Rigorous Interpretation Is a Form of Evaluation

Isabelle Lee[*] USC
Emmy Liu CMU
Cathy Jiao CMU
Brihi Joshi USC
Dani Yogatama USC
Fazl Barez Oxford & WhiteBox
Michael Saxon UW[†]

Abstract

Current machine learning models are evaluated through behavioral snapshots, with benchmark accuracies, win rates and outcome-based metrics. Model explanations and evaluations, however, are fundamentally intertwined: understanding why a model produces a behavior can be as important as measuring what it produces. If we trusted interpretability, we argue that it can serve not merely as diagnostics but as a richer and more principled form of model evaluation beyond surface-level performance metrics. We explore three ways interpretability can function evaluatively: (1) fixing problems by identifying the root causes of unwanted behavior, (2) detecting subtly faulty mechanisms that invalidate model outputs, and (3) predicting potential issues before they arise by fully understanding the model's weaknesses. To fulfill its evaluative potential, we argue that interpretability methods must generate claims that are falsifiable, reproducible, and predictive—that is, interpretability must meet scientific standards.

1 Introduction

Current machine learning models are evaluated on their outputs. Typically, they are behaviorally assessed by performances on held-out data, win rates on benchmarks—some aggregate of these task metric scores. While these forms of evaluations capture immediate summary of behavior in a few numbers, they only capture surface competence. Two models may achieve identical behavior while relying on radically different internal mechanisms for example: one grounded in sound reasoning and computation and therefore structurally robust, the other brittle, heuristic driven.

This position paper argues that *interpretability* has the potential to expand current model assessment beyond surface-level behavioral evaluation to mechanism-level scrutiny. This is not to suggest that benchmarks are uninformative; rather, interpretability can deepen and extend them. However, for interpretability to serve this evaluative role, it must meet scientific standards. Interpretability claims must be *falsifiable, reproducible, and predictive.* Without these properties, interpretability remains descriptive rather than evaluative.

If interpretability satisfies these standards, it can transform evaluation from a retrospective summary of behavior into a dynamic, mechanism-sensitive process. By making internal structures accessible, interpretability enables: (1) causal diagnosis and repair of known failures, (2) detection of faulty reasoning even when outputs appear correct, and (3) anticipation of failures before they manifest behaviorally. More specifically:

1. **Falsifiability enables debugging.** In this scenario, we identified undesirable model behavior using well-defined evaluation criterion, and we want to use interpretability to debug and fix it. Falsifiability, defined as the quality of a claim being structured so it can in principle be proven wrong through targeted tests or interventions, is essential here. If we can make falsifiable, causal interpretability claims, we can reliably use model-based approach to correct unwanted associations at an architectural level, thereby addressing the root cause of the problem. Static benchmarks may reveal when models fail and sometimes characterize the nature of those failures, but they do not explain why the failure occurs or how to fix it. Interpretability may offer a pathway for more principled debugging.

2. **Reproducibility ensures detection of faulty mechanisms.** In this scenario, the evaluative criterion is unknown or poorly defined, so we use interpretability to identify the reasoning

[*]Correspond to lee.isabelle.g@gmail.com
[†]Now at Google Deepmind.

or associations behind a model decision. Interpretability here functions as a *mechanism-level evaluation*, because it can assess internal reasoning rather than just output, which can detect when a model produces a plausible output while relying on spurious correlation or heuristic shortcuts—that is, *if our interpretability methods met the standards of robustness, reliability, and reproducibility.* To serve this role as a scientific tool, interpretability methods must be reproducible and reliable, consistently identifying the same mechanisms across runs and remaining robust to perturbations; otherwise, apparent findings may reflect noise rather than genuine faulty reasoning.

3. **Predicting future failures.** In this scenario, we don't have a specific pre-set evaluative criterion or particular decision we wish to examine. Instead, we aim to predict potential problematic behavior or criteria that the model might struggle with in general. By applying the framework of *scientific prediction* as testing of our understanding, we generate interpretability claims that are predictive. By analyzing the model's internal behavior on a natural or in-distribution input set, interpretability techniques can help generate challenge sets or anticipate which types of inputs might cause issues. Practically, this can be seen as a way to stress-test the model before it is released to the public, allowing further refinement.

Taken together, interpretability can be reframed not merely as an exploratory add-on, but as a scientifically grounded and reliable explanation of behavior: an evaluation in its own right. Current evaluations are largely a retrospective summary of behaviors, but interpretability has the potential to transform evaluations into a forward-looking, mechanism sensitive discipline. When held to rigorous standards, interpretability can be evaluation.

2 Falsifiability enables debugging

One direct application of interpretability is to debug a model by identifying the internal origins of a model's problematic output and correcting them by direct modification or retraining. Unlike evaluation, interpretability can not only characterize the problematic behavior, but also may expose the internal failures underlying a particular behavior. However, for the interpretability claims to be testable and therefore reliably correctable, they must be falsifiable (Popper, 2005; Leavitt and Morcos, 2020)—that is, they can in principle be counterfactually *disprovable*.

Leavitt and Morcos critique a tendency in parts of the interpretability literature to prioritize intuition-building over the formulation of falsifiable hypotheses. In particular, they argue that attention heatmaps, component visualizations, and similar input-based explanations often function as descriptive aids rather than as testable claims about a model's internal mechanisms. Such approaches may help researchers build intuitions about how models behave, but they do not necessarily generate hypotheses that can be rigorously confirmed or disconfirmed. More broadly, they caution against explanatory claims that merely redescribe behavior without exposing themselves to potential refutation. In this respect, causal and mechanistic interpretability approaches appear to better meet Leavitt and Morcos's standard of hypotheses strength, as they more frequently articulate hypotheses that range from weakly to moderately falsifiable.

Current causal interpretations, however, face structural limitations that undermine this falsifiable debugging ideal. Many proposed features are **non-specific**: intervening on them affects multiple behaviors rather than isolating a single target function, leaving their representational role unclear. They are often **inconsistent**: the same intervention does not reliably produce the same behavioral shift across similar inputs. And they are subject to **underdetermination**: multiple distinct features can appear to influence the same behavior, with no principled way to identify which, if any, constitutes the true causal mechanism. When these conditions obtain, the debugging analogy collapses—the "feature" no longer behaves like a stable component but like a correlated direction in a high-dimensional space. A genuinely falsifiable debugging claim would require a stable causal unit, predictable intervention effects, and clear criteria for success or failure. Many current mechanistic interpretations fall short of these standards.

In this section, we first consider cases where interpretability methods do succeed in identifying meaningful, actionable features (§2.1). We then turn to cases where interpretability, particularly causal methods, breaks down (§2.2), and examine the specific modes of failure in those settings.

2.1 *Sometimes* interpretation can fix problems

Interpretability, at its best, helps identify the root cause of model misbehavior by illuminating how predictions are made. For example, explanations produced by *concept bottleneck* methods learn concepts specified by domain experts; these concepts are then used to constrain the model outputs, preventing errors at test time (Koh et al., 2020; Losch et al., 2019). Concept bottleneck interpreters are developed through *supervised* training on explicit concept labels provided by domain experts. These highlighted concepts outputted demonstrate what a model uses to arrive at a certain output. To fix the incorrect model output, we can directly intervene on weights of the model corresponding to found explanations. Concept bottleneck methods have shown promise in medical imaging; Koh et al. (2020) reported significantly improved performance in modeling knee osteoarthritis from X-rays after domain experts edited the model's predictions. However, this test-time debugging approach relies on experts' curated concepts, which are not easily transferable across tasks or, in some cases, even datasets (Hayashi and Sawada, 2023; Shin et al., 2023).

Another recent example is Sieve (Karvonen et al., 2024a), a code-specific suite of SAEs trained on public coding datasets. By promoting LM activation features related to regular expressions, Sieve improved performance on examples that required regular expressions. As in this scenario, interpretability at its best can enable direct debugging through interventions that have an explainable effect on model behavior.

2.2 Interpretation often can't fix problems

In mechanistic interpretability, an interpretability method is seen as faithful if it highlights features *causally* linked to an understood model behavior in output, as verified by interventions on model components or modifications to input data distributions (Saphra and Wiegreffe, 2024; Mueller et al., 2024). While these methods have shown promise, causal interpretations often fail to be falsifiable. They are underdetermined and inconsistent—often breaking down under distributional shifts, where the assumptions underpinning these causal relationships no longer hold. In this section, we examine one widely used mechanistic interpretation: causal interventions using features found with SAEs through steering. We then examine evidence of the brittleness of SAE feature interpretations. Such interventions cannot be reliably used in their current form for debugging.

Sparse Autoencoders Sparse autoencoders (SAEs) are a popular method used to interpret learned concepts in models. An SAE functions as a regular autoencoder with additional constraints that encourage the intermediate dimension to have sparse activations (Ng, 2011). In practice, an SAE takes in an activation vector from a language model of size N, and sparsely reconstructs the original input activation in the hidden layer of size $M > N$. Recently, SAEs have been employed to analyze the activations of large language models by treating the hidden layer as a dictionary of concepts after training (Cunningham et al., 2023; Gao et al., 2024).

Because SAEs allow us to examine a dictionary of concepts learned by a model, we may potentially use this dictionary to examine how learned concepts interact. For example, we could see if the features for gender and profession tend to co-occur in certain ways, when debugging and correcting for gender bias. However, for these associations to be useful in debugging models, they must remain stable across contexts and distributions.

Limitations of SAEs SAEs can suggest causal interventions through steering, or modifying a model's behavior by adjusting its activations in directions mapped to particular concepts. In practice, however, the effectiveness of steering can vary widely across different inputs, with the intervention failing or causing unintended consequences on specific examples (Tan et al., 2025). It may be possible to improve on naive steering, for instance by trying to minimize the effects on other latent features when finding a steering vector as in (Chalnev et al., 2024), but this does not fundamentally solve the problem of inconsistent feature interpretations in different contexts.

Although there is some evidence that SAEs trained on different models may share some common concepts and patterns of feature organization (Lan et al., 2024; Li et al., 2024), at a more granular level the concepts recovered from SAEs depend on what data the SAE was trained on (Paulo and Belrose, 2025). An SAE is trained on activations from a layer of the target model being interpreted. But what inputs is the target model generating its activations from? Model behavior and activations may vary under different data distributions, and therefore SAE concepts found in one domain may

not apply to others.

In one instance (Kissane et al., 2024), when training SAEs on the same model, a guardrail "refusal" feature (Arditi et al., 2024) identified by activations from chat-formatted data failed to generalize when the target model was provided with webtext input instead. For example, suppose we trained an SAE on model activations using input data in which profession and gender were strongly correlated. If our analysis revealed a feature that linked gender and profession, we could apply an intervention to mitigate this bias. However, interventions based on this feature might fail because the same feature might have different functions, and the appropriate feature set may even differ, when the model is handling contexts where the presumed correlation between gender and profession no longer holds.

3 Reproducibility ensures detecting of faulty mechanisms

We have now described how to analyze an output error in evaluation. What about cases where there is no observable error? Sometimes the model generates a seemingly plausible output, but relies on flawed internal mechanisms to produce that result. For example, we might need to address concerns of biased judgments on ambiguous inputs; determine whether a learned algorithm might fail on edge cases; or comply with regulations that ban incorporating protected characteristics for certain decisions. These concerns might not be apparent on individual outputs, prohibiting common metrics like accuracy, but they can affect behavior patterns across a distribution. We argue that *reproducible interpretability* can aid with this ambiguity in evaluation, where perhaps the evaluative objective has not been clarified or is ambiguous by nature.

An example: gender bias in machine translation Let's consider an example where a model which translates Spanish to English: *The doctor lost a stethoscope* as *El doctor perdió un estetoscopio*. The English sentence does not provide any hints as to the gender of the doctor, so the model produces a valid translation which assumes the doctor is male. Because gender marking is mandatory in Spanish, the model has to assume some gender on the part of the doctor, and is perhaps relying on the linguistic default of masculine. However, the gender may have been assumed on the basis of the doctor's profession from the training data distribution, provoking concerns of model bias. Did the model choose *el* based on the profession of *doctor*? Perhaps we can answer this question using interpretability tools.

We hope to use interpretability to assess the following bias claim: "this model chose a masculine token because it referred to a doctor." We would find that our interpretations are only useful for this claim if they are robust and reliable even when shifting away from the training data distribution.For simplicity, let's assume a toy training data where all male referents are doctors and all doctors are male. Say we base the explanation on a specific neuron causally related to the masculine output token *el*; this neuron's activation determines the output token to be masculine. To define the *explanation distance*, then, a pair of inputs have similar explanations if they feature similar activation for this particular neuron.

We seek to explain this neuron through a particular semantic relation between input (profession) and output (gender), such that the explanation distance is proportional to the semantic distance between two input and output pairs. To claim that the output gender depends on the referent being a doctor (and the model is therefore biased), we need this neuron to activate if and only if the referent is a doctor. Suppose that for every example similar to our toy training distribution, this condition holds—the activation is similar when the semantic relation between the profession and the referent is doctor-male. However, we may also find that the neuron also activates when the relation instead simply identifies the referent's gender as male (i.e. "*the doctor is the man.*" translates to "*El doctor es el hombre.*"). As a result, we can interpret the neuron as identifying a doctor referent or a male referent. The neuron's interpretation is underdetermined in the train set because the concepts "doctor" and "male" happen to be equivalent. If we choose the former interpretation, the model appears to be deciding gender based on profession—a potentially undesirable bias. However, if we choose the latter interpretation, the neuron is directly expressing the gender of the referent—so its role in selecting output gender does not imply bias.

In choosing a useful interpretation, we must aim for reproducibility: the interpretation should still hold for inputs out-of-distribution as well as for examples similar to the training distribution. If the robust, reliable, and therefore, reproducible interpretation defines the relation as referent profession rather than referent gender, we can use this neu-

ron to detect biased behavior on a single example. Without reproducible interpretations, bias can only be measured—if at all—by studying the model's behavioral patterns across a whole evaluation set.

In this situation, as opposed to Section 2, the model can produce plausible outputs using spurious correlations or shortcut learning. These heuristics are often undetectable in current evaluation practices. In §3.1, we will examine the case study of shortcut heuristics in Natural Language Inference (NLI). Then, in §3.2, we explore some counterfactual interpretability methods that can identify and validate underlying faulty reasoning patterns. Finally, in §3.3, we argue that reproducible and robust interpretability may ultimately yield faithfulness.

3.1 Sometimes evaluative objectives are ill-defined

In addition to invalid reasoning under specific regulatory requirements, some reasoning might be invalid because it fails to be reproducible under distribution shift. Previously, we discussed how to debug these errors after observation with falsifiable claims. However, these issues can be difficult to identify on in-distribution test sets; they are often discovered manually by chance or intuition. In §2 we discussed how falsifiable interpretability may resolve errors discovered through standard evaluation practices. Here, we discuss how reproducible, *counterfactual* interpretation may enable us to discover shortcuts and biases more elegantly, without blind experimentation.

One task in NLP where shifting evaluative goals and interpretability needs have intersected is natural language inference (NLI). NLI is the task of predicting if a pair of statements logically entail or contradict each other. For a language model to handle this entailment task, it must learn natural language semantics (Merrill et al.); in fact, the strength of next token predictors depends on their implicitly modeling entailment. Popular datasets for this task permitted models to rely on shortcut heuristics, allowing them to correctly predict the entailment of a sentence pair even when only one sentence was shown (Gururangan et al., 2018; Poliak et al., 2018). Because the task is defined in terms of both sentences, these heuristics would be brittle under some domain shifts.

An array of interventions were proposed to remove these shortcuts in the models directly and in the datasets themselves. Model-in-the-loop dataset modification processes were used to either select new hard samples (Nie et al., 2020) or filter shortcut-exhibiting samples (Saxon et al., 2023). At train time, residual learning (He et al., 2019), adversarial training (Stacey et al., 2020), and automated counterfactual data augmentation (Wang et al., 2021) all demonstrated that this ability can be unlearned or avoided during training altogether, even for models trained on biased datasets.

However, both this problem and its solutions were guided through experimentation and incremental changes to the evaluation objective by carefully selecting input examples. What if we could identify these reasoning flaws *without manually creating counterfactual test sets*? If functional interpretability methods for debugging (§2) and detecting subtle failures (§3) can be applied to a trained NLI model, they could identify shortcut reasoning without human discovery, measure its use during inference, and efficiently patch the bias to convert a flawed model into a correct one.

The fundamental purpose of evaluation is to answer questions about what a model can do. Test sets are a necessary step along a claim-supporting chain of evidence for this purpose. Poor within-task, cross-benchmark generalization is so well known that robustness to it is a means to benchmark base models (Yang et al., 2023). Models learning shortcuts and heuristics rather than generalizable mechanisms are an example of *internal validity dangers*—benchmark-internal breaks in the chain of evidence (Liao et al., 2021). These internal validity issues are often symptomatic of a problematic gap between the perception and truth of what an evaluation measures, or its *construct validity* (O'Leary-Kelly and Vokurka, 1998). In the era of benchmarks attempting to measure abstract, generalized capabilities this problem has been exacerbated (Raji et al., 2021). A desire to measure generalized, abstract human-like cognition in models is natural in a field which aims to replicate human-like intelligence, although this goal is flawed in many ways (Saxon et al., 2024). An evaluation's inability to map to a real-world task is a textbook *external validity danger* (Liao et al., 2021).

3.2 But counterfactual interpretation could help frame evaluative objectives

Proactively identifying model internal vs. strictly externally diagnosable issues in evaluation would be a valuable goal for interpretability research. Instead of relying on manual efforts to mitigate problematic test set issues in §3.1, the most promising

path forward lies in identifying shared mechanisms that, across tasks, degrade performance in complex ways that are understandable to humans. Interpretability work in the NLI bias explored which shortcut biases models learned during training and whether mitigation was necessary.While NLI models are brittle under tests requiring non-heuristic syntactic (McCoy et al., 2019) or lexical (Glockner et al., 2018) generalization, Srikanth and Rudinger (2022) showed that models trained on the standard two-sentence task—even when exposed to shortcut features—do not necessarily rely on the single-sentence heuristics identified by Poliak et al. (2018) at test time. This illustrates a broader point: evaluation alone often cannot settle debates about what models understand. Here, counterfactual interpretations that generalize across test distributions can step in where evaluations leave ambiguity. When evaluative objectives are underdefined, interpretability provides a way to probe whether a shared underlying mechanism governs performance across tasks.

When robust and trustworthy, counterfactual interpretability can help clarify whether performance across tasks reflects general capabilities or merely task-specific heuristics, especially when evaluation metrics alone are insufficiently precise. In this way, interpretability complements evaluation: evaluations elicit behaviors, and interpretability probes the causal mechanisms behind them. Thus, interpretability and evaluation are not separate tracks but complementary tools: evaluations act as diverse elicitation settings, while interpretability methods interrogate what is actually driving success.

3.3 Robust and reproducible explanations yield faithfulness in evaluation

Without robustness, multiple such counterfactual interpretations can exist for a particular evaluation of behavior, each potentially revealing different results. Here, we argue an interpretation that satisfies robustness should help uncover causally faithful explanations——interpretations that accurately reflect the model's reasoning process (Wiegreffe and Pinter, 2019; Jacovi and Goldberg, 2020; Lyu et al., 2022)——even when the evaluative objective underlying observed behaviors is ill-defined. While the stability of interpretations could significantly vary depending on what criterion we use to measure faithfulness, robustness has shown to be quite effective (Yin et al., 2021). Although different interpretability methods serve distinct purposes and may highlight different aspects of the model, we should be able to trust their explanations if they are faithful under *some* robustness guarantee. From these faithful, robust explanations, we should then be able to establish insights that are commonly agreed upon——some aspects of explanations should hold across different interpretations. If robust interpretations are generally faithful, their insights should reveal common themes in the behavior being measured, even when evaluative objectives are ill-defined. This, in turn, can help refine under-specified evaluative objective through robust explanations.

3.4 From identification to intervention

Falsifiability (§2) and reproducibility (§3) are minimal criteria for action, necessary but not sufficient. Our gender-MT example makes this concrete: both the profession-proxy and direct-gender readings satisfy reproducibility, yet only their differing OOD predictions can adjudicate which licenses an intervention. Predictivity (§4) is what closes this gap: a predictive interpretation specifies what should happen under counterfactual conditions, transforming a diagnosed mechanism into one we can act on.

4 Predicting future failures

Beyond the two previously discussed scenarios—debugging failures on known evaluative objectives (§2) and detecting subtle failures under ill-defined evaluative objectives (§3)—there are cases in which the evaluation objectives themselves are unknown. In §4, we explore a third scenario: potential model errors that go unnoticed because existing evaluations fail to capture the full range of data conditions and failure modes. We argue that if interpretability can produce predictive claims, we may use them to anticipate model failures from internal signals. Specifically, we may generate evaluations that are sufficiently challenging, tailored to the model. By systematically exploring and formalizing criteria for problematic behaviors or challenging inputs, interpretability facilitates a proactive strategy for uncovering model limitations.

A clarification is necessary about what we mean by "prediction." In machine learning, "prediction" refers to a model's output: given an input, the system produces a label, token, or score. This notion of prediction is in some sense purely behavioral and operational. By contrast, in the scientific sense, prediction plays a fundamentally different role. A scientific theory and hypothesis demonstrates its depth

by its testable expectations about what should occur under sufficiently specified assumptions and conditions. The success or failure of these predictions becomes evidence for—or against—the adequacy of our understanding. Under this framing, prediction becomes the strongest form of evaluation: rather than retrospectively summarizing performance, we generate mechanism-grounded hypotheses and design evaluations to validate our understanding.

Extending the gender-translation example Consider a hypothetical scenario in which no researcher has yet discovered gender bias in machine translation; for example, the possibility of conflating profession and gender has never been considered. Without exhaustively enumerating evaluation scenarios, we could instead examine model internal responses to the training data to anticipate likely failure cases. We may leverage interpretability to discover geometric entanglements between profession- and gender-related features, with which we generate evaluation sets that the models may struggle to generalize when associations from the training data no longer hold. Under sufficient robustness assumptions and conditions, we could leverage that very geometric properties of these relations to propose this form of bias as an evaluative objective even if we never encounter a female doctor in the existing data.

We propose that, if we framed interpretability claims in terms of scientific hypotheses to generate predictions about model behaviors, predictive interpretability can be leveraged to design evaluative objectives that specifically target areas where the model is likely to underperform. We begin by examining how geometric properties identifiable by interpretability can support predictive evaluation (§4.1). This motivates the need for a precise understanding of spurious correlations and their relationship to model internals (§4.2). Finally, we explore how to construct such evaluation sets by employing mechanisms for predicting out-of-distribution (OOD) behavior (§4.3).

4.1 We can use geometric properties of representations for *prediction*

Models learn geometric representations that reflect patterns in the training data, often aligning with how humans intuitively relate certain concepts. The classic example is the embedding vector relation "man + woman = king + queen," which captures gendered relationships in a way that mirrors human intuition (Vylomova et al., 2015; Liu and Neubig, 2022). Earlier work demonstrated that word embeddings can encode female and male gender stereotypes along a specific direction in the embedding space (Bolukbasi et al., 2016). This insight has already been used to address issues in model behavior: understanding the geometry of bias enables targeted interventions in embedding spaces for debiasing (Kaneko and Bollegala, 2021; Gonen and Goldberg, 2019).

Recently, the geometry of learned representations has been further elucidated, characterized in terms of the intrinsic dimensionality of multi-head attention embeddings and identified per-layer affine mappings of feedforward networks, facilitating a deeper understanding of model behavior in tasks like toxicity detection (Balestriero et al., 2023). These examples illustrate how uncovering the geometric structure of representations can not only reveal problematic model behaviors but also suggest it provides avenues for proactively defining and predicting them.

These geometric properties are not only observable but also mathematically definable and structurally persistent (Gardinazzi et al., 2024; Park et al., 2023). Another significant recent development is the linear representation hypothesis (Elhage et al., 2022; Park et al., 2023; Li et al., 2022), which posits that high-level concepts are encoded as linear directions in representation space. However, this hypothesis may break down for significantly out-of-domain data (lewis smith, 2024), where robustness failures undermine linear representation. Consider the most extreme scenario, in which we identify a feature that appears to represent gender, but in fact no longer corresponds to gender at all when the context relates to individual profession. In such a situation, even if we have interpreted features representing gender and profession, we cannot predict model behavior from their interaction because their interpretations do not hold in combination. If we aim to use these mathematical definitions of representations to define evaluative objectives and guide generations of challenge sets, we therefore want robustness guarantees.

Operationalizing this claim requires committing to a specific geometric quantity. Candidates include subspace alignment angles, intrinsic dimensionality, or distributional distances between concept subspaces; each making a different falsifiable prediction about OOD failure. We do not advocate a single choice, but observe that the geometric

argument is only as predictive as the quantity it commits to. Geometric analysis also inherits the underdetermination we diagnosed for SAEs in §2.2: rank collapse and manifold shift are consistent with the same failure pattern, so predictivity requires not just a geometric quantity but a discriminating one.

4.2 To use interpretation for prediction, we should understand spurious correlations

In embedding space, concepts exhibit geometric properties that allow certain representations to be combined or separated to form other intuitive concepts (Lappin and Bernardy, 2022; Wattenberg and Vi'egas, 2024; Lepori et al., 2023). This structure not only enables interpretability but also helps predict and mitigate biases or other problematic behaviors. However, even if geometric interpretations are stable, a key challenge remains: models often learn spurious correlations that may not be evident through geometric analysis alone.

Detecting these spurious correlations remains a significant challenge for current interpretability methods. SAEs have shown promise in low-data or corrupted-data settings—particularly when spurious features are simple—but often fail on OOD data, frequently returning null results (Kantamneni et al., 2024; Karvonen et al., 2024b). When effective, SAEs may still help uncover dataset defects or mislabeled examples (Kantamneni et al., 2024). Moreover, spurious correlations might pose a more fundamental challenge. Prior work suggests that spurious correlations may reflect deeper limitations of Empirical Risk Minimization (ERM), which can lead to "causally confused" models that overfit to patterns that generalize poorly—particularly on OOD inputs (Kirk and Krueger, 2023; Krueger et al., 2020). Since ERM minimizes average risk, models may exploit spurious features to reduce training loss—an issue not easily resolved through scaling or naive fine-tuning. Even increasing data diversity may be insufficient, as concept shifts driven by unobserved causal factors can persist in large datasets (Krueger et al., 2020).

4.3 Predicting out-of-domain behavior with in-domain interpretability

Full understanding of a model through interpretability may enable us predict how it behaves with OOD data (Juneja et al., 2022). Estimating performance in OOD scenarios, especially where labeled data is scarce, is critical for safe deployment. A deeper understanding of how models generalize across distributions allows us to anticipate their behavior on unfamiliar inputs. In this section, we explore analyses that can be extended to predict evaluative criteria and generate targeted OOD test cases.

Detecting generalization failures remains a significant challenge for current interpretability methods. At model internal level, predicting OOD behavior may be possible by leveraging various forms of model invariance. Prior work has shown that such invariances can improve generalization under distribution shifts, including domain changes (Gulrajani and Lopez-Paz, 2020), causal interventions (Arjovsky et al., 2019), data augmentation (Cubuk et al., 2019), and local interpolations (Luo et al., 2017). Models invariant to local transformations tend to factorize input space into a base space and a transformation set, effectively reducing input dimensionality and model complexity—thereby enhancing generalization. One example is neighborhood invariance, a complexity measure for learned representations. Ng et al. showed that neighborhood invariance remains robust even in OOD settings where other methods fail. Because it only requires selecting appropriate data transformations, this approach presents a promising direction for extending model performance beyond the training distribution.

5 Conclusion

	Falsifiability	Reproducibility	Predictability
SAEs	Maybe	✗	Maybe
CBMs	✔	Maybe	✗
Attention	✗	✗	✗
Probing	Maybe	✔	Maybe

Table 1: Interpretability against scientific standards.

In sum, behavioral benchmarks tell us what models do, but not how or why. Interpretability, when held to scientific standards of falsifiability, reproducibility, and predictability, offers a path toward mechanism-level evaluation that extends beyond traditional metrics. Table 1 makes the gap concrete: across four common method families, none currently meets all three criteria. By enabling causal debugging, reliable detection of faulty reasoning, and anticipation of future failures, interpretability can transform evaluation from a surface-level scorekeeping into a scientifically grounded explanation of internal process. If developed rigorously, interpretability can become evaluation in its own right.

Acknowledgements

IL is supported by Coefficient Giving's Technical AI Safety Research Grant. We are grateful to Naomi Saphra for insightful discussions and detailed comments.

References

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

Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *ArXiv*, abs/1907.02893.

Randall Balestriero, Romain Cosentino, and Sarath Shekkizhar. 2023. Characterizing large language model geometry solves toxicity detection and generation. *ArXiv*, abs/2312.01648.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Neural Information Processing Systems*.

Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. 2024. Improving steering vectors by targeting sparse autoencoder features. *Preprint*, arXiv:2411.02193.

Ekin Dogus Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. 2019. Randaugment: Practical automated data augmentation with a reduced search space. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3008–3017.

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

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Baker Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *ArXiv*, abs/2209.10652.

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. *Preprint*, arXiv:2406.04093.

Yuri Gardinazzi, Giada Panerai, Karthik Viswanathan, Alessio Ansuini, Alberto Cazzaniga, and Matteo Biagetti. 2024. Persistent topological features in large language models. *ArXiv*, abs/2410.11042.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *ArXiv*, abs/1903.03862.

Ishaan Gulrajani and David Lopez-Paz. 2020. In search of lost domain generalization. *ArXiv*, abs/2007.01434.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2018*, pages 107–112. Association for Computational Linguistics (ACL).

Naoki Hayashi and Yoshihide Sawada. 2023. Bayesian generalization error in linear neural networks with concept bottleneck structure and multitask formulation. *ArXiv*, abs/2303.09154.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Annual Meeting of the Association for Computational Linguistics*.

Jeevesh Juneja, Rachit Bansal, Kyunghyun Cho, João Sedoc, and Naomi Saphra. 2022. Linear connectivity reveals generalization strategies. *ArXiv*, abs/2205.12411.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. *ArXiv*, abs/2101.09523.

Subhash Kantamneni, Josh Engels, Senthooran Rajamanoharan, and Neel Nanda. 2024. Sae probing: What is it good for? absolutely something!

Adam Karvonen, Dhruv Pai, Mason Wang, and Ben Keigwin. 2024a. Sieve: Saes beat baselines on a real-world task (a code generation case study). *Tilde Research Blog*. Blog post.

Adam Karvonen, Can Rager, Samuel Marks, and Neel Nanda. 2024b. Evaluating sparse autoencoders on targeted concept erasure tasks. *ArXiv*, abs/2411.18895.

Robert Kirk and David Krueger. 2023. Causal confusion as an argument against the scaling hypothesis.

Connor Kissane, Robert Krzyzanowski, Neel Nanda, and Arthur Conmy. 2024. Saes are highly dataset dependent: A case study on the refusal direction. Alignment Forum.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348. PMLR.

David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Rémi Le Priol, and Aaron C. Courville. 2020. Out-of-distribution generalization via risk extrapolation (rex). *ArXiv*, abs/2003.00688.

Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. 2024. Sparse Autoencoders Reveal Universal Feature Spaces Across Large Language Models. *arXiv preprint*. ArXiv:2410.06981 [cs].

Shalom Lappin and Jean-Philippe Bernardy. 2022. A neural model for compositional word embeddings and sentence processing. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*.

Matthew L. Leavitt and Ari Morcos. 2020. Towards falsifiable interpretability research. *Preprint*, arXiv:2010.12016.

Michael A. Lepori, Thomas Serre, and Ellie Pavlick. 2023. Break it down: Evidence for structural compositionality in neural networks. *ArXiv*, abs/2301.10884.

lewis smith. 2024. The 'strong' feature hypothesis could be wrong.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Vi'egas, Hanspeter Pfister, and Martin Wattenberg. 2022. Emergent world representations: Exploring a sequence model trained on a synthetic task. *ArXiv*, abs/2210.13382.

Yuxiao Li, Eric J. Michaud, David D. Baek, Joshua Engels, Xiaoqing Sun, and Max Tegmark. 2024. The geometry of concepts: Sparse autoencoder feature structure.

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. 2021. Are we learning yet? a meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Emmy Liu and Graham Neubig. 2022. Are representations built from the ground up? an empirical examination of local composition in language models. *ArXiv*, abs/2210.03575.

Max Maria Losch, Mario Fritz, and Bernt Schiele. 2019. Interpretability beyond classification output: Semantic bottleneck networks. *ArXiv*, abs/1907.10882.

Yucen Luo, Jun Zhu, Mengxi Li, Yong Ren, and Bo Zhang. 2017. Smooth neighbors on teacher graphs for semi-supervised learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8896–8905.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2022. Towards faithful model explanation in nlp: A survey. *Computational Linguistics*, 50:657–723.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

William Merrill, Zhaofeng Wu, Norihito Naka, Yoon Kim, and Tal Linzen. Can you learn semantics through next-word prediction? the case of entailment. *Preprint*, arxiv:2402.13956 [cs].

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. 2024. The quest for the right mediator: A history, survey, and theoretical grounding of causal interpretability. *ArXiv*, abs/2408.01416.

Andrew Ng. 2011. Sparse autoencoder. `https://web.stanford.edu/class/cs294a/sparseAutoencoder_2011new.pdf`. CS294A Lecture Notes, Stanford University.

Nathan Ng, Neha Hulkund, Kyunghyun Cho, and Marzyeh Ghassemi. 2022. Predicting out-of-domain generalization with neighborhood invariance. *Trans. Mach. Learn. Res.*, 2023.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Scott W O'Leary-Kelly and Robert J Vokurka. 1998. The empirical assessment of construct validity. *Journal of operations management*, 16(4):387–405.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. *ArXiv*, abs/2311.03658.

Gonccalo Paulo and Nora Belrose. 2025. Sparse autoencoders trained on the same data learn different features. *ArXiv*, abs/2501.16615.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *NAACL HLT 2018*, page 180.

Karl Popper. 2005. *The Logic of Scientific Discovery*. Routledge, London.

Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. Ai and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Naomi Saphra and Sarah Wiegreffe. 2024. Mechanistic? In *The 7th BlackboxNLP Workshop*.

Michael Saxon, Xinyi Wang, Wenda Xu, and William Yang Wang. 2023. PECO: Examining single sentence label leakage in natural language inference datasets through progressive evaluation of cluster outliers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3061–3074, Dubrovnik, Croatia. Association for Computational Linguistics.

Michael Stephen Saxon, Ari Holtzman, Peter West, William Yang Wang, and Naomi Saphra. 2024. Benchmarks as microscopes: A call for model metrology. *ArXiv*, abs/2407.16711.

Sungbin Shin, Yohan Jo, Sungsoo Ahn, and Namhoon Lee. 2023. A closer look at the intervention procedure of concept bottleneck models. *ArXiv*, abs/2302.14260.

Neha Srikanth and Rachel Rudinger. 2022. Partial-input baselines show that NLI models can ignore context, but they don't. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4753–4763, Seattle, United States. Association for Computational Linguistics.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Sebastian Riedel, and Tim Rocktäschel. 2020. Avoiding the Hypothesis-Only Bias in Natural Language Inference via Ensemble Adversarial Training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8281–8291, Online. Association for Computational Linguistics.

Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adria Garriga-Alonso, and Robert Kirk. 2025. Analyzing the generalization and reliability of steering vectors. *Preprint*, arXiv:2407.12404.

Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin. 2015. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. *ArXiv*, abs/1509.01692.

Xinyi Wang, Wenhu Chen, Michael Saxon, and William Yang Wang. 2021. Counterfactual maximum likelihood estimation for training deep networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 25072–25085. Curran Associates, Inc.

Martin Wattenberg and Fernanda Vi'egas. 2024. Relational composition in neural networks: A survey and call to action. *ArXiv*, abs/2407.14662.

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.

Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie, and Yue Zhang. 2023. GLUE-X: Evaluating natural language understanding models from an out-of-distribution generalization perspective. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12731–12750, Toronto, Canada. Association for Computational Linguistics.

Fan Yin, Zhouxing Shi, Cho-Jui Hsieh, and Kai-Wei Chang. 2021. On the faithfulness measurements for model interpretations. *ArXiv*, abs/2104.08782.