Mechanistic?

Naomi Saphra[*]
The Kempner Institute at Harvard University
nsaphra@fas.harvard.edu

Sarah Wiegreffe[*]
Ai2 & University of Washington
wiegreffesarah@gmail.com

Abstract

The rise of the term *mechanistic interpretability* has accompanied increasing interest in understanding neural models—particularly language models. However, this jargon has also led to a fair amount of confusion. So, what does it mean to be *mechanistic*? We describe four uses of the term in interpretability research. The most narrow technical definition requires a claim of causality, while a broader technical definition allows for any exploration of a model's internals. However, the term also has a narrow cultural definition describing a cultural movement. To understand this semantic drift, we present a history of the NLP interpretability community and the formation of the separate, parallel *mechanistic* interpretability community. Finally, we discuss the broad cultural definition—encompassing the entire field of interpretability—and why the traditional NLP interpretability community has come to embrace it. We argue that the polysemy of *mechanistic* is the product of a critical divide within the interpretability community.

1 Introduction

The field of mechanistic interpretability is growing dramatically, constantly motivating new workshops, forums, and guides. And yet, many are unsure what the term *mechanistic interpretability* entails. Researchers, whether experienced or new to the field, often ask what makes some interpretability research "mechanistic" (Andreas, 2024; Beniach, 2024; Hanna, 2024; Belinkov et al., 2023). Because both fields study language models (LMs), the distinction between traditional NLP interpretability (NLPI) and mechanistic interpretability (MI) is unclear. In fact, when work is labelled as *mechanistic* interpretability research, the label may refer to:

1. **Narrow technical definition:** A technical approach to understanding neural networks through their causal mechanisms.
2. **Broad technical definition:** Any research that describes the internals of a model, including its activations or weights.
3. **Narrow cultural definition:** Any research originating from the MI community.
4. **Broad cultural definition:** Any research in the field of AI—especially LM—interpretability.

Exacerbating this confusion, *mechanistic interpretability* in the narrow cultural definition describes the authors of a paper, rather than their methods or objectives. We must therefore discuss the landscape of the interpretability community itself in order to clarify the usage of *mechanistic interpretability*.

To that end, we will begin by characterizing the *narrow technical definition* (§2.1) and subsequently explain how the coinage of the term *mechanistic interpretability* led inevitably to its *broad technical definition* (§2.2). Both technical definitions characterize subsets of research from the NLPI community, but their work is not always classified as MI, illustrating the term's contextual application.

In order to understand how semantic drift eventually gave rise to the cultural definitions, we overview the history of the two distinct communities (NLPI and MI) (§3). We describe how a new movement of AI safety researchers, motivated by philosophical arguments for the importance of interpretability, differentiated themselves as the MI community in its *narrow cultural definition* (§3.2). The resulting financial and social context of the field now incentivizes NLPI researchers to bridge this gap by embracing the label in its *broad cultural definition* (§3.3).

Mechanistic is just one example of the imprecise and ambiguous language used in interpretability research. Although clarity is key for distilling and communicating insights about neural networks, we

[*] Equal contribution. Order chosen for aesthetics.

Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 480–498
November 15, 2024

compare it to a number of similarly vague terms in the history of NLPI (Appendix A). However, in contrast with other cases of lexical ambiguity in the area, we argue *mechanistic* is notable because it exposes a cultural divide—one which is worth bridging for the sake of scientific progress.

2 So what is *mechanistic*?

Before the term *mechanistic* described a cultural movement, NLPI researchers occasionally used the term *mechanisms* to refer to internal algorithmic implementation (Belinkov, 2018), as suggested by Marr's levels of analysis (Marr and Poggio, 1976). The earliest uses of *mechanistic interpretability* also draw on this technical meaning, as do most current explicit definitions of the term. What, then, is this precise technical meaning?

2.1 From causality and psychology to NLP

Mechanistic interpretability derives its name from *causal mechanisms*. In a causal model, a causal mechanism is a function—governed by "lawlike regularities" (Little, 2004)—that transforms some subset of model variables (causes) into another subset (outcomes or effects). Causal mechanisms are a necessary component of any causal model explaining an outcome (Halpern and Pearl, 2005a,b).

The *narrow technical definition* of MI thus describes research that discovers causal mechanisms explaining all or some part of the change from neural network input to output at the level of intermediate model representations. For example, one mechanistic interpretation explains how an LM can predict "B" from the input sequence "ABABA" using *induction heads* (Olsson et al., 2022): attention heads that search for a previous occurrence of "A" in combination with other heads that attend to the token that follows it. This narrow definition of MI requires causal methods of understanding, but excludes those that do not investigate intermediate neural representations, such as behavioral testing with input-output pairs (e.g., Ribeiro et al., 2020; Xie et al., 2022). It also excludes non-causal methods, such as describing representational structure or correlating activation features with particular inputs and outputs.

Psychology and philosophy have long stressed the importance of causal mechanisms in explanations. Psychology studies show (Legare and Lombrozo, 2014; Vasilyeva and Lombrozo, 2015) that humans prefer explanations containing causal mechanisms underlying an event over Aristotle's other modes (Lombrozo, 2016) of explanation. Tan (2022) argues that likewise, explanations in machine learning should focus on causal mechanisms linking input and output. Real-world causal models are complex and have many possible pathways to outcomes (Hesslow, 1988); complete explanations of such models would be burdensome and counterproductive. Therefore, explanation requires distillation (Jacovi and Goldberg, 2021). Human explanations distill by capturing only *proximal* mechanisms (Lombrozo, 2006)—those which are closest to, or immediately responsible for, the outcome.

Unlike the human brain, neural networks can be rigorously studied due to our ability to perform causal interventions on them. Because we are not limited to proximal mechanisms in our efforts to discover causal mechanisms in neural networks, we instead rely on *causal abstraction* (Beckers and Halpern, 2019; Beckers et al., 2020) for distillation: the theory that causal models at higher levels of granularity can be faithful simplifications of the true causal model, and thus serve as mechanistic interpretations of the network (Geiger et al., 2024a).

In an attempt to bring definitional rigor to MI, recent work in causal interpretability of neural networks has advocated for an even narrower technical definition of MI: explanation through a *complete end-to-end causal pathway from model inputs to outputs* via intermediate neural representations (Geiger et al., 2021, 2024b; Mueller et al., 2024). This definition excludes most early work in MI (§2.2), and has not yet been widely adopted. Induction heads, for example, only describe one component in the causal pathway—under the end-to-end definition, one would also need to explain how the model identifies the input "ABABA" as a 2-token repeating pattern, and then how the model predicts "B" after attending to earlier occurrences of it.

2.2 The coinage of *mechanistic interpretability*

The term *mechanistic interpretability* was coined by Chris Olah and first publicly used in the Distill.pub **Circuits thread**, a series of blogposts by OpenAI researchers between March 2020–April 2021. The first post (Olah et al., 2020) set out to "understand the *mechanistic* implementations of neurons in terms of their weights." After researchers involved in the Circuits thread moved to Anthropic, their subsequent reports (the **Transformer Circuits thread;** Elhage et al., 2021; Olsson et al., 2022; Hernandez et al., 2022; Elhage et al., 2022a) leaned into the terminology heavily and eventually it became mainstream.

Elhage et al. (2021) provided the first explicit definition of MI: "attempting to *reverse engineer the detailed computations* performed by Transformers, similar to how a programmer might try to reverse engineer complicated binaries into human-readable source code." The analogy to reverse engineering, building on Olah (2015)'s earlier comparisons to code via functional programming, has had staying power. Recent definitions, such as that of the ICML 2024 MI workshop (Barez et al., 2024), use similar wording:

> "...reverse engineering the algorithms implemented by neural networks into human-understandable mechanisms, often by examining the weights and activations of neural networks to identify circuits ... that implement particular behaviors."

While this definition still implicitly focuses on causal mechanisms (the key technical distinction one can draw between MI and some other subtypes of interpretability), current MI research rarely makes reference to causality. Reflecting both the emphasis above on examining weights and activations and the definition's lack of specificity about acceptable methods, many recent works have adopted a *broad technical definition* of MI to mean *any* inspection of model internals.[1] This semantic drift may have been inevitable—how could we reverse engineer a network without first inspecting its internal components? However, the further generalization of the term to label the output of a community, rather than its characteristic approach, was perhaps less inevitable.

3 How did we get here? A history of two LM interpretability communities

Our current terminological confusion results from a historical[2] accident: MI started as a movement with distinct technical objectives in computer vision, but ultimately moved into NLP without engaging the existing community which was already pursuing similar objectives.[3]

3.1 The nascent field of NLP Interpretability

NLP researchers published focused analyses of linguistic structure in neural models as early as 2016, primarily studying recurrent architectures like LSTMs (Ettinger et al., 2016; Linzen et al., 2016; Li et al., 2016; Hupkes et al., 2017; Ding et al., 2017). The growth of the field, however, also coincided with the adoption of Transformers, which were initially developed for machine translation and constituent parsing (Vaswani et al., 2017) but rapidly dominated rankings across standard NLP tasks (Radford et al., 2018, 2019; Devlin et al., 2019), drawing wide interest in understanding how they worked.

To serve the expanding NLPI community, the first BlackBoxNLP workshop (Alishahi et al., 2019) was held in 2018 and immediately became one of the most popular workshops at any ACL conference. Whereas many NLPI researchers had previously struggled to find an audience, ACL implemented an "Interpretability and Analysis" main conference track in 2020 (Lawrence, 2020), reflecting the mainstream success of the field.

In many ways, the early NLPI field—which related model behavior to particular components, layers, and geometric properties—would be familiar to anyone in the current MI community. Not only is current research often reinventing their methods and rediscovering their findings (§3.2.2), it is also repeating the same epistemological debates. These debates pit correlation against causation, simple features against complex subnetworks, and expressive mappings against constrained interpretations.

3.1.1 Distributional semantics and representational similarity

Interest in vector semantics exploded in the NLP community after word2vec (Mikolov et al., 2013a) popularized many approaches to interpreting word embeddings (Mikolov et al., 2013b,c).[4] The en-

[1]The minimal overlap between causality and MI has been previously noted (Mueller, 2024; Mueller et al., 2024).

[2]Note that our "history" turns on events barely two years before the time of writing. We are not overreaching, however, by assuming that many new researchers in this rapidly growing field are unfamiliar with its history. Popular MI tutorials and guides often begin their LM literature review in 2021-2022 (e.g., Docker and Nanda, 2023; Li, 2024; Nanda, 2024b), providing a limited window for many new entrants to the field.

[3]Here, we discuss MI and NLPI work under the *narrow cultural definition*. Although some of these MI papers fall outside of the *technical definitions*, most either self-label as MI or appear in MI venues. Regardless, not all of it is referred to as MI by the authors themselves, who may be more prescriptive in their own definitions. Our categorization of culture is based on the authors' networks and background: A paper's lead authors are MI if they entered the field through the MI or associated alignment community and NLPI if they are closely tied to the ACL interpretability community.

[4]These methods, first introduced in distributional semantics (Louwerse and Zwaan, 2009; Jurgens et al., 2012; Turney,

thusiasm for unigram embedding analysis proved transient, but still influences neural interpretability methods (Ethayarajh, 2019; Reif et al., 2019; Hernandez and Andreas, 2021). Distributional semantics has generalized to representational similarity methods (Saphra and Lopez, 2019b; Raghu et al., 2017; Wu et al., 2020) and vector space analogical reasoning has left clear marks on methods like task vectors (Ilharco et al., 2023) and steering vectors (Subramani et al., 2022; Turner et al., 2023). Many works in MI similarly leverage additive properties in representations (Marks and Tegmark, 2024; Tigges et al., 2023; Arditi et al., 2024).

Despite the brief excitement around distributional semantics, critics quickly noted that not all interesting properties of word embeddings correlated with downstream model behavior (RepEval, 2016). Furthermore, geometric analysis revealed these representations to be anisotropic and heavily influenced by word frequency rather than meaning (Mimno and Thompson, 2017). These critiques remain salient to modern correlational interpretability methods, including similarity-based metrics (Davari et al., 2023).

3.1.2 Attention maps

Attention, originally developed for recurrent machine translation models (Bahdanau et al., 2015), was rapidly adopted across language tasks. Even before the switch to fully attentional Transformers, attention modules offered new avenues of explanation (Bahdanau et al., 2015; Wang et al., 2016). In BERT models, the concurrent discovery of both a correlational (Clark et al., 2019; Htut et al., 2019) and causal (Voita et al., 2019) relationship between syntax and attention demonstrated the case for attention maps as a window into how Transformer LMs handled complex linguistic structure. However, NLPI researchers also identified some limitations of attention for interpretability (Serrano and Smith, 2019; Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Bibal et al., 2022). Some issues have longstanding solutions, such as incorporating the context of the model when computing attention metrics (Brunner et al., 2020; Kobayashi et al., 2020; Abnar and Zuidema, 2020).

MI work has continued to attribute specific stereotyped behavior to attention heads (Olsson et al., 2022) and to present attention patterns as input saliency maps (Wang et al., 2023; Lieberum et al., 2023; Hanna et al., 2023), though more frequently with results that are causally validated.

3.1.3 Neuron analysis and localization

Early works on localizing concepts in NLP often associated individual neurons with sentiment, syntax, bias, or specific token sequences (Radford et al., 2017; Na et al., 2019; Bau et al., 2019; Lakretz et al., 2019; Dalvi et al., 2019; Durrani et al., 2020). Many such studies validated their findings by using causal interventions, though few proposals were causal by design (Sajjad et al., 2022). MI research has largely pursued similar goals of localizing model behaviors to fine-grained model components, including neurons, through its focus on finding "circuits": groups of components forming a sub-network that closely (or faithfully) replicate the full model's performance on a fine-grained task (Olah et al., 2020; Wang et al., 2023).

Single neuron analysis has been subject to criticism arguing that it is infeasible to reduce large, complex models to the sum of their parts (Antverg and Belinkov, 2022; Sajjad et al., 2022). One core problem is polysemanticity: the observation that a single neuron can activate for multiple disparate classes or concepts (Olah et al., 2020; Mu and Andreas, 2020; Bolukbasi et al., 2021). Not only are these concepts ambiguous, but they can combine nonlinearly according to a sequence's underlying latent structure (Saphra and Lopez, 2020; Csordás et al., 2024), making them difficult to disentangle and isolate. MI struggles with many of the same neuron analysis concerns as earlier work, but has taken a particular interest in resolving polysemanticity (Elhage et al., 2022b; Gao et al., 2024). One popular method for this purpose, the sparse autoencoder (SAE) (Bricken et al., 2023; Cunningham et al., 2024), still relies on assumptions of linearity (Park et al., 2024; Millidge, 2023) and naturally emerging feature sparsity (Saphra and Lopez, 2019a; Puccetti et al., 2022; Elhage et al., 2023). Like earlier neuron analysis methods, it also requires expensive causal validation (Mueller et al., 2024).

3.1.4 Component analysis and probing

Probes were first applied in NLPI to extract linguistic information from the hidden states of neural models (Ettinger et al., 2016; Kádár et al., 2017; Shi et al., 2016; Adi et al., 2017; Hupkes et al.,

2012), had previously relied on Latent Semantic Allocation (Turney, 2005) or other word representations derived from matrix factorization—a class that also, implicitly, includes word2vec itself (Levy and Goldberg, 2014).

2017; Belinkov et al., 2017a,b; Giulianelli et al., 2018). Many early papers observed that lower layers encode local features, echoing findings in computer vision (Yosinski et al., 2014) and reflecting the classical NLP pipeline (Tenney et al., 2019).

The probing literature quickly came under scrutiny (Belinkov, 2022) for its lack of proper baselines (Hewitt and Liang, 2019) or informative structural constraints (Saphra and Lopez, 2019b). Without proper experiment design, many probing methods appeared to extract more information from random embeddings than from trained representations (Zhang and Bowman, 2018; Wieting and Kiela, 2019). To manage these issues, newer probes incorporated information complexity (Pimentel et al., 2020; Voita and Titov, 2020) or used simple geometric mappings (Hewitt and Manning, 2019; White et al., 2021). Some designs reflected the model's own processing (Pimentel et al., 2022), as now exemplified by methods like the logit lens (nostalgebraist, 2020; Geva et al., 2022; Chuang et al., 2024) used widely in MI research. However, the logit lens—like other probing methods before it (Belinkov, 2022)— has been criticized for providing a largely incomplete causal explanation (Nanda, 2024b; Zhu et al., 2024; Wiegreffe et al., 2024).

Recent MI work has focused on projecting LM hidden states to interpretable subspaces using linear probes. These probes may be supervised (Li et al., 2023; Marks and Tegmark, 2024) or unsupervised, using an SAE. These methods inherit many critiques from the classic probing literature, including a lack of causal grounding. Recent proposals have therefore argued for validation by causal interventions for SAEs (Mueller et al., 2024), echoing previous efforts to validate probed structures (Giulianelli et al., 2018; Elazar et al., 2021).

3.2 The beginnings of mechanistic interpretability

As NLPI researchers continued investigating language model features and weights, their community and scientific understanding grew rapidly. However, they could not have predicted how the field would grow and change with the infusion of MI researchers into the area. To fully understand the lexical landscape of the NLPI field, we must consider how *mechanistic* historically came to denote a cultural split from the previous NLPI community in the term's *narrow cultural definition*.

3.2.1 The historical context of *mechanistic*

Though it may be surprising in the modern era of LLM hype, not long ago "machine learning" referred primarily to computer vision research. When Saphra (2021) analyzed the proceedings of ICML 2020, they found that over three times as many papers referenced CVPR as any *ACL conference, demonstrating that the language modality was relegated to an application while computer vision results were seen as core machine learning.

The presumed unmarked nature of image classification research shaped the landscape of interpretability research as well: In computer vision work at the time, the dominant interpretability method was gradient-based saliency, which highlighted the importance of specific pixels in an input image (Simonyan et al., 2014; Bach et al., 2015; Springenberg et al., 2015; Zintgraf et al., 2017; Ribeiro et al., 2016; Shrikumar et al., 2017). Meanwhile, NLP researchers (and other ML researchers experimenting on text) occasionally borrowed saliency methods from computer vision (Karpathy et al., 2016; Li et al., 2016; Arras et al., 2016; Lei et al., 2016; Alvarez-Melis and Jaakkola, 2017), but primarily sought to understand models through representational geometry, attention maps, probing, and causal or correlational neuron analysis—all methods employed by the MI community today.

When Chris Olah first described "mechanistic interpretability" in 2020, then, this was the cultural landscape of the ML field: Machine learning mostly meant image classification and interpretability mostly meant feature saliency. Olah has confirmed on multiple occasions (Olah, 2024a,b) that he coined the term to differentiate circuit analysis from saliency methods, which were subject to increasing skepticism at the time (Kindermans et al., 2016; Adebayo et al., 2018; Kindermans et al., 2019; Ghorbani et al., 2019; Heo et al., 2019; Slack et al., 2020; Zhang et al., 2020). The MI paradigm was crucial and novel within computer vision—but the community around it didn't stay in computer vision.

3.2.2 Two LM interpretability communities

As excitement grew around new breakthroughs in NLP and dialogue systems, particularly with the rise of powerful Transformer language models such as GPT-3+ (Brown et al., 2020), many researchers migrated domains. The Circuits thread itself changed focus from vision to language in

2021 (Elhage et al., 2021), with the subsequent discovery of induction heads (Olsson et al., 2022) moving beyond existing efforts to characterize individual predictable attention heads (Kovaleva et al., 2019) to instead interpret the interaction between pairs of such heads. These new discoveries excited the NLPI community, but—unlike in computer vision—MI's goals and methods represented a direct continuation of the existing field.

Instead of a difference in methodology, the MI community brought a distinct *culture* to LM analysis. They came from outside of NLP or even from outside of ML entirely, often drawn by arguments that LMs posed an existential risk which could be tempered by deeper understanding.[5] Until mid-2023, most MI research was shared on blogs or forums such as LessWrong and The AI Alignment Forum; on Discord [6] and Slack [7]; or at invite-only workshops (MIT, 2023). Findings were rarely published on arXiv or in academic venues—and some members of the alignment community even advocated against publication, arguing that it would advance dangerous AI capabilities (Hobbhahn and Chan, 2023), though others advocated for more engagement with academics (@typedfemale, 2023). While MI researchers may have taken NLPI researchers' absence in online forums as a sign that they were uninterested in MI, many NLPI researchers were unaware of the MI community growing outside traditional research and publication venues.

As the MI community expanded and largely switched focus to language models, technical distinctions became less important than these cultural differences. In his popular guide to the field, Nanda (2022, ref. "The Broader Interpretability Field") avoided a strict technical definition of mechanistic interpretability, instead stating it "*feels* distinct," differentiated by its "culture," "research taste," and epistemics. Attempts to differentiate mechanistic from non-mechanistic interpretability quickly became untenable, leading to incongruent ontologies. For example, Nanda (2022) categorized *activation patching*—which Nanda attributed to the ROME paper (Meng et al., 2022)[8]—as MI but ROME itself—which uses activation patching to perform model editing—as non-MI. The modern MI community has even abandoned the early definitional goal of distinguishing MI from saliency—gradient-based feature attribution has re-emerged as another tool in the MI toolbox (Nanda, 2023a; Kramár et al., 2024).

To whatever degree *mechanistic* originally reflected a formal notion of causal mechanisms (§2.2), few researchers retain such a strict definition today. Instead, the formation of a separate, parallel language model interpretability community has led the term to its *narrow cultural definition*.

3.2.3 The clash of communities

The MI community eventually began publishing in academic conferences (Nanda, 2023b; Nanda et al., 2023; Wang et al., 2023). However, new engagement with academia only served to highlight bifurcated norms in the field. Researchers in the NLPI community expressed frustration on social media with the MI community's unfamiliarity with LM interpretability work prior to Anthropic's 2021 Circuits thread. Belinkov (2023a) argued that one paper "fail[ed] to engage with a large body of work on these topics from the past ~5 years," including direct precedents and improved baselines. Saxon (2023) alluded to a "contingent of people studying LLMs [who] don't meaningfully engage with *ACL literature." Others publicly stated that specific work from the MI community was "not new" (Artzi, 2023) or "published in the past" (Ravfogel, 2023). Posts often highlighted a tendency to "reinvent" (Andreas, 2023) or "rediscover" (Davidson, 2024) existing tools.

And yet, despite these tensions, the energy and resources of the growing MI community could not be denied. Many NLPI researchers subsequently began to use the term *mechanistic interpretability* to signal their engagement with the MI conversation (Nanda, 2024a).

3.3 We are all mechanistic now

Who wouldn't want to work on mechanistic interpretability? Students need advisors.[9] Funders

[5]Since 2021, the ML Alignment & Theory Scholars program (MATS), supported by the Berkeley Existential Research Initiative, has become a key point of entry for new researchers entering the interpretability field from outside of machine learning.

[6]e.g., `https://mechinterp.com/reading-group`

[7]e.g., `https://opensourcemechanistic.slack.com`

[8]Although the technique was first applied to neural networks by Vig et al. (2020) and Geiger et al. (2020).

[9]Prof. Sasha Rush of Cornell Tech noted, "pre-PhD researchers. . . [are] most excited about. . . 'mechanistic interpretability"' (Rush, 2024).

need grant recipients.[10] There is free pizza.[11] Is it any surprise that the traditional NLPI community increasingly embraces the term? Because the MI community accounts for much of the current growth in interpretability research (Räuker et al., 2023), the term has ceased to distinguish two separate communities and has grown into its *broad cultural definition*, encompassing the work of all interpretability researchers.

Simply embracing the term, however, has not fully unified these communities. Although MI forum posts are often methodologically similar to papers at ACL, some differences persist.[12] The traditional NLPI community tends, for example, to be interested in using linguistics (Sarti et al., 2024; Mohebbi et al., 2023; Katinskaia and Yangarber, 2024) and automata theory (Weiss et al., 2018, 2021; Merrill et al., 2022, 2020, 2024) as analytic tools. These topics are niche—but growing—in MI.

The MI community has its own characteristic interests, such as training dynamics (Olsson et al., 2022; Liu et al., 2023; Nanda et al., 2023; Zhong et al., 2023)—though the NLPI community has also studied this topic (Chiang et al., 2020; Saphra and Lopez, 2019b; Murty et al., 2023; Chen et al., 2024; Merrill et al., 2023). MI still strongly builds on the circuit paradigm that operates at the level of module interactions (Lieberum et al., 2023; Marks et al., 2024; Merullo et al., 2024; Tigges et al., 2024; Hanna et al., 2024; Dunefsky et al., 2024)—a framework which also inspires NLPI researchers (Ferrando et al., 2024; Ferrando and Voita, 2024). Work from Anthropic often becomes an MI focus, such as when promising results using SAEs (Bricken et al., 2023) inspired a flurry of followup work (Templeton et al., 2024; Gao et al., 2024; Lieberum et al., 2024; Belrose, 2024; Rajamanoharan et al., 2024a,b; Karvonen et al., 2024; Braun et al., 2024; Kissane et al., 2024; Gorton, 2024; Makelov, 2024)—though sparse encoding is another longstanding interest of NLPI (Subramanian et al., 2018; Niculae et al., 2018; Panigrahi et al., 2019; Meister et al., 2021; Prouteau et al., 2022; De Cao et al., 2022; Guillot et al., 2023).

Fortunately, there are signs of increasing unity in scientific focus. Some academics connected to the MI community have promoted interest in tools from linguistics and cognitive science (Wang et al., 2023; Arora et al., 2024). Speaker lineups at MI meetings often include longstanding NLPI researchers (MIT, 2023; Bau et al., 2024; Barez et al., 2024). MI researchers have also begun to engage more deliberately with peer-reviewed general ML conferences (Nanda, 2023b), though this effort has not extended to the specialized NLPI tracks and venues that focus on similar objectives and methods.[13]

4 Conclusion

Whatever terminological confusion and ideological tension they have brought to the interpretability field, the MI community is also responsible for its newfound popularity. The interest, energy, and opportunities MI brings to the field cannot be understated, nor should they be taken for granted. NLPI and MI researchers alike are motivated by social responsibility, intellectual curiosity, and the possibility of improving our tools. However, many MI researchers are also members of the alignment community concerned about catastrophic AI risk, where the value of MI is questioned (Greenblatt et al., 2023; Kross, 2023; Segerie, 2023).

There may come a time when alignment community consensus turns away from MI. Though many current MI researchers may leave—and some generous resources could disappear—others are likely to continue pursuing our shared objectives. Our communities have too much in common: scientific curiosity and a belief that we should understand the tools we use. We will all continue striving for that objective as long as there are opaque models to understand. Why not, therefore, also aim to connect?

[10] Effective Altruist charities have distributed millions in MI research grants (Open Philanthropy; EA Grants; Future of Life Institute).

[11] Many elite institutions have student societies where existential risk and MI are discussed over meals (Washington Post, 2023).

[12] In addition to cultural differences around scientific practice, there are also differences in preferred venues. ACL and BlackBoxNLP have struggled to engage the MI community, who prefer ML venues and the creation of new workshops (Barez et al., 2024). Prof. Yonatan Belinkov of Technion, a BlackBoxNLP founder, posted a call for MI researchers to submit to ACL venues (Belinkov, 2023b) and BlackBoxNLP 2023 (Belinkov et al., 2023) attempted to bridge the gap by inviting MI researchers to participate in a panel, where this divide was discussed.

[13] A point conceded by Neel Nanda, a leading MI researcher (Belinkov et al., 2023). The ACL preprint policy was a discouraging factor, but this is fortunately no longer the case (ACL Executive Committee, 2024).

Acknowledgments

We thank the authors of all tweets cited in this paper for granting us permission to reprint their tweets. We'd also like to thank (in alphabetical order): Fazl Barez, Yonatan Belinkov, Yanai Elazar, Thomas Fel, Neel Nanda, Chris Olah, Yuval Pinter, Ashish Sabharwal, Oyvind Tafjord, and the anonymous reviewers for engaging in discussion with us and providing valuable feedback. Thanks to Lelia Glass, Peter Hase, and Aryaman Arora for providing references.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.

ACL Executive Committee. 2024. Acl policies for review and citation. Webpage.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations*.

Afra Alishahi, Grzegorz Chrupała, and Tal Linzen. 2019. Analyzing and interpreting neural networks for nlp: A report on the first blackboxnlp workshop. *Natural Language Engineering*, 25(4):543–557.

David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 412–421, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Andreas. 2023. "Excited to see that it's that time of the year when we reinvent probing again.". Tweet. Accessed: 2024-08-28.

Jacob Andreas. 2024. "I still don't totally understand the difference between "mechanistic" and "non-mechanistic" interpretability but it seems to be mainly a distinction of the authors' social network?". Tweet. Accessed: 2024-08-09.

Omer Antverg and Yonatan Belinkov. 2022. On the pitfalls of analyzing individual neurons in language models. In *International Conference on Learning Representations*.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. ArXiv:2406.11717.

Aryaman Arora, Dan Jurafsky, and Christopher Potts. 2024. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand. Association for Computational Linguistics.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. Explaining predictions of non-linear classifiers in NLP. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 1–7, Berlin, Germany. Association for Computational Linguistics.

Yoav Artzi. 2023. "Distributional semantics? Reminds me of the "florida" example in the @omerlevy_ and @yoavgo paper from 2014. Granted, contemporary LLMs probably do it much better, but the ability is likely not new". Tweet. Accessed: 2024-08-28.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):e0130140.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Fazl Barez, Mor Geva, Lawrence Chan, Atticus Geiger, Kayo Yin, Neel Nanda, and Max Tegmark. 2024. Mechanistic Interpretability Workshop at the 41st International Conference on Machine Learning (ICML).

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2019. Identifying and controlling important neurons in neural machine translation. In *International Conference on Learning Representations*.

David Bau, Max Tegmark, Koyena Pal, Kenneth Li, Eric Michaud, and Jannik Brinkmann. 2024. New England Mechanistic Interpretability (NEMI) Workshop Series.

Sander Beckers, Frederick Eberhardt, and Joseph Y. Halpern. 2020. Approximate causal abstractions. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 606–615. PMLR.

Sander Beckers and Joseph Y. Halpern. 2019. Abstracting causal models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):2678–2685.

Yonatan Belinkov. 2018. *On internal language representations in deep learning: An analysis of machine translation and speech recognition*. Ph.D. thesis, Massachusetts Institute of Technology.

Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219.

Yonatan Belinkov. 2023a. "Excited to see important work from @andyzou_jiaming, @DanHendrycks..., on interpreting & controlling language models at representation level, to improve fairness & safety of LMs. Unfortunately it fails to engage with a large body of work on these topics from the past 5 years.". Tweet. Accessed: 2024-08-28.

Yonatan Belinkov. 2023b. "We are interested! #blackboxNLP has been the largest #nlproc workshop for several years now. And we have an interpretability track in all main #nlproc confs! Please submit your work to be reviewed in such venues...Even if you disagree with other approaches to interpretability, I think engagement through common conferences would help the community grow.". Tweet. Accessed: 2024-08-16.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017a. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Yonatan Belinkov, Sophie Hao, Jaap Jumelet, Najoung Kim, Arya McCarthy, and Hosein Mohebbi. 2023. Panel discussion on "mechanistic interpretability" at the 6th BlackboxNLP workshop at emnlp 2023. Video recording from 8:05.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017b. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Nora Belrose. 2024. "The @AiEleuther interpretability team is releasing a set of top-k sparse autoencoders for every layer of Llama 3 8B: `https://huggingface.co/EleutherAI/sae-llama-3-8b-32x`. We are working on an automated pipeline to explain the SAE features, and will start training SAEs for the 70B model shortly.". Tweet. Accessed: 2024-08-27.

Nathan Beniach. 2024. "is mechanic [sic] interpretability a sexier way of saying interpretability?". Tweet. Accessed: 2024-08-10.

Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*.

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

Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. 2021. An interpretability illusion for bert. ArXiv: 2104.07143.

Dan Braun, Jordan Taylor, Nicholas Goldowsky-Dill, and Lee Sharkey. 2024. Identifying functionally important features with end-to-end sparse dictionary learning. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread Blogpost.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, page 1877–1901. Curran Associates, Inc.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On identifiability in transformers. In *International Conference on Learning Representations*.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L. Leavitt, and Naomi Saphra. 2024. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *The Twelfth International Conference on Learning Representations*.

Cheng-Han Chiang, Sung-Feng Huang, and Hung yi Lee. 2020. Pretrained language model embryology: The birth of albert. *Preprint*, arXiv:2010.02480.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Róbert Csordás, Christopher Potts, Christopher D. Manning, and Atticus Geiger. 2024. Recurrent neural networks learn to store and generate sequences using non-linear representations. ArXiv: 2408.10920.

Hoagy Cunningham, Aidan Ewart, Logan Riggs Smith, Robert Huben, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. 2019. What Is One Grain of Sand in the Desert? Analyzing Individual Neurons in Deep NLP Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6309–6317. Number: 01.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. 2023. Reliability of CKA as a similarity measure in deep learning. In *The Eleventh International Conference on Learning Representations*.

Tim Davidson. 2024. "watching the mechanistic interpretability community rediscover manifolds with non-trivial topologies in real time is simultaneously amazing/exciting and concerning — highly recommend anyone in (mech-int) ML read @naturecomputes stunning work below to skip some steps!". Tweet. Accessed: 2024-08-28.

Nicola De Cao, Leon Schmid, Dieuwke Hupkes, and Ivan Titov. 2022. Sparse interventions in language models with differentiable masking. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 16–27. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Visualizing and understanding neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1150–1159, Vancouver, Canada. Association for Computational Linguistics.

Gus Docker and Neel Nanda. 2023. Neel Nanda on Avoiding an AI Catastrophe with Mechanistic Interpretability. Future of Life Institute Podcast.

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

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. 2024. Transcoders find interpretable llm feature circuits. ArXiv:2406.11944.

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)*, pages 4865–4880, Online. Association for Computational Linguistics.

EA Grants. Grants database: Long term futures fund.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Nelson Elhage, Tristan Hume, Catherine Olsson, Neel Nanda, Tom Henighan, Scott Johnston, Nicholas Joseph, Nova DasSarma, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022a. Softmax linear units. Transformer Circuits Thread Blogpost.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. 2022b. Toy models of superposition. ArXiv:2209.10652.

Nelson Elhage, Robert Lasenby, and Christopher Olah. 2023. Privileged bases in the transformer residual stream. Transformer Circuits Thread Blogpost.

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

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, Berlin, Germany. Association for Computational Linguistics.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. 2024. A primer on the inner workings of transformer-based language models. arXiv:2405.00208.

Javier Ferrando and Elena Voita. 2024. Information flow routes: Automatically interpreting language models at scale. ArXiv:2403.00824.

Future of Life Institute. PhD fellowships. Webpage. Accessed: 2024-08-26.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. ArXiv:2406.04093.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. 2024a. Causal abstraction: A theoretical foundation for mechanistic interpretability. ArXiv:2301.04709.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, page 9574–9586. Curran Associates, Inc.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 163–173, Online. Association for Computational Linguistics.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. 2024b. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 160–187. PMLR.

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

Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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

Liv Gorton. 2024. The missing curve detectors of inceptionv1: Applying sparse autoencoders to inceptionv1 early vision. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Ryan Greenblatt, Neel Nanda, Buck, and habryka. 2023. How useful is mechanistic interpretability? LessWrong Blogpost.

Simon Guillot, Thibault Prouteau, and Nicolas Dugue. 2023. Sparser is better: one step closer to word embedding interpretability. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 106–115. Association for Computational Linguistics.

Joseph Y Halpern and Judea Pearl. 2005a. Causes and explanations: A structural-model approach. part i: Causes. *The British journal for the philosophy of science*.

Joseph Y Halpern and Judea Pearl. 2005b. Causes and explanations: A structural-model approach. part ii: Explanations. *The British journal for the philosophy of science*.

Michael Hanna. 2024. What is mechanistic interpretability? You're not the only one asking! Transformer-specific Interpretability tutorial, part 3. Slides 2-6. Presented at the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL).

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Advances in Neural Information Processing Systems*, volume 36, pages 76033–76060. Curran Associates, Inc.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. 2024. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Juyeon Heo, Sunghwan Joo, and Taesup Moon. 2019. Fooling neural network interpretations via adversarial model manipulation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Bernease Herman. 2017. The promise and peril of human evaluation for model interpretability. In *Symposium on Interpretable Machine Learning at NeurIPS*, Long Beach, USA.

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

Evan Hernandez and Jacob Andreas. 2021. The low-dimensional linear geometry of contextualized word representations. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 82–93, Online. Association for Computational Linguistics.

Germund Hesslow. 1988. The problem of causal selection. *Contemporary science and natural explanation: Commonsense conceptions of causality*, pages 11–32.

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)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Marius Hobbhahn and Lawrence Chan. 2023. Should we publish mechanistic interpretability research? Blogpost on LessWrong and AI Alignment Forum. Accessed: 2024-08-19.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R Bowman. 2019. Do attention heads in bert track syntactic dependencies? ArXiv:1911.12246.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2017. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research (JAIR)*, 61:907–926.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

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*, pages 4198–4205, Online. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2021. Aligning faithful interpretations with their social attribution. *Transactions of the 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, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Sarthak Jain and Sarah Wiegreffe. 2023. Is "attention = explanation"? past, present, and future. Talk at "The Big Picture: Crafting a Research Narrative" workshop at EMNLP 2023.

David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. SemEval-2012 task 2: Measuring degrees of relational similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 356–364, Montreal, Canada. Association for Computational Linguistics.

Ákos Kádár, Grzegorz Chrupała, and Afra Alishahi. 2017. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4):761–780.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2016. Visualizing and understanding recurrent networks. In *Workshop at International Conference on Learning Representations (ICLR)*.

Adam Karvonen, Benjamin Wright, Can Rager, Rico Angell, Jannik Brinkmann, Logan Riggs Smith, Claudio Mayrink Verdun, David Bau, and Samuel Marks. 2024. Measuring progress in dictionary learning for language model interpretability with board game models. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Anisia Katinskaia and Roman Yangarber. 2024. Probing the category of verbal aspect in transformer language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3347–3366, Mexico City, Mexico. Association for Computational Linguistics.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2019. *The (Un)reliability of Saliency Methods*, pages 267–280. Springer International Publishing.

Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. 2016. Investigating the influence of noise and distractors on the interpretation of neural networks. In *NeurIPS 2016 Workshop on Interpretable Machine Learning in Complex Systems*.

Connor Kissane, Robert Krzyzanowski, Joseph Isaac Bloom, Arthur Conmy, and Neel Nanda. 2024. Interpreting attention layer outputs with sparse autoencoders. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

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)*, pages 4365–4374, Hong Kong, China. 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. ArXiv:2403.00745.

Nicholas / Heather Kross. 2023. Why and when interpretability work is dangerous. LessWrong Blogpost. Accessed: 2024-08-16.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 11–20, Minneapolis, Minnesota. Association for Computational Linguistics.

Carolin Lawrence. 2020. Interpretability and analysis of models for NLP @ ACL 2020. Medium Blogpost. Accessed: 2024-08-19.

Cristine H Legare and Tania Lombrozo. 2014. Selective effects of explanation on learning during early childhood. *Journal of experimental child psychology*, 126:198–212.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, page 107–117, Austin, Texas. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pages 41451–41530. Curran Associates, Inc.

Ruizhe Li. 2024. https://github.com/ruizheliuoa/awesome-interpretability-in-large-language-models. Github repository.

Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. 2023. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. ArXiv:2307.09458.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. ArXiv:2408.05147.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

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

Daniel Little. 2004. *Encyclopedia of Social Science Research Methods*, chapter Causal Mechanisms. Sage Publications. Edited by Michael Lewis-Beck, Alan Bryman, and Tim Futing Liao.

Ziming Liu, Eric J Michaud, and Max Tegmark. 2023. Omnigrok: Grokking beyond algorithmic data. In *The Eleventh International Conference on Learning Representations*.

Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470.

Tania Lombrozo. 2016. Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, 20(10):748–759.

Max M Louwerse and Rolf A Zwaan. 2009. Language encodes geographical information. *Cognitive Science*, 33(1):51–73.

Andreas Madsen, Himabindu Lakkaraju, Siva Reddy, and Sarath Chandar. 2024. Interpretability needs a new paradigm. ArXiv:2405.05386.

Aleksandar Makelov. 2024. Sparse autoencoders match supervised features for model steering on the IOI task. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. 2024. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *The Twelfth International Conference on Learning Representations*.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2024. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. ArXiv:2403.19647.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Models (COLM)*.

David Marr and Tomaso Poggio. 1976. From understanding computation to understanding neural circuitry.

Clara Meister, Stefan Lazov, Isabelle Augenstein, and Ryan Cotterell. 2021. Is sparse attention more interpretable? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 122–129. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, page 17359–17372.

William Merrill, Jackson Petty, and Ashish Sabharwal. 2024. The illusion of state in state-space models. In *Forty-first International Conference on Machine Learning*.

William Merrill, Ashish Sabharwal, and Noah A. Smith. 2022. Saturated transformers are constant-depth threshold circuits. *Transactions of the Association for Computational Linguistics*, 10:843–856.

William Merrill, Nikolaos Tsilivis, and Aman Shukla. 2023. A tale of two circuits: Grokking as competition of sparse and dense subnetworks. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

William Merrill, Gail Weiss, Yoav Goldberg, Roy Schwartz, Noah A. Smith, and Eran Yahav. 2020. A formal hierarchy of RNN architectures. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 443–459, Online. Association for Computational Linguistics.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2024. Circuit component reuse across tasks in transformer language models. In *The Twelfth International Conference on Learning Representations*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. arXiv:1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Beren Millidge. 2023. Deep learning models are secretly (almost) linear. Blogpost.

David Mimno and Laure Thompson. 2017. The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2873–2878. Association for Computational Linguistics.

MIT. 2023. Mechanistic interpretability conference.

Hosein Mohebbi, Grzegorz Chrupała, Willem Zuidema, and Afra Alishahi. 2023. Homophone disambiguation reveals patterns of context mixing in speech transformers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8249–8260, Singapore. Association for Computational Linguistics.

Jesse Mu and Jacob Andreas. 2020. Compositional explanations of neurons. In *Advances in Neural Information Processing Systems*, volume 33, page 17153–17163. Curran Associates, Inc.

Aaron Mueller. 2024. Missed causes and ambiguous effects: Counterfactuals pose challenges for interpreting neural networks. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, et al. 2024. The quest for the right mediator: A history, survey, and theoretical grounding of causal interpretability. ArXiv:2408.01416.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher Manning. 2023. Grokking of hierarchical structure in vanilla transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 439–448, Toronto, Canada. Association for Computational Linguistics.

Seil Na, Yo Joong Choe, Dong-Hyun Lee, and Gunhee Kim. 2019. Discovery of natural language concepts in individual units of cnns. In *International Conference on Learning Representations*.

Neel Nanda. 2022. A comprehensive mechanistic interpretability explainer & glossary. Blogpost. Accessed: 2024-08-09.

Neel Nanda. 2023a. Attribution patching: Activation patching at industrial scale. Blogpost. Accessed: 2024-08-14.

Neel Nanda. 2023b. "We're currently refining my grokking work to better fit academic interests and language, and are submitting it to a peer-reviewed AI venue. This was a bunch of effort, but I'm really hoping this can get more awareness and interest in mech interp from academic communities!". Tweet. Accessed: 2024-08-19.

Neel Nanda. 2024a. "A bunch of people who don't seem to identify as EA at all use the term to describe their work nowadays. I'm happy the field is growing outside of the EA bubble!". Tweet. Accessed: 2024-08-19.

Neel Nanda. 2024b. An extremely opinionated annotated list of my favourite mechanistic interpretability papers v2. Blogpost. Accessed: 2024-08-19.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*.

Vlad Niculae, Andre Martins, Mathieu Blondel, and Claire Cardie. 2018. SparseMAP: Differentiable sparse structured inference. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3799–3808. PMLR. ISSN: 2640-3498.

nostalgebraist. 2020. Interpreting GPT: the logit lens. LessWrong blogpost.

Chris Olah. 2015. Neural networks, types, and functional programming – colah's blog. Blogpost.

Chris Olah. 2024a. "I introduced the term to distinguish the work I was doing on circuits from a lot of other work that was going on circa 2018, notably saliency maps...I was motivated by many of my colleagues at Google Brain being deeply skeptical of things like saliency maps. When I started the OpenAI interpretability team, I used it to distinguish our goal: understand how the weights of a neural network map to algorithms...Since then, I think it's become an umbrella term for a variety of other things.". Tweet. Accessed: 2024-08-10.

Chris Olah. 2024b. "The motivating moment for me was that I went on a walk with a senior colleague shortly before I left Google Brain, and they matter of factly told me that "all interpretability is bullshit" because they'd been so turned off by saliency maps...I wanted a way to get people who were skeptical in this way to realize that I was talking about something pretty different than saliency maps and be willing to give it a second look.". Tweet. Accessed: 2024-08-10.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022. In-context learning and induction heads. Transformer Circuits Thread Blogpost.

Open Philanthropy. Potential risks from advanced artificial intelligence.

Abhishek Panigrahi, Harsha Vardhan Simhadri, and Chiranjib Bhattacharyya. 2019. Word2sense: Sparse interpretable word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5692–5705. Association for Computational Linguistics.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Forty-first International Conference on Machine Learning*.

Tiago Pimentel, Naomi Saphra, Adina Williams, and Ryan Cotterell. 2020. Pareto probing: Trading off accuracy for complexity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3138–3153, Online. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Niklas Stoehr, and Ryan Cotterell. 2022. Attentional probe: Estimating a module's functional potential. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11459–11472, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thibault Prouteau, Nicolas Dugué, Nathalie Camelin, and Sylvain Meignier. 2022. Are embedding spaces interpretable? results of an intrusion detection evaluation on a large french corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4414–4419. European Language Resources Association.

Giovanni Puccetti, Anna Rogers, Aleksandr Drozd, and Felice Dell'Orletta. 2022. Outlier dimensions that disrupt transformers are driven by frequency. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. ArXiv:1704.01444.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, Janos Kramar, Rohin Shah, and Neel Nanda. 2024a. Improving sparse decomposition of language model activations with gated sparse autoencoders. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. 2024b. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. ArXiv:2407.14435.

Shauli Ravfogel. 2023. "Many claims, modulo the autoencoder bit (which I strongly suspect is unnecessary for many of the findings), were published in the past. The post does not acknowledge the existence of much of the previous work. This is not atypical.". Tweet. Accessed: 2024-08-28.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

RepEval, editor. 2016. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. Association for Computational Linguistics, Berlin, Germany.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Mark O Riedl. 2019. Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, 1(1):33–36.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.

Sasha Rush. 2024. "I recently asked pre-PhD researchers what area they were most excited about, and overwhelmingly the answer was "mechanistic interpretability". Not sure how that happened, but I am interested how it came about.". Tweet. Accessed: 2024-08-08.

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)*, page 464–483, Raleigh, NC, USA. IEEE.

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

Naomi Saphra. 2021. Against monodomainism. Blogpost. Accessed: 2024-08-14.

Naomi Saphra and Adam Lopez. 2019a. Sparsity emerges naturally in neural language models. In *ICML Workshop Deep Phenomena*.

Naomi Saphra and Adam Lopez. 2019b. Understanding learning dynamics of language models with SVCCA. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3257–3267, Minneapolis, Minnesota. Association for Computational Linguistics.

Naomi Saphra and Adam Lopez. 2020. LSTMs compose—and Learn—Bottom-up. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2797–2809, Online. Association for Computational Linguistics.

Gabriele Sarti, Grzegorz Chrupała, Malvina Nissim, and Arianna Bisazza. 2024. Quantifying the plausibility of context reliance in neural machine translation. *Preprint*, arXiv:2310.01188.

Michael Saxon. 2023. "This is what happens when a significant contingent of people studying LLMs don't meaningfully engage with *ACL literature. This is why we need policies that don't drive out ECRs by disadvantaging their ability to preprint and publicize against ICLR/CVPR/NeurIPS-primary ECRs.". Tweet. Accessed: 2024-08-28.

Charbel-Raphaël Segerie. 2023. Against almost every theory of impact of interpretability. LessWrong Blogpost. Accessed: 2024-08-17.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural mt learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, page 1526–1534, Austin, Texas. Association for Computational Linguistics.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations (ICLR)*.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2015. Striving for simplicity: The all convolutional net. In *ICLR Workshops*.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581, Dublin, Ireland. Association for Computational Linguistics.

Anant Subramanian, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard Hovy. 2018. SPINE: SParse Interpretable Neural Embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Chenhao Tan. 2022. On the diversity and limits of human explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2173–2188, Seattle, United States. Association for Computational Linguistics.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Transformer Circuits Thread Blogpost.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 4593–4601, Florence, Italy. Association for Computational Linguistics.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. 2024. Llm circuit analyses are consistent across training and scale. ArXiv:2407.10827.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. 2023. Linear representations of sentiment in large language models. ArXiv:2310.15154.

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

P. D. Turney. 2012. Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research*, 44:533–585.

Peter D. Turney. 2005. Measuring semantic similarity by latent relational analysis. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1136–1141.

@typedfemale. 2023. "anthropic is making a mistake by not better grounding their interpretability research to existing topics like compressed sensing or frames. engaging smart academics who've spent years working on these areas in a different context is far more valuable than lesswrong posters". Tweet. Accessed: 2024-08-26.

Nadya Vasilyeva and Tania Lombrozo. 2015. Explanations and causal judgments are differentially sensitive to covariation and mechanism information. In *CogSci*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. In *The Eleventh International Conference on Learning Representations*.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615. Association for Computational Linguistics.

Washington Post. 2023. How elite schools like stanford became fixated on the AI apocalypse. By Nitasha Tiku. Section: Technology.

Gail Weiss, Yoav Goldberg, and Eran Yahav. 2018. On the practical computational power of finite precision RNNs for language recognition. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 740–745, Melbourne, Australia. Association for Computational Linguistics.

Gail Weiss, Yoav Goldberg, and Eran Yahav. 2021. Thinking like transformers. In *Proceedings of the 38th International Conference on Machine Learning*, pages 11080–11090. PMLR.

Jennifer C. White, Tiago Pimentel, Naomi Saphra, and Ryan Cotterell. 2021. A non-linear structural probe. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 132–138, Online. Association for Computational Linguistics.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*.

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)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. 2024. Answer, assemble, ace: Understanding how transformers answer multiple choice questions. ArXiv:2407.15018.

John Wieting and Douwe Kiela. 2019. No training required: Exploring random encoders for sentence classification. In *International Conference on Learning Representations*.

John Wu, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2020. Similarity Analysis of Contextual Word Representation Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4638–4655, Online. Association for Computational Linguistics.

Zhengxuan Wu, Atticus Geiger, Jing Huang, Aryaman Arora, Thomas Icard, Christopher Potts, and Noah D Goodman. 2024. A reply to makelov et al.(2023)'s" interpretability illusion" arguments. ArXiv:2401.12631.

Kaige Xie, Sarah Wiegreffe, and Mark Riedl. 2022. Calibrating trust of multi-hop question answering systems with decompositional probes. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2888–2902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361, Brussels, Belgium. Association for Computational Linguistics.

Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. 2020. Interpretable deep learning under fire. In *29th USENIX Security Symposium (USENIX Security)*.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. 2023. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Advances in Neural Information Processing Systems*, volume 36, pages 27223–27250. Curran Associates, Inc.

Zining Zhu, Hanjie Chen, Xi Ye, Qing Lyu, Chenhao Tan, Ana Marasovic, and Sarah Wiegreffe. 2024. Explanation in the era of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 19–25, Mexico City, Mexico. Association for Computational Linguistics.

Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. 2017. Visualizing deep neural network decisions: Prediction difference analysis. In *International Conference on Learning Representations*.

A Understanding language about understanding language models

The interpretability field struggled with terminological clarity and consensus long before *mechanistic* entered the lexicon (Doshi-Velez and Kim, 2017; Lipton, 2018; Rudin, 2019; Riedl, 2019; Jacovi and Goldberg, 2020). By even using the word *interpretability*, we implicitly dismiss the distinction drawn by Rudin (2019) between large "black-box" neural models and models which are designed to be understood: particularly, that the latter can be interpreted, but the former only explained.[14]

[14] As AI capabilities have advanced, this position against post-hoc black-box explanation has become less popular: Intrinsically interpretable models are often less performant (Madsen et al., 2024) and cannot always guarantee the human understanding that motivates their use (Lipton, 2018). As machine learning researchers have rejected the argument against black-box explanation (Jacovi and Goldberg, 2020), they have also abandoned any semantic distinction between explanation and interpretation.

While clarity is always important in scientific language, the nature of interpretability research makes it all the more urgent to speak precisely. As a community, we aim to understand the behavior of models and how they work, but how can we shed any light on these inner workings by leveraging confusing jargon? In fact, the ambiguity of *mechanistic* is emblematic of a wider struggle to communicate interpretability research effectively.

Let us consider some other sticking points in the interpretability lexicon. A core part of the "Is attention explanation?" debate (Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019; Bibal et al., 2022; Jain and Wiegreffe, 2023) is a disagreement over whether an *explanation* must be *faithful* by definition (Wiegreffe and Pinter, 2019, sec. 5). Subsequent work (Wiegreffe and Pinter, 2019; Jacovi and Goldberg, 2020) delineated between faithful and *plausible* (Herman, 2017), or human-acceptable (Wiegreffe et al., 2022), explanation. Even the terminology used to describe the format of textual explanations has been a source of discussion and disagreement (Jacovi and Goldberg, 2021; Wiegreffe and Marasović, 2021)—such as whether "extractive" and "abstractive," terms borrowed from the summarization literature, adequately characterize the difference between types of textual explanations.

In the MI literature, there have been terminology overloads or semantic disagreements over words like *feature* and *illusion*. The term *feature* has been used to describe mono-semantic concept representations of neurons derived from SAEs (Mueller et al., 2024), though it is more widely and historically associated with vector representations of data (text) that are either manually designed ("feature engineering") or learned by neural networks (Bereska and Gavves, 2024). A debate about subspace activation patching has centered around the meaning of the word *illusion*, namely, whether it applies to any dimension that becomes clearly causally relevant only when its causal role is tested with an intervention (Makelov et al., 2024), or whether such artifacts are a natural—and even explanatory—product of the model's representational geometry, and therefore informative of its true structure (Wu et al., 2024).

All of these examples, however, center around the need to ground our empirical work in precise vocabulary—not, like *mechanistic interpretability*, around the designation of group identity (§3.2.2). Terminological disagreements are usually resolved through discourse in shared venues. The NLPI community's adoption of the term *mechanistic* did not follow the same pattern (§3.3); its use may give the impression of cohesion and unity, but it masks a deep division which leads to duplicated research efforts and limits shared knowledge. Such outcomes will only hinder progress towards our shared goal: more deeply understanding language models.