not require access to the model internals). Select from the following options:
  - "model-specific", "model-agnostic", "both", "unknown"

Answer format:
```json
{
  "relevant": bool,
  "NLP research": bool,
  "LLM": bool,
  "TL;DR interpretability method": str,
  "field": str,
  "paradigm explanation": str,
  "paradigm": str,
  "methods": list,
  "explaining what": str,
  "causal": bool,
  "local or global": str,
  "specific or agnostic": str
}
```

Title: [PAPER_TITLE]
Abstract: [PAPER_ABSTRACT]

Answer: