Interpreting Language Models Through Concept Descriptions: A Survey

Nils Feldhus*[1,2] **Laura Kopf***[1,2]
[1]BIFOLD – Berlin Institute for the Foundations of Learning and Data
[2]Technische Universität Berlin
{feldhus,kopf}@tu-berlin.de
*Equal contribution

Abstract

Understanding the decision-making processes of neural networks is a central goal of mechanistic interpretability. In the context of Large Language Models (LLMs), this involves uncovering the underlying mechanisms and identifying the roles of individual model components such as neurons and attention heads, as well as model abstractions such as the learned sparse features extracted by Sparse Autoencoders (SAEs). A rapidly growing line of work tackles this challenge by using powerful generator models to produce open-vocabulary, natural language concept descriptions for these components. In this paper, we provide the first survey of the emerging field of concept descriptions for model components and abstractions. We chart the key methods for generating these descriptions, the evolving landscape of automated and human metrics for evaluating them, and the datasets that underpin this research. Our synthesis reveals a growing demand for more rigorous, causal evaluation. By outlining the state of the art and identifying key challenges, this survey provides a roadmap for future research toward making models more transparent.

1 Introduction

The interpretability of large generative models, and specifically LLMs, poses a central challenge to understanding their internal computations and enabling transparent decision-making. A key goal of this effort, often termed mechanistic interpretability, is to reverse-engineer the algorithms learned by these models through analysis of their fundamental components, such as individual neurons and attention heads (Saphra and Wiegreffe, 2024; Ferrando et al., 2024). A central problem in component analysis is assigning human-understandable meaning to these building blocks. Early approaches sought to map component activations to predefined linguistic properties through probing classifiers (Conneau et al., 2018; Belinkov, 2021) or to test their alignment with human-defined concepts (Kim et al., 2018; Lee et al., 2025). While foundational, these methods are limited by their reliance on a fixed set of concepts, potentially missing the novel representations learned by the model itself.

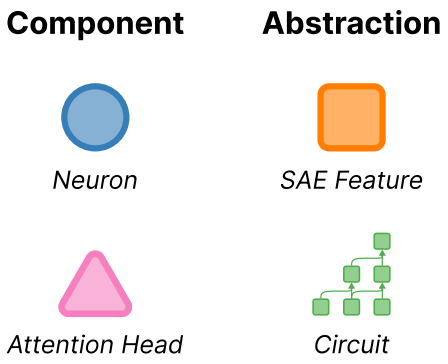

Figure 1: Illustration of model components (left column) and model abstractions (right column) of a language model. Components are individual units of a language model, such as neurons or attention heads. Abstractions go beyond the individual components of a model encompassing higher-level representations such as those learned by SAEs, or subgraphs involving multiple components or sparse features, as in circuits. Each component or abstraction can be associated with a human-understandable concept description.

A recent paradigm shift leverages the generative power of LLMs to overcome this limitation. Instead of testing for predefined concepts, this new line of work uses LLMs to produce open-vocabulary concept descriptions for the components of another model. This approach elicits a natural language explanation of a component's function by prompting the generator with data on when that component activates (Bills et al., 2023; Cunningham et al., 2024; Choi et al., 2024). For instance, given text fragments that maximally activate a specific neuron, an LLM synthesizes a description of the concept that neuron appears to detect, such as "legal clauses" or "references to the

Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 149–162

1980s". This method can be applied to both native model components and learned abstractions like Sparse Autoencoder (SAE) features (Figure 1).

This survey provides a structured overview of this rapidly emerging field. We focus on the methods, datasets, and evaluation techniques for generating concept descriptions for LLM internals, and we aim to answer the following questions:

(1) What are common methods and data sources for generating concept descriptions for interpreting language models? (§3)

(2) What are best practices for evaluating such concept descriptions? (§4)

(3) What are the common trends that can be observed in concept description research? (§5)

(4) What are the key gaps and promising future directions for concept-based descriptions of language models? (§6)

2 Definitions

2.1 Language Models

Neuron At the heart of modern LLMs lies the Transformer architecture (Vaswani et al., 2017), which stacks layers composed of two primary submodules: multi-head self-attention (MHA) and a position-wise feed-forward network (FFN). While MHA computes context-aware token representations, the FFN is responsible for further transforming these representations. An FFN layer typically consists of two linear transformations with a nonlinear activation function in between:

$$\boldsymbol{h}^i = \text{act}_{\text{fn}}(\tilde{\boldsymbol{h}}^i \boldsymbol{W}_1^i) \cdot \boldsymbol{W}_2^i, \quad (1)$$

where $\tilde{\boldsymbol{h}}^i$ is the output from the attention sub-layer, and $\boldsymbol{W}_1^i \in \mathbb{R}^{d \times d_{ff}}$ and $\boldsymbol{W}_2^i \in \mathbb{R}^{d_{ff} \times d}$ are weights matrices (d_{ff} is typically $4d$).

Within this framework, a neuron is defined as one of the intermediate dimensions in the FFN; specifically, it corresponds to a single column in the first weight matrix $\boldsymbol{W}_1^i$ and its subsequent nonlinear activation (Geva et al., 2022; Sajjad et al., 2022). A neuron is considered to have *fired* or *activated* for a given input if its activation value, $\text{act}_{\text{fn}}(\tilde{\boldsymbol{h}}^i \boldsymbol{W}_1^i)_j$, is positive.

However, the neuron as a unit of interpretation is fundamentally challenged by the phenomenon of polysemanticity. A single neuron often activates for a diverse and seemingly unrelated set of inputs, making its function difficult to summarize with a single, coherent description. This is a consequence of superposition, a strategy where models represent a vast number of concepts by encoding them as linear combinations of neuron activations (Elhage et al., 2022). While this is an efficient use of limited parameters, it means that individual neurons are inherently entangled and do not typically correspond to clean, understandable concepts.

Attention Head A level up from the individual neuron is the attention head. As a core component of the Transformer architecture, the multi-head attention mechanism allows a model to process information from different representation subspaces in parallel. Each head can be conceptualized as a specialist that learns to focus on different parts of the input sequence to capture specific relational patterns between tokens.

A significant body of research has demonstrated that these heads often acquire specialized and interpretable roles. For example, Voita et al. (2019) categorized attention heads in a machine translation model into distinct functional groups, including heads that attend to adjacent tokens (positional), heads that implement syntactic dependencies, and heads that focus on rare words. Subsequent work has identified heads responsible for more complex linguistic phenomena like subject-verb agreement, coreference resolution (Clark et al., 2019), and dependency parsing (Shen et al., 2022), as well as newer paradigms such as in-context retrieval augmentation (Kahardipraja et al., 2025).

2.2 Concepts

In the context of LLM interpretability, concepts constitute a human-understandable description of a pattern or property that a model has learned to represent. This notion aligns with the definition in Dalvi et al. (2022) who describe concepts as meaningful groups of words that can be clustered by a shared linguistic relationship. Such relationships can span a wide spectrum of abstraction, from low-level lexical features (e.g., *words starting with "anti-"*) and syntactic roles (e.g., *direct objects*) to high-level semantic categories (e.g., *names of capital cities*, *legal terminology*, or *financial terms*).

In terms of selecting ground-truth concepts, many works rely on pre-defined concepts, e.g., by using metaclasses in Wikipedia (Schwettmann et al., 2023; Dumas et al., 2025), or concept classification datasets (Antognini and Faltings, 2021; Abraham et al., 2022; Jourdan et al., 2023; Singh

et al., 2023; Sun et al., 2025). Others manually determine ground-truth concept labels with the help of human annotators (Dalvi et al., 2022; Mousi et al., 2023) or validate concept explanations post-hoc, e.g., in terms of usefulness for understanding the model's classification decision in a downstream task (Yu et al., 2024). Recent advances for topic modeling involving LLMs include SEAL (Rajani et al., 2022), a labeling tool for identifying challenging subsets in data and assigning human-understandable semantics to them. LLooM (Lam et al., 2024), which extracts interpretable, high-level concepts from unstructured text. It uses model embeddings with clustering methods and is shown to be better aligned with human judgment of semantic similarity and topic quality. Goal-driven explainable clustering (Wang et al., 2023b) assigns free-text explanations to each cluster with an LLM. Such approaches focus on providing a thematic summarization of textual inputs.

2.3 Natural Language Explanations

The automatic generation of human-readable text has been a long-standing goal in NLP. Natural Language Generation (NLG) systems often relied on templates and hand-crafted rule sets to convert structured data into prose (Gatt and Krahmer, 2018). This landscape shifted dramatically going via sequence-to-sequence architectures to LLMs and the prompting paradigm, also effecting how to generate explanations: Instead of simply predicting extractive rationales (Lei et al., 2016), a natural language explanation (NLE) explains model predictions with free text (Camburu et al., 2018; Wiegreffe et al., 2022). This gives them greater expressive power in terms of the reasoning they can convey, especially with complex reasoning tasks necessitating implicit knowledge.

NLEs generated by LLMs exceed other explanation methods in plausibility (Jacovi and Goldberg, 2020), but lack in faithfulness guarantees, as evidenced by a slew of recent studies (Turpin et al., 2023; Lanham et al., 2023; Chen et al., 2024; Madsen et al., 2024; Bentham et al., 2024; Parcalabescu and Frank, 2024; Bartsch et al., 2023). Furthermore, the resulting descriptions are highly sensitive to the specific prompt and the choice of the generator model, so the same prompt can yield different explanations. As noted by Bills et al. (2023), LLMs often produce overly broad or generic summaries and do not capture the nuance of the specific concept(s) represented by a model component. Figure 2 presents examples of such descriptions for different model components and abstractions.

2.4 Concept Datasets

The concepts a language model learns are fundamentally constrained by the data on which it was trained. Any dataset, from large-scale web scrapes to curated classification benchmarks, is an implicit repository of concepts. The frequency, context, and co-occurrence of words and phrases in the training corpus determine which patterns the model learns to represent internally. Consequently, interpretability methods that aim to describe internal components are, in effect, attempting to reverse-engineer and label these data-driven concepts. This process inherently creates an interpretability illusion: A model likely cannot represent a concept that is absent from its training data and the descriptions are therefore bounded by the conceptual scope of the underlying dataset (Bolukbasi et al., 2021).

Popular examples for datasets with explicit, human-labeled concepts include CEBaB (Abraham et al., 2022) and aspect-based sentiment datasets like BeerAdvocate (McAuley et al., 2012) and Hotel reviews (Wang et al., 2010). These resources are invaluable for controlled experiments where the goal is to see if a model's internal representations align with pre-defined, human-validated concepts. On the other hand, they heavily restrict the vocabulary in concept descriptions and certainly exclude the potential of detecting concepts that are not obvious to the human observer (Hewitt et al., 2025).

3 Description Methods

This section surveys the main targets of natural language concept description methods. These methods aim to make language models more transparent to humans in terms of understanding individual roles and underlying mechanisms. We organize these targets into native model components (§3.1) and learned model abstractions (§3.2). A comparative overview is provided in Table 1.

3.1 Model Components

Neurons A growing line of research focuses on fully automated descriptions of model components, such as neurons and their associated functions. In computer vision, early work explored neuron-level interpretability through visual concept alignment and feature visualization, generating text-based explanations and concept anno-

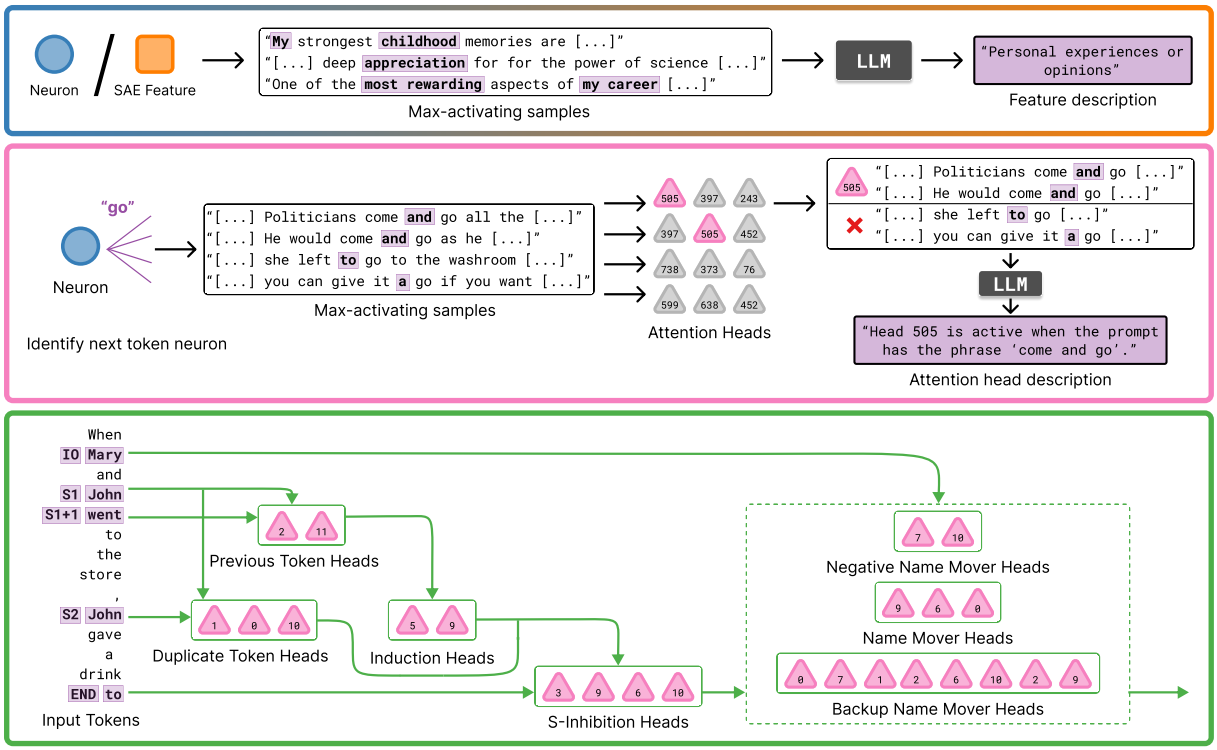

Figure 2: Overview of descriptions for model components (N neurons, A attention heads) and model abstractions (S SAE features, C circuits). The top panel shows a schematic example of an automatically generated feature description for a neuron or SAE feature, based on top-activating text samples (same process for both). The middle panel shows an example from Neo et al. (2024) of how attention head descriptions are generated: first a token-predicting neuron is identified, then prompts that highly activate it are found, the attention heads responsible for its activation are determined, and explanations for these heads are generated. The bottom panel shows a circuit from Wang et al. (2023a) implementing indirect object identification (IOI), where input tokens enter the residual stream and attention heads move information between streams, query/output arrows indicate where they write, key/value arrows where they read, and each class of head has an associated description.

tations to make individual neurons more human-interpretable (Bau et al., 2017; Mu and Andreas, 2020; Hernandez et al., 2022; Oikarinen and Weng, 2023; Bykov et al., 2023; Kopf et al., 2024). For language models, the seminal auto-interpretability approach by Bills et al. (2023) proposed labeling all neurons in GPT-2 XL. Their method uses GPT-4 to generate textual explanations from input samples that strongly activate each neuron. Since its introduction, this approach has been widely adopted and extended in follow-up work on automated neuron description (Choi et al., 2024; Gur-Arieh et al., 2025; Kopf et al., 2025).

A Attention Heads Another potential target component is the attention head. The MAPS framework, introduced by Elhelo and Geva (2025), provides a powerful example for inferring a head's functionality directly from its parameters, by projecting the head's weight matrices into the vocabulary space to create a token-to-token mapping matrix. This reveals the transformations the head has learned to perform, which lets us automatically map how strongly a given head implements a predefined relation, such as the knowledge-based *Country to capital* mapping or the linguistic *Word to antonym* operation. Neo et al. (2024), illustrated in Figure 2, investigate attention-MLP interactions. For example, they identify heads that perform "copying" from previous tokens in the same context, and use an LLM to generate and validate hypotheses about these functions.

3.2 Model Abstractions

S SAE Features Most of the neuron description methods adopt decomposition-based approaches that assign a single description to each neuron (Bills et al., 2023; Choi et al., 2024; Gur-Arieh et al., 2025), thereby limiting interpretability to the latent dimensions of the original model. However, this strategy struggles with neuron polysemanticity, so recent research has shifted toward

Study	Explained Model				Description Source	Target Dataset
	N Neurons	S SAEs	C Circuits	A Attention H.		
Bills et al. (2023)	GPT-2 XL				GPT-4	WebText
Cunningham et al. (2024)		Pythia-70M Pythia-410M			GPT-4	OpenWebText
Paulo et al. (2024)		Llama-3.1 8B, Gemma-2 9B			Claude Sonnet 3.5, Llama-3.1 70B	RedPajama-v2
Rajamanoharan et al. (2024)		Gemma-2 9B			Gemini Flash	*unspecified*
Choi et al. (2024)	Llama-3.1 8B				GPT-4o-mini	LMSYS-Chat-1M, FineWeb
He et al. (2024)		Llama-3.1 8B			GPT-4o	SlimPajama
Gur-Arieh et al. (2025)	Gemma-2 2B, Llama-3.1 8B, GPT-2 small,	Gemma Scope, Llama Scope, OpenAI SAE			GPT-4o-mini	The Pile
Kopf et al. (2025)	GPT-2 XL, Llama-3.1 8B	GPT-2 small, Gemma Scope			Gemini-1.5-Pro, GPT-4o-mini	C4
Chen et al. (2025)		Gemma-2 2B, Gemma-2 9B			GPT-4o-mini	PrivacyParaRel
Heap et al. (2025)		Pythia 70M – 7B			Llama-3.1 70B	RedPajama-v2
Muhamed et al. (2025)		Pythia 70M			Claude 3.5 Sonnet	arXivPhysics
Movva et al. (2025)		OpenAI SAE			GPT-4o	Wiki, Bills, Headlines, Yelp, Congress
Wang et al. (2023a)			GPT-2 small		*human*	IOI
Marks et al. (2025)			Pythia 70M SAE, Gemma Scope		*human*	Bias in Bios
Elhelo and Geva (2025)				Pythia 6.9B, GPT-2 XL	GPT-4o	*custom*
Neo et al. (2024)				GPT-2, Pythia 160M, Pythia 1.4B	GPT-4	The Pile

Table 1: Concept description techniques categorized by component/abstraction (N Neurons, S SAEs, C Circuits, A Attention Heads), description source, and target dataset.

learning more disentangled representations. One prominent direction involves sparse coding using SAEs (Bricken et al., 2023; Shu et al., 2025), which decompose model activations into higher-dimensional, sparsely activated feature spaces. These representations enable capturing a wider range of more interpretable, potentially monosemantic concepts (Bricken et al., 2023; Templeton et al., 2024; Gao et al., 2025). Auto-interpretability techniques initially developed for neurons (Bills et al., 2023) have since been successfully extended to these learned SAE features (Cunningham et al., 2024; Bricken et al., 2023; Gao et al., 2025; He et al., 2024; McGrath et al., 2024). However, SAEs have recently been subject to critique, with negative results reported by Smith et al. (2025) showing their underperformance relative to linear probes.

C Circuits A further step beyond studying individual components or SAE features is the analysis of circuits, which are computational subgraphs composed of multiple interacting components. The goal is to understand how the components interact to accomplish a specific task, and represent this interaction as a human-interpretable graph that captures the computation responsible for the task. Wang et al. (2023a) and Ameisen et al. (2025) use manual analysis and human labeling to trace and understand circuits, such as those responsible for indirect object identification. More recently, Marks et al. (2025) have pioneered methods to automatically discover and describe "sparse feature circuits", demonstrating how compositions of SAE features can implement complex logic, such as detecting gender bias. The well-studied IOI circuit has since been formalized as a benchmark task in the MIB suite (Mueller et al., 2025), providing a standardized way to evaluate circuit discovery methods. However, the fully automated generation of natural language descriptions for arbitrary, higher-order circuits remains an open challenge.

Evaluation Type	Measure	Study
Predictive Simulation	Simulator Correlation	Bills et al. (2023) Lee et al. (2023) Cunningham et al. (2024) He et al. (2024) Neo et al. (2024) Choi et al. (2024) Rajamanoharan et al. (2024) Chen et al. (2025) Muhamed et al. (2025) Movva et al. (2025) Poché et al. (2025)
	Detection / Fuzzing	Paulo et al. (2024)
Input-based Evaluation	AUROC	Heap et al. (2025) Kopf et al. (2025)
	Mean Activation Difference	Gur-Arieh et al. (2025) Kopf et al. (2025)
	Specificity	Templeton et al. (2024)
	Purity, Responsiveness	Puri et al. (2025)
	F1	Gao et al. (2025)
Output-based Evaluation	Intervention / Steering	Paulo et al. (2024) Gur-Arieh et al. (2025) Puri et al. (2025)
	Surprisal Score	Paulo et al. (2024)
Semantic Similarity	Cosine Similarity	Lee et al. (2023) Paulo et al. (2024) Heap et al. (2025) Kopf et al. (2025)
Human Evaluation	Correctness Correctness / Preference Readability	Bills et al. (2023) Lee et al. (2023) Li et al. (2024)
	Mono-/ Polysemanticity Rating	Rajamanoharan et al. (2024) Kopf et al. (2025)
	Plausibility Faithfulness	Elhelo and Geva (2025) Gur-Arieh et al. (2025)

Table 2: Concept description *evaluation* techniques categorized by metric, study, and the underlying quality being measured. Metrics are grouped into conceptual families: predictive simulation, input-based evaluation, output-based evaluation, semantic similarity, and human judgment.

4 Evaluating Concept Descriptions

Evaluating the quality of a concept description is a critical and non-trivial challenge. A good description should be accurate, faithful to the model's internal processing, and understandable to humans. In recent years, the field has developed a diverse toolkit of evaluation techniques, which we categorize into five main families: predictive simulation, input-based evaluation, output-based evaluation, semantic similarity, and human evaluation. Table 2 provides a comprehensive overview of these methods, the studies that use them, and the specific quality they aim to measure.

4.1 Predictive Simulation

Automated metrics are essential for evaluating descriptions at scale. The most widespread automated evaluation paradigm tests a description's predictive power: how well can it be used to simulate the feature's behavior? The canonical example is the simulator correlation method from Bills et al. (2023). Here, a "simulator" LLM (e.g., GPT-4) is given a feature's description and a text sample, and it must predict the feature's activation value for each token. The quality of the description is determined by the correlation between the simulated and the true activations. The simulator correlation scores in Bills et al. (2023) are generally very low (only 1,000 out of 307,200 neurons score at least 0.8, on a $[0.0, 1.0]$ scale). Adjusting the activations to a $[0, 10]$ scale creates a further loss of precision and nuance. While widely adopted (Table 2), this method has a key limitation: its reliance on another opaque LLM introduces a layer of confounding abstraction, making it difficult to know if a high score reflects a good description or simply the simulator's own pattern-matching prowess. Variations on this theme frame the task as classification instead of regression. For instance, the Detection and Fuzzing metrics (Paulo et al., 2024) ask a simulator to make a binary decision about whether a given

text would activate the feature described. Similarly, Automated Simulatability (Poché et al., 2025) tests if a description allows a simulator to predict the final output of the explained model, providing a more holistic measure of predictive utility.

4.2 Input-based Evaluation

A second family of metrics, building on the framework proposed by Huang et al. (2023), evaluates how accurately a description characterizes the inputs a feature activates on. The goal is to measure the description's purity (it only covers things the feature responds to) and coverage (it covers everything the feature responds to). Metrics like Specificity (Templeton et al., 2024) and Purity (Puri et al., 2025) quantify how selectively a feature fires for inputs matching its description versus random inputs. These approaches are often contrastive, testing a feature's response to on-concept examples versus "distractor" examples (McGrath et al., 2024). This property is often measured with standard classification metrics like AUROC, which assess how well a feature's activation score separates concept-positive from concept-negative examples (Heap et al., 2025; Kopf et al., 2025). Other metrics assess the mean activation difference between on-concept and off-concept inputs, regarding a description as better when activating on-concept examples have higher mean activations than off-concept examples (Gur-Arieh et al., 2025; Kopf et al., 2025).

4.3 Output-based Evaluation

The most rigorous metrics test a description's causal faithfulness: does it correctly predict the feature's effect on the model's output? These methods often rely on interventions or steering (Paulo et al., 2024; Gur-Arieh et al., 2025; Puri et al., 2025). For example, an evaluator might artificially activate a feature and check if the model's output distribution shifts in the way the description predicts (e.g., making a specific word more likely). The framework by Paulo et al. (2024) also includes causal metrics, such as direct Intervention Scoring and Surprisal Scoring. The latter measures whether providing the description reduces the model's loss on relevant inputs, offering a proxy for causal impact. These causal evaluations are crucial for verifying that a description is not merely correlational but reflects the feature's actual function.

4.4 Semantic Similarity

When ground-truth concepts are known or can easily be labeled, a straightforward automated evaluation is to measure the semantic similarity between the generated description and the ground-truth label. This is typically implemented by embedding both the generated text and the ground-truth concept name using a sentence embedding model (Muennighoff et al., 2023) and then calculating their cosine similarity. A high score indicates that the generated description is semantically aligned with the expected concept. This method is used by several studies as a sanity check or for evaluating performance on features with known correlates (Paulo et al., 2024; Heap et al., 2025). Semantic similarity can also be measured in the context of polysemanticity, when multiple descriptions per feature are available (Kopf et al., 2025). In this setting, the reference is the set of feature descriptions, which are descriptive annotations rather than ground-truth labels, and their similarity is measured by comparing them to one another.

4.5 Human Evaluation

Ultimately, concept descriptions are for humans, making human judgment essential for validating the meaningfulness of automated metrics. We identify several distinct roles for human evaluation:

- **Accuracy and Plausibility:** The most common use is asking humans to rate if a description is a correct and plausible summary of what a feature does, by showing them high-activating examples (Bills et al., 2023; Elhelo and Geva, 2025).
- **Readability and Clarity:** A description can be accurate but incomprehensible. Works like Li et al. (2024) focus specifically on evaluating the linguistic quality and clarity of the generated text.
- **Ground-Truth Annotation:** Instead of validating a generated description, humans can be used to create the ground truth itself, e.g., Kopf et al. (2025) employ human annotators to label polysemantic neurons with multiple concepts and apply similarity measures.
- **Faithfulness and Usefulness:** Humans can also validate causal claims. Gur-Arieh et al. (2025) ask humans to assess if a feature's effect on model outputs aligns with its descrip-

tion, providing a human-centric measure of faithfulness.

Despite its importance, human evaluation is less common in the mechanistic interpretability community compared to related NLP subfields (Geva et al., 2022; Simhi and Markovitch, 2023). We posit this is due to two factors: (1) the inherent difficulty, cost, and ambiguity in labeling what a polysemantic component "does", and (2) the community's traditional focus on expert-centric, debugging-oriented explanations over layperson-understandable ones (Saphra and Wiegreffe, 2024).

5 Findings

Our survey of the landscape of concept descriptions reveals several key trends. First, the recognition that neurons can be polysemantic (Elhage et al., 2022) has notable effects on methodological choices. It has driven the widespread adoption of model abstractions like SAEs (Bricken et al., 2023; Cunningham et al., 2024), which aim to provide more monosemantic interpretive units than the neurons themselves. This insight has also inspired the development of new frameworks that can capture multiple concepts per feature, such as Kopf et al. (2025).

Second, the evaluation of these descriptions is maturing, moving beyond simple correlation scores. The community is developing multi-faceted evaluation metrics to assess descriptions from different angles: their predictive power via simulation (§4.1), their accuracy at capturing activating inputs (§4.2), and their causal faithfulness to the model's behavior (§4.3). Frameworks that combine these perspectives are becoming the new standard (Gur-Arieh et al., 2025; Puri et al., 2025).

6 Recommendations for Future Work

From Components to Circuits The current focus on describing individual neurons or features is a necessary first step, but true understanding requires knowing how these parts compose. Future work should focus on scaling automated description methods to circuits. While circuit discovery is an active research area (Ameisen et al., 2025; Marks et al., 2025), the next frontier is to generate a natural language description for an entire computational subgraph, explaining how multiple features interact to implement a more complex function.

Scaling to New Data Domains and Modalities Most concept description methods have focused on models trained on general web text. A significant opportunity lies in applying these methods to specialized domains. Describing the features of models trained on legal text, medical data, or source code could yield domain-specific insights. Similarly, as models become increasingly multilingual and multimodal, future work should explore how concepts are represented across languages and whether unified descriptions can be found for features that respond to multiple modalities.

Analyzing the Interpreters Themselves As we increasingly rely on LLMs to generate explanations, we must critically analyze the "interpreters" themselves. What types of concepts are LLM-based describers biased towards? Do they tend to produce simple, atomic descriptions (e.g., *names of cities*) while missing more abstract or relational ones (e.g., *syntactic subject-verb agreement*)? Future work should include a meta-analysis of the generated descriptions to understand their linguistic properties, potential biases, and conceptual limitations, ensuring that our window into one model is not distorted by the lens of another.

A Finer-Grained View of Polysemanticity SAEs have become the default solution to polysemanticity, but they are not a silver bullet. Future research should explore more nuanced models of feature activation. For instance, rather than assuming a feature is simply "on" or "off", its activation level may matter; a feature might represent different concepts at different intensities or in different ranges (Haider et al., 2025). This calls for description methods that can capture this fine-grained detail, such as PRISM (Kopf et al., 2025), to provide a more complete picture of a feature's function.

Rigorous Causal Evaluation While evaluation methods are improving, most automated metrics remain correlational. The field must continue to push towards more rigorous tests of faithfulness. This includes scaling up intervention-based methods that test the causal effects of features on model outputs (Paulo et al., 2024) and developing "stress tests" that assess whether descriptions hold up under adversarial or out-of-distribution contexts. Critiques of the steerability of SAE features suggest that their causal impact is not yet fully understood, making this a critical area for future work (Wu et al., 2025).

Standardized Benchmarks for Evaluation Finally, a significant accelerator for progress in concept description methods would be the development of standardized benchmarks inspired by the Mechanistic Interpretability Benchmark of Mueller et al. (2025). Future work could build on this by creating benchmarks designed specifically to evaluate the quality of natural language descriptions. Such a benchmark could include a suite of components with agreed-upon "gold" descriptions, allowing for a more systematic comparison of different description and evaluation methods.

7 Conclusion

In conclusion, the use of LLMs to generate concept descriptions for model components and abstractions represents a notable step towards demystifying the internal workings of LLMs. Our survey has charted a clear trajectory in this emerging field: from early attempts to label individual, often polysemantic, neurons to higher-level abstractions like SAEs and circuits. Concurrently, the methods for evaluating these descriptions have matured from simple predictive correlations to a multi-faceted toolkit encompassing input-based purity, causal faithfulness, and nuanced human judgment.

For practitioners, this field offers a powerful new lens for model analysis, debugging, and auditing. However, the path from a generated description to a verifiable causal mechanism is not yet fully paved. The high computational costs and the ongoing debates around the faithfulness of these descriptions mean they should be applied with a critical eye. By continuing to refine the methods for generating these descriptions and, crucially, developing more rigorous and standardized benchmarks for their evaluation, the research community can forge these techniques into indispensable tools for building more robust, transparent, and trustworthy NLP systems.

Limitations

For scoping this survey, we limit our focus to natural language concept descriptions for internal model components and abstractions. We intentionally exclude other important families of interpretability work, such as purely mathematical analyses of model properties, feature visualization techniques that do not produce textual output, and methods focused on explaining final predictions rather than internal functions.

We also note that the computational and financial costs associated with the methods surveyed are substantial. Training high-quality SAEs requires plenty of GPU resources, and the subsequent steps of generating descriptions and performing automated evaluations often rely on expensive API calls to proprietary models. These costs currently pose a barrier to wider adoption and reproducibility, particularly in academic settings.

The vast majority of the research surveyed here focuses on English-language models and text corpora. The extent to which these methods and the concepts they uncover generalize to other languages remains a largely open and important question for future investigation.

Acknowledgements

We thank the anonymous reviewers at the BlackboxNLP Workshop for their feedback. We acknowledge support by the Federal Ministry of Research, Technology and Space (BMFTR) for BIFOLD (ref. 01IS18037A) and news-polygraph (ref. 03RU2U151C).

References

Eldar David Abraham, Karel D'Oosterlinck, Amir Feder, Yair Ori Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. 2022. CEBab: Estimating the causal effects of real-world concepts on NLP model behavior. In *Advances in Neural Information Processing Systems*.

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, and 8 others. 2025. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*.

Diego Antognini and Boi Faltings. 2021. Rationalization through concepts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 761–775, Online. Association for Computational Linguistics.

Henning Bartsch, Ole Jorgensen, Domenic Rosati, Jason Hoelscher-Obermaier, and Jacob Pfau. 2023. Self-consistency of large language models under ambiguity. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 89–105, Singapore. Association for Computational Linguistics.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network Dissection:

Quantifying Interpretability of Deep Visual Representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6541–6549.

Yonatan Belinkov. 2021. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, pages 1–13.

Oliver Bentham, Nathan Stringham, and Ana Marasovic. 2024. Chain-of-thought unfaithfulness as disguised accuracy. *Transactions on Machine Learning Research*.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. *OpenAI*.

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

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

Kirill Bykov, Laura Kopf, Shinichi Nakajima, Marius Kloft, and Marina Höhne. 2023. Labeling Neural Representations with Inverse Recognition. In *Advances in Neural Information Processing Systems*, volume 36, pages 24804–24828. Curran Associates, Inc.

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

Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen McKeown. 2024. Do models explain themselves? counterfactual simulatability of natural language explanations. In *Forty-first International Conference on Machine Learning*.

Yuheng Chen, Pengfei Cao, Kang Liu, and Jun Zhao. 2025. The knowledge microscope: Features as better analytical lenses than neurons. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10493–10515, Vienna, Austria. Association for Computational Linguistics.

Dami Choi, Vincent Huang, Kevin Meng, Daniel D. Johnson, Jacob Steinhardt, and Sarah Schwettmann. 2024. Scaling Automatic Neuron Description. *Transluce AI*.

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.

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

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, Abdul Rafae Khan, Firoj Alam, Nadir Durrani, Jia Xu, and Hassan Sajjad. 2022. Discovering latent concepts learned in bert. In *International Conference on Learning Representations*.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2025. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31822–31841, Vienna, Austria. Association for Computational Linguistics.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/toy_model/index.html.

Amit Elhelo and Mor Geva. 2025. Inferring functionality of attention heads from their parameters. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17701–17733, Vienna, Austria. 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*, abs/2405.00208.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2025. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core

tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

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.

Yoav Gur-Arieh, Roy Mayan, Chen Agassy, Atticus Geiger, and Mor Geva. 2025. Enhancing automated interpretability with output-centric feature descriptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5757–5778, Vienna, Austria. Association for Computational Linguistics.

Muhammad Umair Haider, Hammad Rizwan, Hassan Sajjad, Peizhong Ju, and A. B. Siddique. 2025. Neurons speak in ranges: Breaking free from discrete neuronal attribution. *arXiv*, abs/2502.06809.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. 2024. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *arXiv*, abs/2410.20526.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. 2025. Sparse autoencoders can interpret randomly initialized transformers. *arXiv*, abs/2501.17727.

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

John Hewitt, Robert Geirhos, and Been Kim. 2025. Position: We can't understand AI using our existing vocabulary. In *Forty-second International Conference on Machine Learning Position Paper Track*.

Jing Huang, Atticus Geiger, Karel D'Oosterlinck, Zhengxuan Wu, and Christopher Potts. 2023. Rigorously assessing natural language explanations of neurons. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 317–331, Singapore. Association for Computational Linguistics.

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.

Fanny Jourdan, Agustin Picard, Thomas Fel, Laurent Risser, Jean-Michel Loubes, and Nicholas Asher. 2023. COCKATIEL: COntinuous concept ranKed ATtribution with interpretable ELements for explaining neural net classifiers on NLP. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5120–5136, Toronto, Canada. Association for Computational Linguistics.

Patrick Kahardipraja, Reduan Achtibat, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. 2025. The atlas of in-context learning: How attention heads shape in-context retrieval augmentation. *arXiv*, abs/2505.15807.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2668–2677. PMLR.

Laura Kopf, Philine Lou Bommer, Anna Hedström, Sebastian Lapuschkin, Marina MC Höhne, and Kirill Bykov. 2024. Cosy: Evaluating textual explanations of neurons. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Laura Kopf, Nils Feldhus, Kirill Bykov, Philine Lou Bommer, Anna Hedström, Marina M. C. Höhne, and Oliver Eberle. 2025. Capturing polysemanticity with PRISM: A multi-concept feature description framework. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Michelle S. Lam, Janice Teoh, James A. Landay, Jeffrey Heer, and Michael S. Bernstein. 2024. Concept induction: Analyzing unstructured text with high-level concepts using lloom. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, and 1 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv*, abs/2307.13702.

Jae Hee Lee, Sergio Lanza, and Stefan Wermter. 2025. From neural activations to concepts: A survey on explaining concepts in neural networks. *Neurosymbolic Artificial Intelligence*, 1:NAI–240743.

Justin Lee, Tuomas Oikarinen, Arjun Chatha, Keng-Chi Chang, Yilan Chen, and Tsui-Wei Weng. 2023. The importance of prompt tuning for automated neuron explanations. In *NeurIPS Workshop on Attributing Model Behavior at Scale*.

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

Meng Li, Haoran Jin, Ruixuan Huang, Zhihao Xu, Defu Lian, Zijia Lin, Di Zhang, and Xiting Wang. 2024. Evaluating readability and faithfulness of concept-based explanations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 607–625, Miami, Florida, USA. Association for Computational Linguistics.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025. Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models. In *The Thirteenth International Conference on Learning Representations*.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining*, pages 1020–1025.

Thomas McGrath, Daniel Balsam, Liv Gorton, Murat Cubuktepe, Myra Deng, Nam Nguyen, Akshaj Jain, Thariq Shihipar, and Eric Ho. 2024. Mapping the latent space of llama 3.3 70b. *Goodfire Research*.

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

Rajiv Movva, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson. 2025. Sparse autoencoders for hypothesis generation. In *Forty-second International Conference on Machine Learning*.

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

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, and 4 others. 2025. MIB: A mechanistic interpretability benchmark. In *Forty-second International Conference on Machine Learning*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Aashiq Muhamed, Mona T. Diab, and Virginia Smith. 2025. Decoding dark matter: Specialized sparse autoencoders for interpreting rare concepts in foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1604–1635, Albuquerque, New Mexico. Association for Computational Linguistics.

Clement Neo, Shay B Cohen, and Fazl Barez. 2024. Interpreting context look-ups in transformers: Investigating attention-MLP interactions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16681–16697, Miami, Florida, USA. Association for Computational Linguistics.

Tuomas Oikarinen and Tsui-Wei Weng. 2023. CLIP-dissect: Automatic description of neuron representations in deep vision networks. In *The Eleventh International Conference on Learning Representations*.

Letitia Parcalabescu and Anette Frank. 2024. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6048–6089, Bangkok, Thailand. Association for Computational Linguistics.

Gonçalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. 2024. Automatically Interpreting Millions of Features in Large Language Models. *arXiv*.

Antonin Poché, Alon Jacovi, Agustin Martin Picard, Victor Boutin, and Fanny Jourdan. 2025. ConSim: Measuring concept-based explanations' effectiveness with automated simulatability. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5594–5615, Vienna, Austria. Association for Computational Linguistics.

Bruno Puri, Aakriti Jain, Elena Golimblevskaia, Patrick Kahardipraja, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. 2025. FADE: Why bad descriptions happen to good features. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17138–17160, Vienna, Austria. Association for Computational Linguistics.

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

Nazneen Rajani, Weixin Liang, Lingjiao Chen, Margaret Mitchell, and James Zou. 2022. SEAL: Interactive tool for systematic error analysis and labeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 359–370, Abu Dhabi, UAE. Association for Computational Linguistics.

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 and Sarah Wiegreffe. 2024. Mechanistic? In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 480–498, Miami, Florida, US. Association for Computational Linguistics.

Sarah Schwettmann, Tamar Rott Shaham, Joanna Materzynska, Neil Chowdhury, Shuang Li, Jacob Andreas, David Bau, and Antonio Torralba. 2023. FIND: A function description benchmark for evaluating interpretability methods. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yikang Shen, Shawn Tan, Alessandro Sordoni, Peng Li, Jie Zhou, and Aaron Courville. 2022. Unsupervised dependency graph network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4767–4784, Dublin, Ireland. Association for Computational Linguistics.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. 2025. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. *arXiv*, abs/2503.05613.

Adi Simhi and Shaul Markovitch. 2023. Interpreting embedding spaces by conceptualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1704–1719, Singapore. Association for Computational Linguistics.

Chandan Singh, Aliyah Hsu, Richard Antonello, Shailee Jain, Alexander Huth, Bin Yu, and Jianfeng Gao. 2023. Explaining black box text modules in natural language with language models. In *XAI in Action: Past, Present, and Future Applications*.

Lewis Smith, Senthooran Rajamanoharan, Arthur Conmy, Callum McDougall, Tom Lieberum, János Kramár, Rohin Shah, and Neel Nanda. 2025. Negative results for saes on downstream tasks and deprioritising sae research.

Yifan Sun, Danding Wang, Qiang Sheng, Juan Cao, and Jintao Li. 2025. Enhancing the comprehensibility of text explanations via unsupervised concept discovery. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14695–14713, Vienna, Austria. Association for Computational Linguistics.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, and 3 others. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. (Transformer) Attention Is All You Need. *arXiv:1706.03762 [cs]*.

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.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, page 783–792, New York, NY, USA. Association for Computing Machinery.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023a. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small. In *The Eleventh International Conference on Learning Representations*.

Zihan Wang, Jingbo Shang, and Ruiqi Zhong. 2023b. Goal-driven explainable clustering via language descriptions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10626–10649, Singapore. 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.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. 2025. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Forty-second International Conference on Machine Learning*.

Xuemin Yu, Fahim Dalvi, Nadir Durrani, Marzia Nouri, and Hassan Sajjad. 2024. Latent concept-based explanation of NLP models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12435–12459, Miami,

Florida, USA. Association for Computational Linguistics.