From Attribution to Action: A Human-Centered Application of Activation Steering

Tobias Labarta
Fraunhofer Heinrich-Hertz-Institut

Maximilian Dreyer
Fraunhofer Heinrich-Hertz-Institut

Katharina Weitz
Fraunhofer Heinrich-Hertz-Institut

Wojciech Samek
Fraunhofer Heinrich-Hertz-Institut
Technische Universität Berlin
BIFOLD – Berlin Institute for the Foundations of Learning and Data

Sebastian Lapuschkin
Fraunhofer Heinrich-Hertz-Institut

Abstract

Explainable AI (XAI) methods reveal which features influence model predictions, yet provide limited means for practitioners to act on these explanations. Activation steering of components identified via XAI offers a path toward actionable explanations, although its practical utility remains understudied. We introduce an interactive workflow combining SAE-based attribution with activation steering for instance-level analysis of concept usage in vision models, implemented as a web-based tool. Based on this workflow, we conduct semi-structured expert interviews (N=8) with debugging tasks on CLIP to investigate how practitioners reason about, trust, and apply activation steering. We find that steering enables a shift from inspection to intervention-based hypothesis testing (8/8 participants), with most grounding trust in observed model responses rather than explanation plausibility alone (6/8). Participants adopted systematic debugging strategies dominated by component suppression (7/8) and highlighted risks including ripple effects and limited generalization of instance-level corrections. Overall, activation steering renders interpretability more actionable while raising important considerations for safe and effective use.

1. Introduction

Unlike traditional engineered systems, deep neural networks develop their internal structures through optimization over billions of parameters rather than through deliberate architectural specification [2, 19, 23]. The semantic knowledge encoded by individual neurons, channels, or attention heads remains largely unknown [12, 25], creating challenges in safety-critical applications where comprehending failure modes is essential [6, 42]. The explainable AI (XAI) research community has produced diverse methodologies to explain machine learning (ML) models; spanning local attribution techniques [3, 35, 46], counterfactual explanations [24], and concept-based approaches [1, 5, 31].

These methods enable what we term *correlational inspection*: practitioners can observe which components or features receive high attribution for a prediction, identifying candidates that may drive model behavior [26]. However, precise identification of relevant components is only half the picture. To verify whether highly attributed components *causally drive* a prediction rather than merely correlate with it, practitioners need interventive tools that act on the components explanations have surfaced [30, 36]. Such tools remain largely unavailable in practice [9, 36].

Advances in mechanistic interpretability (MI) offer a path toward closing this gap, from correlational inspection to what we term *actionable investigation*: directly ma-

nipulating model internals to test causal hypotheses about functional component roles. Sparse autoencoders (SAEs) can be applied to decompose model representations into interpretable components [15, 16], and attribution methods quantify each component's relevance to a given prediction. Activation steering complements this with an actionable capability by modifying these components to test their actual influence on model behavior [22, 44, 47]. In principle, combining attribution with steering enables such actionable explainability previous work have called for [30, 36]. However, existing work on steering has focused on dataset-level component selection [29], system usability [34], and perceptibility of steering effects [13] rather than its application for practitioners. Whether practitioners actually make the transition from correlational to causal reasoning when given steering access [8], and with what consequences for trust, strategy, and risk awareness, remains open.

In this work, we operationalize and evaluate the transition from attribution inspection to actionable investigation for instance-level model debugging. We structure this transition into a four-step workflow, from prediction review through attribution analysis and hypothesis formation to hypothesis testing via activation steering, and implement it in SemanticLens, a web-based tool for instance-level investigation of vision-language models. Using this workflow as experimental vehicle, we conducted semi-structured expert interviews with ML researchers and engineers (N = 8) with two debugging tasks on CLIP [45] to investigate whether the transition occurs in practice and what reasoning patterns, trust dynamics, investigation strategies, and perceived risks emerge.

2. Related Work

Mechanistic Foundation of Model Interventions The field of mechanistic interpretability seeks to reverse-engineer neural networks by decomposing them into understandable components and causal mechanisms [21, 59, 61]. A recent survey [59] organizes mechanistic MI methods into a "Locate, Steer, and Improve" pipeline, categorizing localization techniques (including gradient-based attribution) and intervention approaches (including different steering methods).

Model intervention can be applied for both understanding model representations and controlling model behavior. Early work on indiscriminate ablations [49, 51] applied random interventions as a tool to understand general network properties, while later targeted ablations [38, 39] were used for testing functional hypotheses. More recently, activation and relevance patching [28, 58], also known as causal tracing [37] and interchange intervention [20], emerged as a method to isolate which components mediate specific computations by replacing activations from counterfactual inputs [39].

Beyond model understanding, intervention has been applied to steer models towards desired behavior. Contrastive Activation Addition (CAA) enables steering of language model behavior via direction vectors [22, 44, 47]. Recent work demonstrates that SAE features in CLIP can be effectively steered to influence model outputs [29], with approximately 10-15% of features exhibiting meaningful steerability.

Bhalla et al. [8] observe that while causal intervention has been employed to assess explanation faithfulness [7, 10, 40, 43], these approaches rarely examine intervention as a method for practical control and debugging.

Interactive Visual Analytics for Model Validation and Error Analysis The goal of actionable model interventions has motivated development of several interactive visual analytics systems that combine explanation techniques with user-facing interfaces for systematic debugging. AttributionScanner [55] identifies data slices through attribution-based clustering without requiring metadata, while VISLIX [56] generates natural language explanations of error patterns using foundation models. SLIM [54] combines attention-weighted feature representations with human feedback to actively filter and rebalance datasets, and SUNY [53] analyzes necessity and sufficiency of learned features as causal explanations. These systems collectively establish that bridging correlational observation with actionable intervention through human feedback, data curation, or direct model steering is central to practical explainability.

Human-Centered Evaluation of Steering While the MI field has developed intervention techniques opening a new human-AI interaction space for model understanding and control, their human evaluation remains limited. To our knowledge, only two works [13, 34] have measured user-facing outcomes.

ConceptViz [34] is a visual analytics system for LLM concept exploration and validation with activation steering. The authors conduct a user study (N = 12) to evaluate the effectiveness of ConceptViz system components in supporting users to explore and understand features as well as the overall system usability and workflow. The evaluation focuses on interface and high-level usability aspects without investigating the impact of direct model intervention with steering on users' mental models and trust.

Diallo et al. [13] conduct the most comprehensive human evaluation of steering to date (N = 190), measuring whether users perceive emotion control in language model outputs. In a user study, participants rate perceived emotional intensity and text comprehensibility across six emotions and eight steering strengths. Results demonstrate that steering successfully amplifies target emotions, with sig-

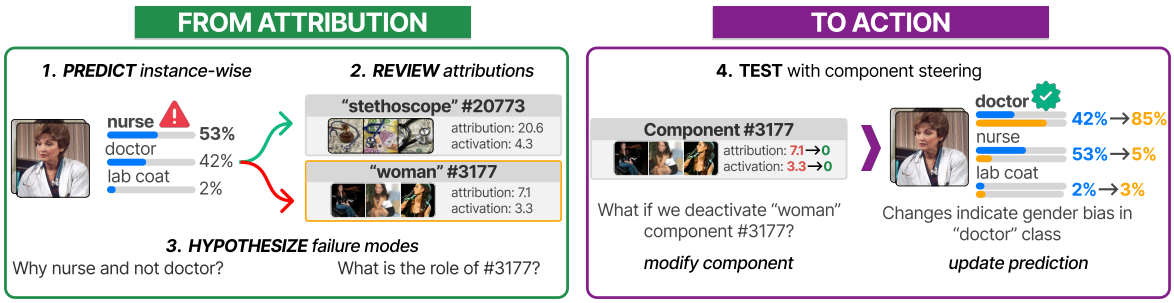

Figure 1. The four-step workflow from attribution to action: practitioners review component attributions, form causal hypotheses about components' roles, and test them via activation steering. The resulting prediction updates provide immediate feedback for hypothesis evaluation. Example shows suppressing the "woman" component to assess its influence on job role classification for "nurse".

nificant main effects of steering strength across five emotions. However, this evaluation measures perceptibility (i.e., can users detect the steering effect?) and identified a quality tradeoff (i.e., steering strengths beyond a threshold progressively degrade coherence), but does not assess whether emotion control improves task performance, trust calibration, or decision quality.

The related field of XAI offers established frameworks for human-centered evaluation that can inform steering assessment [14, 32, 33, 41]. Doshi-Velez & Kim [14] define functionally-grounded evaluation, where "real humans perform real tasks" to assess whether explanations improve decision quality. Building on this, Lage et al. [33] propose measuring trust calibration, cognitive workload, and behavioral patterns as core metrics for a human evaluation of interpretability. Empirical work demonstrates that explanations can both improve [4] and harm [60] decision quality depending on their fidelity, underscoring the need to measure human outcomes.

3. Investigating Model Predictions via Steering

This section operationalizes the transition from attribution to action as an interactive workflow for instance-level model investigation. We first describe the technical foundation that enables component attribution and steering, building on prior work on SAE-based decomposition and attribution [16], and then follow with the workflow.

3.1. Technical Foundation

Individual neurons in transformers are often polysemantic [18, 39]. Following Dreyer et al. [16], we train SAEs to extract monosemantic latent components from CLIP's representations. Each latent CLIP embedding $\mathbf{x}$ is decomposed as:

$$\mathbf{x} = \sum_{j=1}^{d_{\text{SAE}}} a_j(x)\mathbf{v}_j + \mathbf{b} + \boldsymbol{\epsilon}(\mathbf{x}),$$

where $a_j(x)$ denotes the activation of component j with feature direction $\mathbf{v}_j$, b is a bias term, and $\epsilon(x)$ represents reconstruction error. To assign human-interpretable descriptions, we compute semantic alignment scores against textual labels $t \in \mathcal{T}$ using the average visual embedding $\bar{\boldsymbol{x}}_j$ of the top-k most activating samples:

$$s_j(\boldsymbol{t}) = \frac{\bar{\boldsymbol{x}}_j \cdot \boldsymbol{t}}{\|\bar{\boldsymbol{x}}_j\|\|\boldsymbol{t}\|} - \frac{\bar{\boldsymbol{x}}_j \cdot \boldsymbol{t}_{\text{empty}}}{\|\bar{\boldsymbol{x}}_j\|\|\boldsymbol{t}_{\text{empty}}\|},$$

where $\boldsymbol{t}_{\text{empty}}$ is the embedding of an empty prompt. We quantify each component's influence on predictions through instance-wise attribution using Activation×Gradient (resembling Input×Gradient [50]):

$$R_j(\boldsymbol{x}, \boldsymbol{t}) = a_j \frac{\partial y(\boldsymbol{x}, \boldsymbol{t})}{\partial a_j},$$

for a given image-text (embedding) pair $(\boldsymbol{x}, \boldsymbol{t})$ and model output $y(\boldsymbol{x}, \boldsymbol{t})$, providing two key insights per component: its semantic meaning via $s_j(\boldsymbol{t})$ and its relevance to the current prediction via $R_j(\boldsymbol{x}, \boldsymbol{t})$.

Based on this information, users can formulate and test causal hypotheses about component roles through interactive steering. For a selected SAE component j with original activation a_j, a continuous control parameter $m_j \in [-1, 1]$ rescales the activation as:

$$a'_j = a_j(1 + m_j),$$

where $m_j = -1$ suppresses the component entirely ($a'_j = 0$), $m_j = 0$ leaves it unchanged, and $m_j = 1$ doubles its activation ($a'_j = 2a_j$). Unlike prior work that selects features based on dataset-level activation statistics [29], our approach uses instance-wise attribution scores $R_j(x, t)$ to identify components most relevant to specific predictions. Users select components for steering based on their attribution scores and semantic descriptions, forming a set $\mathcal{S} = \{j_k\}_{k=1}^{n}$ with corresponding steering values

$\{m_{j_k}\}_{k=1}^{n}$. Re-running inference with these modified activations reveals prediction changes.

Following the locate-and-steer paradigm described by Zhang et al. [59], our approach combines activation-weighted gradient attribution for localization and amplitude manipulation for steering. The continuous parameter m enables exploring dose-response relationships between component activation and prediction outcome, allowing users to test whether highly-attributed components causally drive instance-level predictions.

3.2. From Attribution to Action: Workflow

We implement the technical components described above as an interactive workflow in SemanticLens, a web-based tool for instance-level model investigation (Figure 2). The workflow guides users through four steps: (1) prediction review, (2) attribution analysis, (3) hypothesis formation, and (4) hypothesis testing through steering.

Step 1: Predict. The workflow begins with users selecting an input image for analysis. The system performs real-time inference, presenting the predicted class and confidence scores. Instance-wise attribution scores $R_j(x,t)$ are computed for all model components, establishing the foundation for exploration.

Step 2: Review. Components are ranked by their attribution and displayed in an interactive table (Figure 2). For each component j, users can inspect: (i) its attribution $R_j(x,t)$ indicating influence on the current prediction, (ii) its activation value a_j, (iii) its semantic description derived from alignment scores $s_j(t)$, and (iv) highly activating example images revealing what visual patterns trigger the component. This step enables correlational inspection: practitioners can identify which components are associated with a prediction and what they encode, but cannot yet act on this information to test whether these associations are causal.

Step 3: Hypothesize. By examining attribution rankings alongside semantic descriptions, users can formulate hypotheses about model behavior. For instance, when a component shows high attribution for a melanoma classification and its visualization indicates "textual markings", a user might hypothesize that the model is vulnerable to typographic attacks. However, attribution alone cannot confirm this hypothesis, since highly attributed components may correlate with but not cause the prediction. Transitioning from attribution to action requires a final verification step.

Step 4: Test. Users test their hypotheses through targeted interventions by adjusting activations via m_{j_k} for selected components $\mathcal{S} = \{j_k\}_{k=1}^{n}$. After defining modifications, users re-run the prediction and observe how outputs change, revealing whether the targeted components causally influence the prediction. This step closes the gap between inspection and action: correlational inspection of component attributions becomes actionable investigation of the causal roles of components.

Figure 2 illustrates this workflow for a typographic attack on the medical WhyLesionCLIP model [27, 57]: attribution analysis reveals a text-responsive component (#30496) as the top contributor to a misclassification induced by overlaying the word "regular". Setting $m_{30496} = -1$ suppresses this component and reverts the prediction to the correct class. The implementation is publicly available as part of the SemanticLens webapp[1].

The workflow structures the transition from attribution inspection to actionable investigation, but how practitioners apply it, and with what epistemic consequences, remains open.

4. Qualitative Evaluation: Expert Interviews

To investigate how practitioners understand and apply activation steering, we conducted semi-structured expert interviews. Our evaluation addresses four research questions:

- **RQ1** How does the transition from attribution inspection to action affect practitioners' reasoning about model failures?
- **RQ2** How does actionable investigation affect the epistemic basis, the grounds on which practitioners justify their trust, and the perceived utility of explanations?
- **RQ3** What investigation strategies emerge when practitioners are given steering capabilities?
- **RQ4** What risks and limitations of the approach do practitioners perceive?

We investigate these questions using the workflow and implementation described in Subsection 3.2 as the experimental setting. While the workflow provides the operational structure, our research questions and expert interviews target the epistemic effects of activation steering.

4.1. Interview Design

Each session lasted approximately 40 minutes and followed a think-aloud protocol [17]. The procedure is illustrated in Figure 3.

First, a pre-study questionnaire captured participants' professional roles, backgrounds, and prior experience with model debugging, bias detection, and explanation methods.

Second, participants completed two debugging tasks in SemanticLens (Figure 2). Participants received a brief introduction explaining that it was an image classification task with the true label shown alongside each case, and were asked to explore the available inspection cases. Task 1 targeted a typographic vulnerability on CLIP ViT-B-32 from Westerhoff et al. [52], where visible text overrides visual classification. Task 2 targeted a gender bias for job role classification in CLIP ViT-L-14, where swapping a male for

[1] https://semanticlens.hhi-research-insights.eu

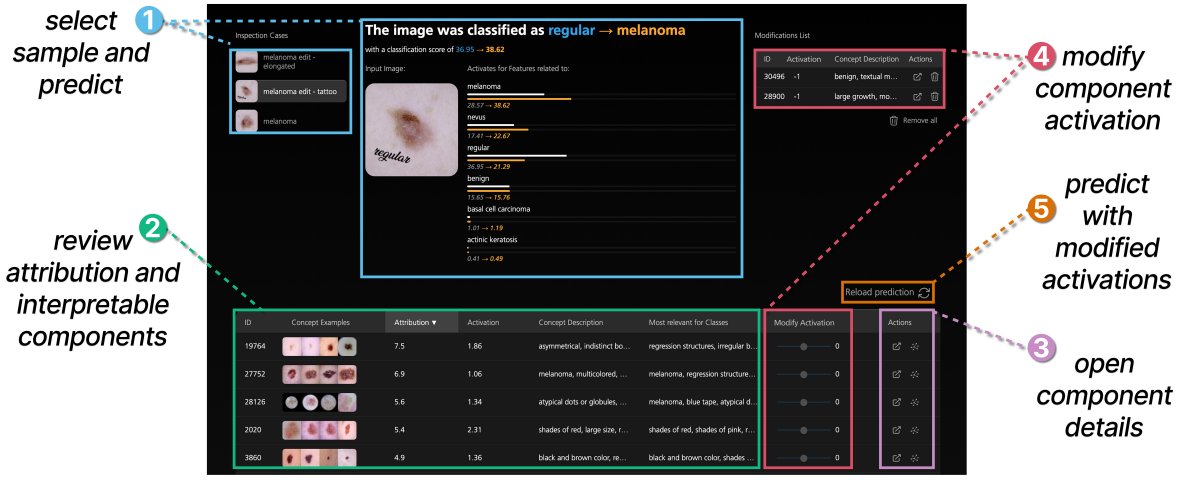

Figure 2. SemanticLens workflow implementation. Users select inspection samples (1), review components ranked by attribution (2), examine component details and visualizations (3), apply steering modifications (4), and observe prediction changes (5). Yellow text and bars indicate post-modification state. Example shown: a melanoma case corrected after suppressing component #30496, which responds to text artifacts.

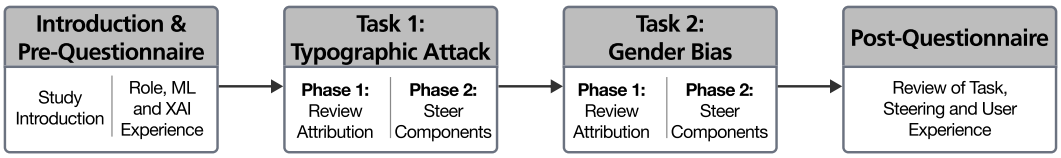

Figure 3. A process diagram of the interview structure: Pre-questionnaire, two debugging tasks each with attribution-only phase (Phase 1) and steering phase (Phase 2), and post-questionnaire.

a female person shifts predictions toward lower-status professions. Both tasks contained two inspection cases each (Figure 4).

Each task consisted of two phases. In Phase 1, participants explored attributions and explanations without access to steering. This phase continued until participants either identified incorrect predictions and began forming hypotheses about root causes, or ceased to make further progress. In Phase 2, steering controls were enabled, allowing participants to manipulate component activations and observe the effect on predictions. Semi-structured questions probed participants' understanding, confidence, trust, and strategy at the end of each phase.

This sequential design was chosen to observe within-subject shifts in reasoning: by first exposing participants to attribution alone, we establish a baseline against which the effect of adding actionability in the form of steering can be assessed. As a consequence, observed shifts in reasoning coincide with, but cannot be causally attributed solely to, the introduction of steering. Other factors such as task familiarity, increased interface comfort, and general learning may contribute to the patterns we report.

Finally, a post-study questionnaire collected participants' reflections on the tasks, steering, and overall user experience.

4.2. Study Participants

We recruited eight ML researchers and engineers (P1–P8) with diverse expertise spanning vision-language models, diffusion models, weather/climate modeling, time-series analysis, and emotion recognition. Experience with XAI methods ranged from extensive (SAEs, feature-attribution methods, probing classifiers) to none. One participant (P1) had prior experience with activation steering; the remaining seven encountered it for the first time during the study.

4.3. Method and Tools

Sessions were recorded and transcribed via Microsoft Teams and anonymized prior to analysis. Responses were coded from transcriptions and supplementary handwritten interviewer notes. Interviews were conducted in English or German depending on participant preference and native language; German responses were translated by the first author whose native language is German. The study design

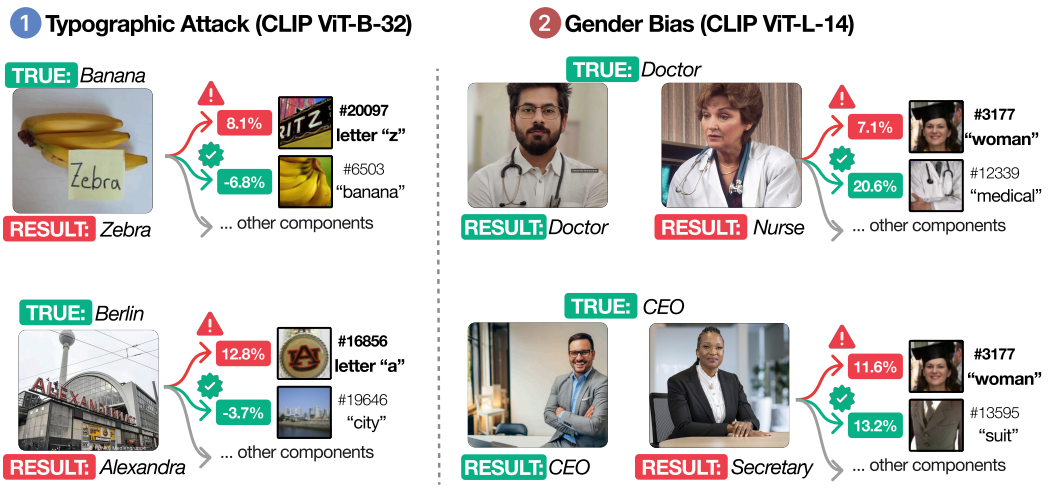

Figure 4. Two debugging tasks: Task 1 (typographic attack on CLIP ViT-B-32, where overlaid text causes misclassification [52]) and Task 2 (gender bias in CLIP ViT-L-14, where gender swap shifts predictions toward lower-status professions). Red badges indicate components contributing to misclassification; green badges indicate components supporting correct class. Component #3177 ("woman") appears among top-attributed features in both tasks.

was reviewed and approved by the Ethics Council of Fraunhofer Heinrich-Hertz-Institut. Given the exploratory nature, limited sample size, and novelty of human-centered evaluations of steering, our results focus on recurring themes and patterns rather than statistical analysis.

We employed thematic analysis with a deductive coding framework [11]. A deductive codebook of 22 codes organized under the four research questions was developed from the interview protocol and iteratively refined during initial coding. Codes capture reasoning patterns (correlational vs. causal), trust bases (plausibility vs. evidence vs. conditional), debugging strategies (suppression, amplification, exploration), and perceived risks. Initial candidate codes were generated with LLM assistance (see section 5) from the interview transcripts and the pre-defined research questions. The authors then reviewed all candidate codes against the transcripts, merging overlapping codes, discarding codes that lacked grounding in the data, and relabeling codes for conceptual precision. This refinement reduced the initial set and consolidated it into the final 22 codes reported in Table 1. All codes were subsequently applied to the transcripts by the authors.

Table 1. The 22 codes used to analyze interview data, organized by research question. *Count* indicates the number of participants for whom each code was identified.

RQ	Code	Count
RQ1	Correlational reasoning (pre-steering)	8
	Causal hypothesis formation	6
	Causal test completed	8
	Causal limitation acknowledged	5
RQ2	Plausibility-based trust	2
	Evidence-based trust (post-steering)	6
	Conditional trust	3
RQ3	Suppression strategy (necessity test)	7
	Amplification strategy (sufficiency test)	2
	Semantic scanning	7
	Exploratory (no prior hypothesis)	2
RQ4	Ripple effects / non-orthogonality	3
	Insufficient instance-level validation	3
	Over-steering / performance degradation	2
	Modification accumulation confounds	2

4.4. Task Outcomes

Table 2 summarizes behavioral outcomes across both tasks. Seven of eight participants identified the typographic attack and all eight identified the gender bias. Critically, all participants successfully corrected both failure modes through steering, including cases where participants had not identified the specific failure mode or responsible component beforehand.

Two observations deserve emphasis. First, P7 corrected the typographic attack through four iterative steering attempts without having identified the problem or the responsible component, demonstrating that steering allowed the participant to explore model behavior simply by steering model components and responding to steering outcomes. Second, participants needed fewer attempts to correct the

Table 2. Task outcomes across participants. *Recog.* = recognized the failure mode; *Comp.* = identified the responsible component; *Fix* = successfully corrected the prediction via steering; *Att.* = number of steering attempts to achieve fix. T1 = typographic attack task, T2 = gender bias task.

	T1: Typographic				**T2: Gender Bias**			
ID	Recog.	Comp.	Fix	Att.	Recog.	Comp.	Fix	Att.
P1	✓	✓	✓	1	✓	✓	✓	1
P2	✓	✘	✓	2	✓	✓	✓	1
P3	✓	✓	✓	2	✓	✓	✓	1
P4	✓	✓	✓	3	✓	✓	✓	1
P5	✓	✓	✓	1	✓	✓	✓	3
P6	✓	✓	✓	3	✓	✓	✓	1
P7	✘	✘	✓	4	✓	✓	✓	2
P8	✓	✓	✓	1	✓	✘	✓	2
Total	7/8	6/8	8/8	$M = 2.1$	8/8	7/8	8/8	$M = 1.5$

prediction via steering for T2 compared to T1 ($M = 1.5$ vs. $M = 2.1$). This likely reflects a learning effect, since T2 always followed T1, and may also be due to most participants perceived T2 as less difficult (see Table 2).

The 100% fix rate across both tasks constitutes a ceiling effect that limits the study's sensitivity to differences in tool effectiveness. Both failure modes were designed with known ground truth and involved perceptually salient cues (visible text overlay, gender swap), which eased identification and correction. These tasks served their intended purpose of enabling systematic observation of reasoning and strategy patterns during a short task, but the results should not be interpreted as evidence that the workflow would achieve similar success rates on complex tasks with subtler failure modes.

4.5. RQ1: Shift in Reasoning Patterns

During the attribution-only Phase 1, all participants engaged in correlational reasoning, trying to identify connections between components and predictions. For instance, P5 noted: *"One can already guess that it is probably because the text was recognized"*.

After receiving steering access, all eight participants completed at least one causal test cycle: formulating or iteratively discovering a hypothesis, performing an intervention, and evaluating the outcome. Six participants (P1, P3–P6, P8) articulated explicit causal hypotheses before intervening. P4 exemplified this pattern for Task 2: *"This concept should be less important, . . . then it should jump to CEO"*; an assumption they confirmed by steering accordingly. P8 used the most explicitly causal wording of all participants. For Task 1, they stated: *"Removing the text-related activation and then it works is a proof for the hypothesis"*. The remaining two participants (P2, P7) arrived at correct fixes through exploratory steering without articulating hypotheses beforehand, suggesting that steering can support discovery beyond hypothesis-driven investigation.

Notably, five participants (P2, P3, P5, P6, P8) acknowledged limitations of their causal claims without being prompted. P5 explicitly characterized the observed relationship as correlation rather than causation: *"The interaction between the slider and the output behavior; a correlation was definitely there"*. P6 questioned whether components are truly independent: *". . . whether the concepts are really orthogonal to each other or also influence other activations"*. This reflection suggests that steering enabled calibrated reasoning rather than inappropriately inflated causal confidence.

4.6. RQ2: Trust and Utility Calibration

We observed a change in the basis of trust reports after steering was introduced, though the sequential design precludes attributing this change solely to the steering capability. During the attribution-only phase, participants who expressed strong trust (2/8) grounded it in plausibility and coherence. P4 reported very high trust in the explanations, stating: *"It was very consistent with what I saw and with the prediction"*. After steering, the same participant grounded trust in testability: *"It is cool when you can directly change the influence and then check whether the model also changes the prediction, which is a kind of verification of my hypothesis"*.

Six of eight participants expressed evidence-based trust after steering, grounding their confidence in observed model responses to steering. Three participants additionally conditioned their trust on external validation. P2 stated that two tasks were insufficient and demanded full test-set metrics. P5 required robustness evaluation before considering safety-critical applications: *"Medical applications where patient lives are at stake . . . it would be too uncertain for me"*.

Regarding perceived utility of the explanation components, concept visualizations (example images showing what activates a component) were the most consistently valued element (7/8), while concept descriptions were frequently criticized as confusing (4/8). All participants assessed steering positively, emphasizing the speed and directness of feedback (P1, P4, P5), the live prediction update (P3, P6, P7), and the capacity to move beyond black-box perception (P8).

4.7. RQ3: Investigation Strategies

Before steering, participants used two complementary approaches to review components and build hypotheses. Most (7/8) used semantic scanning, inspecting concept visualizations and descriptions to identify suspicious components. Four of these participants additionally followed a top-down approach, starting from the highest-attributed component.

When steering, the dominant strategy was suppression of suspect components (7/8): participants tested whether removing a component's activation corrected the prediction. If suppressing a component restored correct classification, participants inferred that the component was responsible for the misclassification.

By contrast, only two participants tested through amplification of desired components. For example, P2 amplified banana- and city-related components to fix the predictions in the first task to "Banana" and "Berlin". P5 combined both approaches, first suppressing text features and then boosting banana features to fix the prediction to "Banana", while following the prediction outcomes and iteratively steering further: *"Zebra is now only at 11% down from 72% ... now we are at banana at 80%"*.

This asymmetry leads to design implications for SemanticLens: By ranking components by attribution to the (incorrect) prediction, the interface may implicitly guide users toward suppression strategies.

4.8. RQ4: Perceived Risks and Limitations

Seven of eight participants reported at least one risk when asked whether they could imagine potential concerns. Notably, most participants (7/8) had no prior knowledge of steering and could therefore not base their assessment on technical specifics. The identified risks cluster into two categories.

Technical risks concern the mechanics of steering itself. Three participants (P2, P6, P8) raised the concern that modifying one component may have unintended effects on others if components are not independent. P8 proposed a concrete design solution: *"If you make a change like this, if you could have some sort of global score ... being able to make a trade-off decision: is fixing this behavior locally worth the global effects?"*. Two participants (P1, P3) noted that excessive steering could potentially degrade overall model performance. These concerns align with established challenges in the MI literature around ripple effects [48].

Methodological risks concern the epistemological status of instance-level findings. Three participants (P1, P2, P5) emphasized that correcting individual predictions does not guarantee generalization. P1, the participant with prior steering experience, articulated the strongest standard: *"The steering direction must generalize across different datasets"*. P5 explicitly cautioned against deployment in safety-critical domains without robustness guarantees. One participant (P3) questioned whether steering would provide value beyond the rather obvious cases from our study, requesting a task where they "genuinely could not figure it out on their own". This highlights a limitation of the present evaluation: both tasks involved failure modes with perceptually salient cues (visible text overlay, explicit gender swap), meaning that the root cause could often be hypothesized from visual inspection of the input alone. Whether the workflow and the observed reasoning patterns transfer to subtler failure modes, such as texture bias, frequency shortcuts, or spurious correlations not visible in the input, remains an open question that future evaluations should address with tasks where the failure mechanism is not apparent a priori.

4.9. Additional Findings

Two further patterns emerged during the study with implications for future development of SemanticLens. Participants' mental models of the steering mechanism ($m \in [-1, 1]$) diverged substantially: only one participant (P8) understood the multiplicative scaling mechanism (where -1 removes and $+1$ doubles activation), while others conceptualized steering as weight adjustment (P3, P4, P6, P7), fine-tuning (P2), or contrastive modification (P5). This heterogeneity suggests explicit interface communication through tooltips or mechanistic animations would reduce ambiguity. Furthermore, participants identified potential applications beyond CLIP debugging in climate attribution, model reparameterization, and time-series analysis, though some questioned whether discrete, interpretable concepts would emerge in non-vision domains.

5. Conclusion

We operationalize and evaluate the transition from inspection to action for instance-level model debugging through semi-structured expert interviews (N = 8). All participants corrected both failure modes through steering, including cases resolved through exploratory steering without prior hypothesis formation. After steering was introduced, participants shifted from plausibility-based to evidence-based trust grounded in observed model responses, and most articulated explicit causal hypotheses while simultaneously acknowledging limitations of their causal claims. Strategy use was asymmetric, with suppression dominating over amplification, and participants surfaced concrete risks including ripple effects, over-steering, and insufficient instance-level validation. These findings provide first empirical evidence that activation steering, embedded in a structured workflow, can support the transition from correlational inspection to actionable investigation.

Several limitations warrant acknowledgment. The small expert sample (N = 8) and perceptually salient failure modes limit generalizability to subtler defects also supported by the high success rates. The sequential design introduces potential learning confounds; counterbalancing was not pursued given the exploratory scope. The gap between instance-level correction and global model improvement requires dataset-level validation. Future work should extend evaluation to larger samples, subtler failure modes, and develop improved communication of the steering mechanism.

Acknowledgements

This work was supported by the Federal Ministry of Research, Technology and Space (BMFTR) as grants [BIFOLD (01IS18025A, 01IS180371I), xJuRAG (16IS25015B)]; the European Union's Horizon Europe research and innovation programme (EU Horizon Europe) as grant ACHILLES (101189689); and the German Research Foundation (DFG) as research unit DeSBi [KI-FOR 5363] (459422098).

Declaration on the Use of Generative AI

During the preparation of this work, Claude Sonnet 4.5 and Grammarly were used for spelling and grammar checks, paraphrasing and rewording. Additionally, Claude Sonnet 4.5 was applied for generating initial qualitative codes during thematic analysis of the interviews. After using these tools, all authors reviewed and edited the content as needed and take full responsibility for the publication's content.

Ethical Statement

The Ethics Commission of the Fraunhofer Heinrich-Hertz-Institut provided guidelines for the study procedure and approved the study design. Informed consent has been obtained from all participants.

References

[1] Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 5(9):1006–1019, 2023. 1

[2] Shun-ichi Amari. Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4):185–196, 1993. 1

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10 (7):e0130140, 2015. 1

[4] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–16, 2021. 3

[5] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017. 1

[6] Vaishak Belle and Ioannis Papantonis. Principles and practice of explainable machine learning. *Frontiers in big Data*, 4:688969, 2021. 1

[7] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023. 2

[8] Usha Bhalla, Suraj Srinivas, Asma Ghandeharioun, and Himabindu Lakkaraju. Towards unifying interpretability and control: Evaluation via intervention. *arXiv preprint arXiv:2411.04430*, 2024. 2

[9] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019. 1

[10] Lawrence Chan, Adria Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: A method for rigorously testing interpretability hypotheses. In *AI Alignment Forum*, page 19, 2022. 2

[11] Victoria Clarke and Virginia Braun. Thematic analysis. *The journal of positive psychology*, 12(3):297–298, 2017. 6

[12] Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*, 2020. 1

[13] Diaoulé Diallo, Katharina Dworatzyk, Sophie Jentzsch, Peer Schütt, Sabine Theis, and Tobias Hecking. The effectiveness of style vectors for steering large language models: A human evaluation. *IEEE Access*, 13:191443–191457, 2025. 2

[14] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. 3

[15] Maximilian Dreyer, Jim Berend, Tobias Labarta, Johanna Vielhaben, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Mechanistic understanding and validation of large ai models with semanticlens. *Nature Machine Intelligence*, 7(9):1572–1585, 2025. 2

[16] Maximilian Dreyer, Lorenz Hufe, Jim Berend, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. From what to how: Attributing clip's latent components reveals unexpected semantic reliance. *arXiv preprint arXiv:2505.20229*, 2025. 2, 3

[17] David W Eccles and Güler Arsal. The think aloud method: what is it and how do i use it? *Qualitative Research in Sport, Exercise and Health*, 9(4):514–531, 2017. 4

[18] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022. 3

[19] M. Beatrice Fazi. Beyond human: Deep learning, explainability and representation. *Theory, Culture & Society*, 38(7): 55–77, 2021-12. 1

[20] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances*

in Neural Information Processing Systems, 34:9574–9586, 2021. 2

[21] Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, et al. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. 2

[22] Asma Ghandeharioun, Ann Yuan, Marius Guerard, Emily Reif, Michael Lepori, and Lucas Dixon. Who's asking? user personas and the mechanics of latent misalignment. *Advances in Neural Information Processing Systems*, 37: 125967–126003, 2024. 2

[23] Ian J Goodfellow, Oriol Vinyals, and Andrew M Saxe. Qualitatively characterizing neural network optimization problems. *arXiv preprint arXiv:1412.6544*, 2014. 1

[24] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38(5):2770–2824, 2024. 1

[25] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, 51(5):93:1–93:42, 2018. 1

[26] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 9(4):e1312, 2019. 1

[27] Lorenz Hufe, Constantin Venhoff, Erblina Purelku, Maximilian Dreyer, Sebastian Lapuschkin, and Wojciech Samek. Dyslexify: A mechanistic defense against typographic attacks in clip. In *The Fourteenth International Conference on Learning Representations*, 2025. 4

[28] Farnoush Rezaei Jafari, Oliver Eberle, Ashkan Khakzar, and Neel Nanda. Relp: Faithful and efficient circuit discovery in language models via relevance patching. *arXiv preprint arXiv:2508.21258*, 2025. 2

[29] Sonia Joseph, Praneet Suresh, Ethan Goldfarb, Lorenz Hufe, Yossi Gandelsman, Robert Graham, Danilo Bzdok, Wojciech Samek, and Blake Aaron Richards. Steering clip's vision transformer with sparse autoencoders. *arXiv preprint arXiv:2504.08729*, 2025. 2, 3

[30] Bernard Keenan and Kacper Sokol. Mind the gap! bridging explainable artificial intelligence and human understanding with luhmann's functional theory of communication. *arXiv preprint arXiv:2302.03460*, 2023. 1, 2

[31] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018. 1

[32] Jenia Kim, Henry Maathuis, and Danielle Sent. Human-centered evaluation of explainable ai applications: a systematic review. *Frontiers in Artificial Intelligence*, 7:1456486, 2024. 3

[33] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. Human evaluation of models built for interpretability. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, pages 59–67, 2019. 3

[34] Haoxuan Li, Zhen Wen, Qiqi Jiang, Chenxiao Li, Yuwei Wu, Yuchen Yang, Yiyao Wang, Xiuqi Huang, Minfeng Zhu, and Wei Chen. Conceptviz: A visual analytics approach for exploring concepts in large language models. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 2

[35] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017. 1

[36] Gennie Mansi, Julia Kim, and Mark Riedl. Evaluating actionability in explainable ai. *arXiv preprint arXiv:2601.20086*, 2026. 1, 2

[37] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35: 17359–17372, 2022. 2

[38] Richard Meyes, Melanie Lu, Constantin Waubert De Puiseau, and Tobias Meisen. Ablation studies in artificial neural networks. *arXiv preprint arXiv:1901.08644*, 2019. 2

[39] Raphaël Millière and Cameron Buckner. Interventionist methods for interpreting deep neural networks. In *Neurocognitive Foundations of Mind*. Routledge, 2025. 2, 3

[40] Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, et al. The quest for the right mediator: A history, survey, and theoretical grounding of causal interpretability. *arXiv e-prints*, pages arXiv–2408, 2024. 2

[41] Sidra Naveed, Gunnar Stevens, and Dean Robin-Kern. An overview of the empirical evaluation of explainable ai (xai): A comprehensive guideline for user-centered evaluation in xai. *Applied Sciences*, 14(23):11288, 2024. 3

[42] Konstantinos Nikiforidis, Alkiviadis Kyrtsoglou, Thanasis Vafeiadis, Thanasis Kotsiopoulos, Alexandros Nizamis, Dimosthenis Ioannidis, Konstantinos Votis, Dimitrios Tzovaras, and Panagiotis Sarigiannidis. Enhancing transparency and trust in AI-powered manufacturing: A survey of explainable AI (XAI) applications in smart manufacturing in the era of industry 4.0/5.0. *ICT Express*, 11(1):135–148, 2025. 1

[43] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020. 2

[44] Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023. 2

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[46] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any

classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016. 1

[47] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, 2024. 2

[48] Roy Rinberg, Usha Bhalla, Igor Shilov, and Rohit Gandikota. Ripplebench: Capturing ripple effects by leveraging existing knowledge repositories. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025. 8

[49] Terrence J Sejnowski and Charles R Rosenberg. Parallel networks that learn to pronounce english text. *Complex Systems*, 1(1):145–168, 1987. 2

[50] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMlR, 2017. 3

[51] Paul Smolensky. Neural and conceptual interpretation of pdp models. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 2:390–431, 1986. 2

[52] Justus Westerhoff, Erblina Purelku, Jakob Hackstein, Jonas Loos, Leo Pinetzki, Erik Rodner, and Lorenz Hufe. Scam: A real-world typographic robustness evaluation for multimodal foundation models. *arXiv preprint arXiv:2504.04893*, 2025. 4, 6

[53] Xiwei Xuan, Ziquan Deng, Hsuan-Tien Lin, Zhaodan Kong, and Kwan-Liu Ma. Suny: A visual interpretation framework for convolutional neural networks from a necessary and sufficient perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8371–8376, 2024. 2

[54] Xiwei Xuan, Ziquan Deng, Hsuan-Tien Lin, and Kwan-Liu Ma. Slim: Spuriousness mitigation with minimal human annotations. In *European Conference on Computer Vision*, pages 215–231. Springer, 2024. 2

[55] Xiwei Xuan, Jorge Piazentin Ono, Liang Gou, Kwan-Liu Ma, and Liu Ren. Attributionscanner: A visual analytics system for model validation with metadata-free slice finding. *IEEE transactions on visualization and computer graphics*, 2025. 2

[56] Xinyuan Yan, Xiwei Xuan, Jorge Piazentin Ono, Jiajing Guo, Vikram Mohanty, Shekar Arvind Kumar, Liang Gou, Bei Wang, and Liu Ren. Vislix: An xai framework for validating vision models with slice discovery and analysis. In *Computer Graphics Forum*, page e70125. Wiley Online Library, 2025. 2

[57] Yue Yang, Mona Gandhi, Yufei Wang, Yifan Wu, Michael S Yao, Chris Callison-Burch, James Gee, and Mark Yatskar. A textbook remedy for domain shifts: Knowledge priors for medical image analysis. In *Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond*, 2024. 4

[58] Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*, 2023. 2

[59] Hengyuan Zhang, Zhihao Zhang, Mingyang Wang, Zunhai Su, Yiwei Wang, Qianli Wang, Shuzhou Yuan, Ercong Nie, Xufeng Duan, Qibo Xue, et al. Locate, steer, and improve: A practical survey of actionable mechanistic interpretability in large language models. *arXiv preprint arXiv:2601.14004*, 2026. 2, 4

[60] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305, 2020. 3

[61] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38, 2024. 2