A Multi-Dimensional Evaluation of Explainability in Media Bias Detection

Anonymous ACL submission

Abstract

Detecting media bias automatically is difficult because biased framing is often subtle, yet in domains such as news analysis, accurate predictions alone are insufficient without explanations that reflect the model's underlying reasoning. We present a multi-dimensional evaluation of explainability in encoder-based media bias detection using the Bias Annotations By Experts (BABE) dataset. Specifically, we study BERT and RoBERTa as classifiers (base and large variants) along three complementary axes: predictive performance, explanation plausibility (token-level alignment with expert rationales), and mechanistic faithfulness (whether compact sets of attention heads recover predictive signal under counterfactual rationale masking). To induce variation in plausibility, we additionally investigate attention-supervised finetuning, which incorporates expert rationale annotations as an auxiliary training signal. Attention supervision serves as an intervention on attribution plausibility, while the effectiveness of attribution methods varies substantially across architectures. Circuit analysis further reveals substantial variation in mechanistic recoverability across architectures, suggesting that model scale alone does not determine circuit compressibility. Taken together, our findings suggest that predictive performance, attribution plausibility, and mechanistic faithfulness characterize different aspects of model behavior and should be evaluated separately when studying explainability in media bias detection.

1 Introduction

As news consumption moves online, the scale of media production has outpaced the capacity for manual bias analysis. Early detection methods relied on content analysis and qualitative frameworks from the social sciences (Hamborg et al., 2019), but these approaches lack scalability and suffer from annotation inconsistencies. Transformer-based models now surpass manual annotation in consistency (Spinde et al., 2024), and large language models (LLMs) can capture the linguistic nuance through which bias is expressed (Wang et al., 2024).

However, strong classification performance alone is insufficient for media bias detection, where predictions carry editorial and civic weight. A system that flags an article as biased without interpretable reasoning may be dismissed by journalists or misapplied by policymakers (Lyu et al., 2024).

One of the main concerns in eXplainable AI (XAI) is that two key dimensions are often conflated but should be kept separate (Lyu et al., 2024). Plausibility measures whether an explanation aligns with human judgment (i.e., do the highlighted tokens match what an expert annotator would select?) Faithfulness measures whether an explanation reflects the model's actual computation (i.e., does the model rely on those tokens to reach its prediction?) A model can produce plausible but unfaithful explanations when the XAI mechanism concentrates on semantically reasonable tokens that play no causal role in the prediction. Faithful explanations may highlight tokens that a human would not intuitively select. Conflating the two can lead to overconfident assessments of the quality of the explanation.

Prior work on explainable bias detection has focused almost entirely on plausibility, lacking consideration on model faithfulness. For example, attention-based XAI methods are widely implemented (Zheng et al., 2025), but whether attention constitutes a faithful explanation has long been contested (Jain and Wallace, 2019). Raw attention scores often assign the highest weights to function words and delimiters, undermining their reliability even when the overall distribution appears plausible.

Mechanistic interpretability provides a complementary perspective to attention-based plausibility evaluation by focusing on the internal computa-

tions responsible for a model's predictions. Circuit discovery methods, such as ACDC (Conmy et al., 2023), identify the minimal computational subgraphs causally responsible for model behavior. However, current work has focused almost exclusively on decoder-only language models, and the relationship between mechanistic faithfulness and attribution plausibility remains largely unexplored for encoder-based text classification.

We bridge this gap by jointly evaluating plausibility and faithfulness for encoder-based media bias classifiers. Our contributions are:

1. We conduct a multi-dimensional evaluation of explainability in media bias detection, jointly studying predictive performance, attribution plausibility, and mechanistic faithfulness.

2. We adapt activation-patching circuit discovery to encoder-only text classification, introducing Retention and Rescue metrics to evaluate circuit recoverability under counterfactual rationale masking.

3. Using attention supervision as an intervention on plausibility, we characterize how the quality of the explanation varies between attribution methods, model architectures, and mechanistic circuit properties.

2 Related Work

Automated Bias Detection. Early work on media bias detection relied primarily on handcrafted linguistic and contextual features. Previous studies have shown that lexical framing and contextual cues can effectively identify bias-inducing language (Spinde et al., 2021b). More recent work shifted to transformer-based architectures, where studies on the BABE benchmark (Spinde et al., 2021a) demonstrated that contextualized transformer models substantially outperform traditional feature-based approaches in fine-grained bias classification tasks.

Recent research has also explored the use of LLMs for political bias analysis. GPT models can achieve high agreement with human political bias ratings even in zero-shot settings (Hernandes and Corsi, 2024). Despite the strong performance of LLMs, concerns also arose. Studies have demonstrated that larger language models may still exhibit systematic political framing biases despite improved generation quality (Huang et al., 2026). Together, these findings motivate the need for more transparent and faithful explainability methods for bias detection systems.

Explainable NLP. Explainability methods in NLP span post-hoc feature attribution, gradient-based analysis, attention interpretation, probing, and mechanistic circuit analysis (Lyu et al., 2024; Luo and Specia, 2024; Tursunalieva et al., 2024). The breadth of available methods makes multidimensional evaluation necessary.

Rationale-based evaluation frameworks, such as ERASER (DeYoung et al., 2020), formalize the assessment of explanations using human-annotated spans. Chain-of-thought prompting (Yeo et al., 2024; Wei et al., 2023) takes a different approach by eliciting intermediate reasoning steps from the model itself.

Attention-based explanations are among the most studied. Hao et al. and Naim and Asher explore the conditions under which attention provides interpretable insights into the information flow of the transformer. More recent work improves the explainability of self-attention distributions in text classification (Mylonas et al., 2022), and studies demonstrate that specific attention heads correspond to semantically meaningful reasoning patterns (Zheng et al., 2025). However, whether attention scores constitute genuine explanations remains debated (Wiegreffe and Pinter, 2019; Jain and Wallace, 2019).

To move beyond the limitations of attention analysis, the mechanistic interpretability shifts the focus from individual weights to localized computational subgraphs known as circuits (Olsson et al., 2022; Elhage et al., 2021). This line of work has focused primarily on autoregressive, decoder-only architectures. A prominent example is the indirect object identification circuit (IOI) in GPT-2, where activation patching revealed how heads cooperatively route linguistic information to produce predictions (Conmy et al., 2023). More recent work has further emphasized the evaluation of circuit discovery methods through behavioral recovery rather than qualitative inspection alone, benchmarking methods by how well recovered circuits preserve model behavior (Syed et al., 2024; Arad et al., 2025). However, the application of circuit discovery to encoder-based text classification and the comparison of the resulting faithfulness measures against plausibility scores on the same models have not been explored.

Incorporating human rationales during finetun-

ing has also been used to shape model behavior beyond raw classification accuracy, including improving reasoning and reducing hallucination in other settings (Just et al., 2025; Cho et al., 2025).

3 Method

Our study progresses in three stages, illustrated in Figure 1. First, we train the BERT and RoBERTa classifiers on the BABE dataset under both standard supervision and attention-supervised finetuning, using rationale supervision as a controlled intervention on explanation plausibility. Second, we evaluated token-level plausibility using three attention-based attribution methods and compared their outputs against expert rationale annotations. Third, we evaluate mechanistic faithfulness using activation-patching circuit discovery, measuring the extent to which compact sets of attention heads recover predictive signal under counterfactual rationale masking.

3.1 Dataset

Bias Annotations By Experts (BABE). We use the BABE dataset (Spinde et al., 2021a), which contains 3,700 English news sentences drawn from 14 U.S. outlets spanning left, center, and right political leanings according to the AllSides media bias chart. Each sentence carries a binary bias label assigned by majority vote among five trained annotators with backgrounds in data science, psychology, and intercultural communication (Krippendorff's $\alpha = 0.40$ for the bias label). We downsample the majority class to produce a balanced subset of 2,762 sentences (1,381 biased, 1,381 unbiased). For biased instances, the annotators additionally highlighted specific words and phrases responsible for the biased framing. We use these token-level rationales both as supervision during attention-guided finetuning and as ground truth for evaluating the plausibility of explanations. We partition the data using a stratified 70:15:15 train, validation, and test split to ensure class balance across all subsets.

3.2 Models

We experiment with two encoder-only transformer architectures: BERT and RoBERTa. BERT (Devlin et al., 2019) uses bidirectional masked language modeling to produce contextual token representations; its `[CLS]` embedding serves as input for sequence-level bias classification. RoBERTa (Liu et al., 2019) modifies BERT's pretraining procedure with dynamic masking, larger batches, and more data, yielding stronger downstream performance on tasks that require sensitivity to subtle lexical choices. For both architectures, we experiment with base and large variants to examine how model capacity affects classification performance, plausibility, and faithfulness.

3.3 Attention-supervised Finetuning

Expert rationales in BABE make it possible to train the model on both bias labels and on *why* a sentence is biased. We incorporate an auxiliary attention supervision loss (DeYoung et al., 2020; Cho et al., 2025) that encourages the model to allocate higher attention mass to expert-highlighted tokens. We use rationale supervision as an intervention on attribution plausibility. The joint training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \cdot \mathcal{L}_{\text{attn}},$$

where $\mathcal{L}_{\text{cls}}$ is the binary cross-entropy loss over the bias label, $\mathcal{L}_{\text{attn}}$ is the mean squared error between the normalized last-layer CLS attention distribution and the binary expert rationale mask:

$$\mathcal{L}_{\text{attn}} = \frac{1}{N} \sum_{i=1}^{N} \left(\bar{a}_i - r_i\right)^2,$$

where $\bar{a}_i$ denotes the min-max normalized attention weight from the CLS token to the token i and $r_i \in \{0, 1\}$ is the expert rationale indicator. λ controls the tradeoff between predictive accuracy and rationale alignment. We fix $\lambda = 0.5$ across all experiments based on a hyperparameter search ($\lambda \in \{0, 0.2, 0.5, 0.7, 1.0, 1.5, 2.0, 5.0\}$).

3.4 Attention-based Explainability

Last-layer CLS Attention. In the final layers of BERT, a large part of the attention heads focus heavily on task-related tokens (Clark et al., 2019). The most straightforward attribution approach examines the attention distribution from the `[CLS]` token at the final transformer layer L. The importance score for token i is:

$$s_i^{\text{CLS}} = \mathrm{A}_{\text{CLS},\,i}^{(L)}.$$

where $\mathrm{A}_{\text{CLS},\,i}^{(L)}$ is the weight of attention from the `[CLS]` token to token i in the final transformer layer L. This formulation treats higher attention as greater importance to the prediction, but ignores how information is routed through earlier layers and may overweight tokens that are prominent only at the final layer.

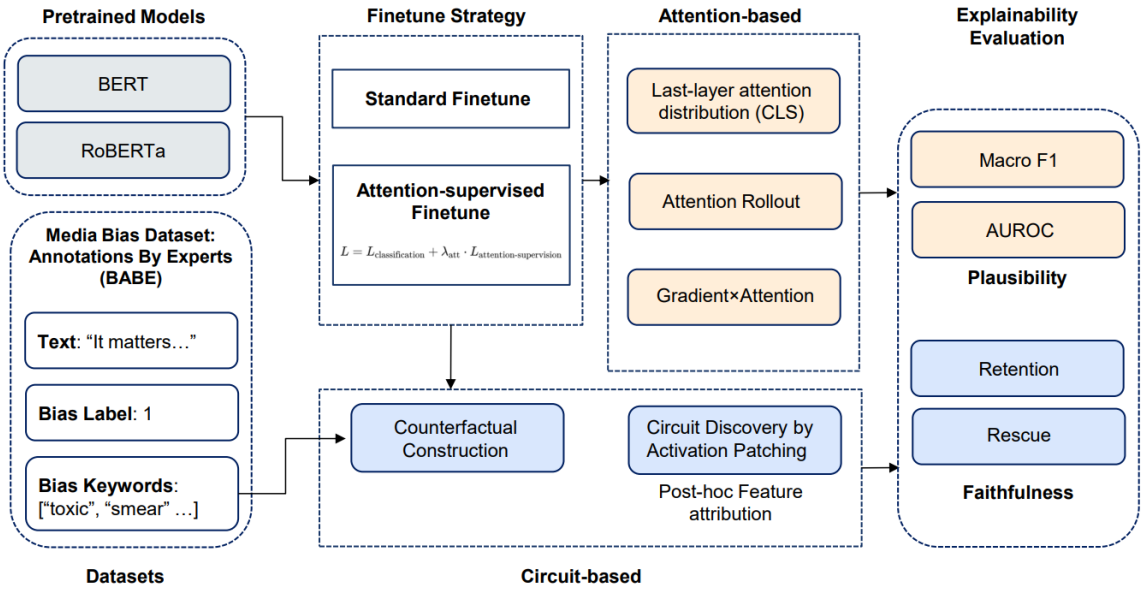


Figure 1: Flowchart of the media bias analysis pipeline.

Attention Rollout. Attention rollout (Abnar and Zuidema, 2020) addresses the limitation of examining only the last single layer by propagating attention across all layers via recursive matrix multiplication, approximating the flow of token information from input to output. At each layer l, we first average the attention matrices across all heads, then add the identity matrix to account for the residual connection:

$$\tilde{\mathrm{A}}^{(l)} = 0.5 \cdot \bar{\mathrm{A}}^{(l)} + 0.5 \cdot I,$$

where $\bar{\mathrm{A}}^{(l)}$ is the head-averaged attention matrix at layer l. The rollout attribution is then computed as follows:

$$\mathrm{Rollout}^{(L)} = \tilde{\mathrm{A}}^{(1)} \tilde{\mathrm{A}}^{(2)} \cdots \tilde{\mathrm{A}}^{(L)},$$
$$s_i^{\text{rollout}} = \mathrm{Rollout}^{(L)}_{\mathrm{CLS},\, i}.$$

Gradient-weighted Attention. Raw attention weights indicate where the model looks, but not whether that attention affects the output. Following Hao et al., we compute gradient-weighted attention (i.e., Gradient×Attention) by scaling the final-layer attention weights by their gradient with respect to the pre-sigmoid logit y for the predicted class using:

$$s_i^{\text{grad}} = \mathrm{A}^{(L)}_{\mathrm{CLS},\, i} \cdot \frac{\partial y}{\partial \mathrm{A}^{(L)}_{\mathrm{CLS},\, i}}.$$

The gradient term measures how sensitive the prediction is to each attention weight, so the product downweights attention heads that are large but causally inert. The result is an attribution that combines the way the model is attended with how much that attention matters to the classification decision.

3.5 Circuit Discovery for Explainability

Attention-based attribution scores measure plausibility, but not whether the model actually relies on the highlighted tokens. To assess faithfulness, we adapt Automated Circuit Discovery (ACDC) (Conmy et al., 2023) to encoder-based classifiers. ACDC identifies the minimal subgraph of a model's computation graph whose activations are sufficient to reproduce the model's predictions.

Counterfactual Masking. We converted sentence-level samples from the BABE dataset (Spinde et al., 2024) into counterfactual pairs. Let $\mathcal{D} = \{(T_i, y_i, K_i)\}$ represent the dataset, where T_i is the raw textual string, $y_i \in \{0, 1\}$ is the binary expert bias flag, and K_i is a set of expert-annotated biased keywords or phrases.

We apply a deterministic corruption function that replaces instances of biased keywords or phrases with an architectural token mask (`[MASK]`). Counterfactual pairs are partitioned using a stratified $70 : 15 : 15$ train, validation, and test split strategy.

Linear Logistic Probing. Circuit analysis is performed after fine-tuning at each model checkpoint. Rather than using the architecture-specific classification heads of BERT and RoBERTa as causal targets, we trained a lightweight linear probe over

the final hidden representations to provide a uniform diagnostic objective across encoder architectures. This design allows the same activation-patching pipeline to operate directly on the shared transformer encoder layers without introducing architecture-specific branching in the downstream readout.

Let

$$\mathbf{h}_i^{(L)} = \text{Encoder}(T_i)_{[\text{CLS}]} \in R^d$$

denote the final-layer representation of the classification token (`[CLS]`) for clean input T_i. We train a logistic regression classifier on a sampled subset of clean hidden representations to predict the target bias label y_i:

$$P(y_i = 1 \mid \mathbf{h}_i^{(L)}) = \sigma(\mathbf{w}^T \mathbf{h}_i^{(L)} + b), \quad (1)$$

where $\mathbf{w} \in R^d$ and $b \in R$ are learned parameters and $\sigma(\cdot)$ denotes the sigmoid activation function. After optimization, the fixed probe parameters $(\mathbf{w}, b)$ serve as the downstream causal readout used during activation patching experiments.

Circuit Extraction by Activation Patching. To quantify the contribution of each attention head, we perform activation patching on all head positions. For each head h at layer ℓ, we run a forward pass on the corrupted input but replace that head's activations with the corresponding clean activations from $\mathcal{C}_{\text{clean}}$.

The causal effect $\mathcal{E}_{(\ell,h)}$ of an individual head is defined as the absolute marginal difference in the expected probability output of the downstream probe:

$$\mathcal{E}_{(\ell,h)} = \Big| \mathbf{E}\left[P(y = 1 \mid \mathbf{h}^{(L)}_{\text{patched}(\ell,h)})\right] - \mathbf{E}\left[P(y = 1 \mid \mathbf{h}^{(L)}_{\text{corrupt}})\right] \Big|$$

Taking the absolute value ensures that heads contributing to predictions of either class (biased or unbiased) are ranked by the magnitude of their causal influence rather than direction. Subsequently, all heads are sorted in descending order of $\mathcal{E}_{(\ell,h)}$ to form a causal importance ranking.

Circuit Verification Modes. To evaluate the collective characteristics of the causal heads ranked in the top-k, we assemble them into a candidate circuit sub-network $\mathcal{G}_k = \{(\ell,h)_1, (\ell,h)_2, \ldots, (\ell,h)_K\}$. We implement two complementary patch styles via forward hooks to assess the circuit's functional traits on the held-out test split:

1. **Circuit Elimination (`no mode`):** we run clean inputs T through the model but replace the outputs of all heads in $\mathcal{G}_k$ with their corrupted counterparts from $\mathcal{C}_{\text{corrupt}}$. This tests whether the circuit is *causally necessary* for the model's predictive gap.

2. **Circuit Isolation (`circuit_only mode`):** We run the corrupted inputs T^{corrupt} but restore clean activations from $\mathcal{C}_{\text{clean}}$ for the heads in $\mathcal{G}_k$ only. This tests whether the circuit is *causally sufficient* (sufficient to recover the original predictive behavior under activation patching.) to recover the model's predictive signal.

Since circuit size is an important hyperparameter, we sweep $k \in \{1, 3, 5, 10, 20, 30\}$ to characterize the faithfulness-compactness corelation for each model.

3.6 Evaluation Metrics

Macro F1. We evaluate classification performance using Macro F1, which computes F1 independently for each class and averages with equal weight. This protects against performance inflation from class imbalances and ensures that reported gains reflect genuine improvement in both biased and unbiased instances.

AUROC for Attention Plausibility. The Area Under the Receiver Operating Characteristic Curve (AUROC) is a standard metric for evaluating token-level ranking performance (Mandrekar, 2010; DeYoung et al., 2020). To evaluate whether attention-based attributions align with human rationales, we computed the AUROC between token-level attribution scores and binary expert rationale masks. An AUROC of 0.5 indicates that the attributions are no better than random when identifying the expert-highlighted tokens; a value of 1.0 indicates perfect alignment. AUROC measures the plausibility of the explanation *plausibility*: how well the model-derived token importance matches the human judgment.

Retention and Rescue for Circuit Faithfulness. To evaluate circuit faithfulness, we compare the predicted probability gap of the model under three evaluation conditions: the unperturbed baseline model ($\Delta P(\text{base})$), the circuit-eliminated model ($\Delta P(\text{none})$), and the circuit-isolated model ($\Delta P(\text{circuit\_only})$). From these conditions, we

adapt the sufficiency and comprehensiveness measures introduced in the ERASER benchmark (DeYoung et al., 2020) and inspired by Conmy et al., reframed here to evaluate causal circuits rather than discrete token rationales. We present two complementary causal metrics, Retention and Rescue. **Retention** quantifies the fraction of the full model's probability gap that the isolated subgraph reproduces:

$$\text{retention} = \frac{\Delta P(\text{circuit\_only})}{|\Delta P(\text{base})| + \epsilon},$$

where $\epsilon = 10^{-12}$ is a regularizing scalar to prevent division by zero. The absolute value in the denominator ensures that retention remains interpretable when the base probability gap is negative (i.e., when the model's mean predicted probability for unbiased instances exceeds that for biased instances). Retention of 1.0 indicates that the circuit fully reproduces the probability gap of the base model, while values near 0 indicate that the circuit captures little predictive signal. **Rescue** measures the signed predictive margin by which the circuit recovers beyond the fully ablated baseline.

$$\text{rescue} = \Delta P(\text{circuit\_only}) - \Delta P(\text{none}).$$

Positive rescue indicates that the circuit recovers the predictive signal beyond what the ablated model retains. The rescue score complements retention by confirming that high retention values are not inflated by a near-zero ablated baseline. Together, retention and rescue quantify faithfulness by measuring the degree to which the isolated circuit is causally responsible for the model's bias predictions.

4 Results

4.1 Attention Analysis and Plausibility

Table 1 reports Macro F1 and AUROC scores for all models. All transformer models substantially outperform the feature-based baseline reported by Spinde et al. (2021a) (Macro F1 = 0.511) and the BERT baseline reported in the same work (Macro F1 = 0.762), with every configuration exceeding 0.80 Macro F1. RoBERTa-large achieved the strongest predictive performance in general, reaching a Macro F1 of 0.863 under attention-supervised finetuning.

The effect of attention supervision on predictive performance is modest and architecture-dependent. Attention supervision does not produce a uniform improvement across architectures. While RoBERTa-large benefits from rationale supervision, other configurations show smaller gains or slight decreases relative to standard finetuning. This suggests that rationale supervision acts as a weak auxiliary signal whose effects depend on the model architecture and capacity rather than fundamentally changing the learned representations.

Plausibility is strongly attribution- and architecture-dependent. We evaluate the plausibility of the explanation using AUROC, comparing CLS attention, Attention Rollout, and Gradient×Attention. No single attribution method consistently dominates across architectures. Gradient×Attention generally performs strongly in BERT models, while Attention Rollout and CLS attention exhibit greater variability in architectures and finetuning conditions. These findings suggest that the quality of the attribution depends not only on the explanation method itself but also on the underlying attention structure of the model being analyzed.

Overall, the alignment between model explanations and expert rationales remains moderate. Most configurations achieve AUROC values only modestly above the random baseline of 0.5, with substantial variance between examples and architectures. These results are consistent with previous work that argued that attention distributions alone do not reliably constitute faithful explanations of model behavior (Jain and Wallace, 2019).

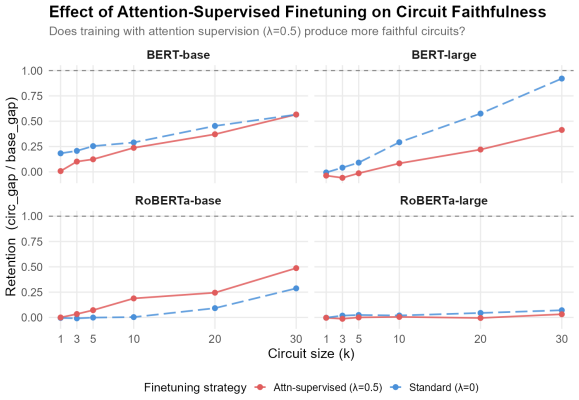
4.2 Circuit Analysis and Faithfulness

Circuit retention varies across architectures and interventions. Figure 2a compares retention under standard and attention-supervised finetuning. The effect of rationale supervision is mixed across architectures. The RoBERTa-base exhibits noticeably improved retention under attention supervision, while BERT-large and RoBERTa-large show weaker or even reduced retention relative to standard finetuning. The BERT-base exhibits similar retention under both training strategies once larger circuits are considered. These results suggest that rationale supervision does not uniformly improve mechanistic faithfulness and that circuit recoverability depends strongly on model architecture and representational structure.

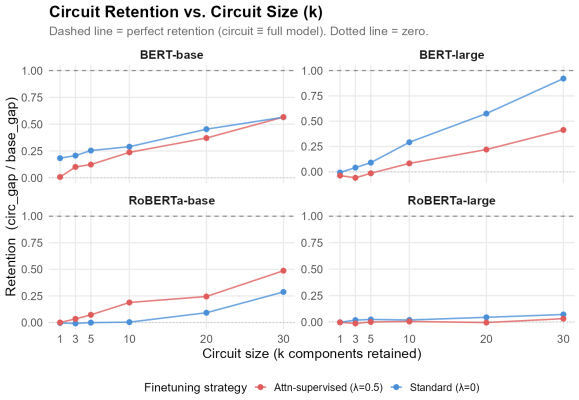
Circuit faithfulness varies substantially across architectures. Figure 2b shows that retention

Model	Finetune Strategy	Macro F1	Explanation Plausibility (AUROC)		
			CLS	Rollout	Grad × Attn
BERT baseline (Spinde et al., 2021a)	–	0.762	–	–	–
BERT (base)	Standard	0.829	0.523 (0.218)	0.497 (0.196)	0.583 (0.297)
	Attention-supervised	0.805	0.476 (0.227)	0.506 (0.196)	0.566 (0.284)
BERT (large)	Standard	0.812	0.554 (0.220)	0.515 (0.162)	0.621 (0.321)
	Attention-supervised	0.809	0.487 (0.211)	0.496 (0.170)	0.535 (0.286)
RoBERTa (base)	Standard	0.843	0.492 (0.283)	0.426 (0.292)	0.687 (0.271)
	Attention-supervised	0.831	0.575 (0.320)	0.465 (0.287)	0.674 (0.317)
RoBERTa (large)	Standard	0.849	0.470 (0.284)	0.484 (0.240)	0.620 (0.271)
	Attention-supervised	**0.863**	0.512 (0.312)	0.522 (0.231)	0.661 (0.245)

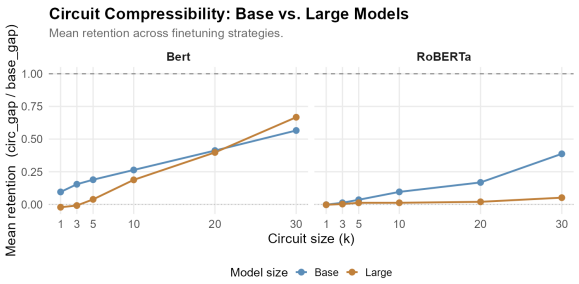
Table 1: Classification performance and explanation plausibility. AUROC values (with standard deviations) measure alignment between attribution method outputs and expert rationales in the BABE dataset.



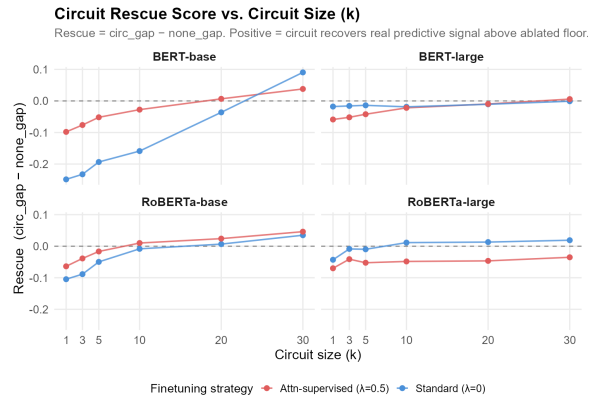
(a) Effect of attention-supervised finetuning.



(b) Circuit retention vs. circuit size.



(c) Circuit compressibility across model sizes.



(d) Circuit rescue scores across circuit sizes.

Figure 2: Circuit faithfulness analysis across encoder-based bias detection models. Attention-supervised finetuning generally improves circuit retention and rescue scores, particularly in smaller models.

generally increases with circuit size k, indicating that the predictive signal is distributed over multiple components rather than concentrated in a single attention head. However, architectures differ substantially in compressibility (Figure 2c). The BERT-family models exhibit a relatively monotonic increase. The RoBERTa-base shows a similar trend under attention supervision but near-zero retention under standard finetuning, whereas the RoBERTa-large remains difficult to compress even for large circuit sizes. These findings suggest that model architecture and scale influence the degree to which predictive behavior can be recovered using compact mechanistic circuits.

Rescue analysis reveals meaningful causal contribution. Retention alone can be inflated when

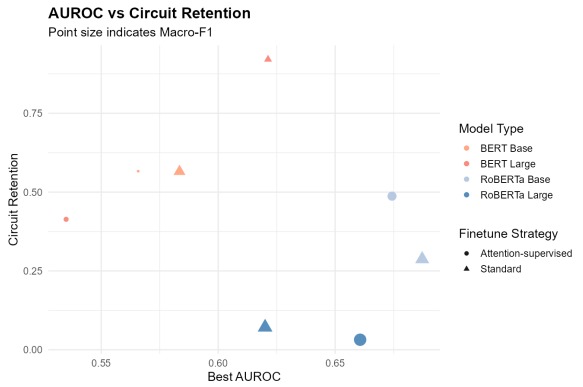


Figure 3: Relationship between attribution plausibility (best AUROC across attribution methods) and circuit retention ($k = 30$) across all model configurations. Point size corresponds to Macro F1.

the ablated baseline is near zero, so we also examine the rescue scores (Figure 2d). Positive rescue confirms that the circuit recovers the predictive signal above the ablated baseline. The attention-supervised models of the BERT-base and the RoBERTa-base show increasingly positive rescue as k grows, consistent with the circuits that capture genuine predictive computation. The Standard-finetuned RoBERTa-large, by contrast, remains near or below zero across most settings, pointing to a weak causal structure. Rescue scores become positive only at larger circuit sizes, which indicates that bias detection is not governed by a single sparse mechanism but emerges from coordination among multiple heads.

Relationships among explanation dimensions. Figure 3 compares plausibility of attribution and circuit retention in all model configurations. We do not observe a consistent monotonic relationship between the two quantities. Models with strong rationale alignment do not necessarily exhibit high circuit retention, while some configurations with only moderate plausibility scores achieve relatively recoverable circuits. In particular, the RoBERTa models achieve some of the strongest plausibility scores in our experiments but often display comparatively weak circuit retention, whereas several BERT configurations exhibit the opposite pattern.

These observations suggest that the plausibility of the attribution and the mechanistic faithfulness capture complementary but distinct aspects of the model behavior. Models with strong rationale alignment do not necessarily exhibit highly recoverable circuits, indicating that improvements in one explanation metric do not automatically transfer to another.

5 Conclusion

We evaluated BERT and RoBERTa classifiers for media bias detection along three axes: predictive performance, plausibility of explanations, and mechanistic faithfulness. Attention-supervised finetuning serves as an intervention for studying explanation behavior, while attribution analyzes reveal only moderate agreement with expert rationales and substantial variation across architectures and explanation methods. Circuit discovery through activation patching further reveals substantial variation in mechanistic recoverability across models. Rescue analysis in particular suggests that bias prediction emerges from coordination among multiple attention heads rather than a single sparse mechanism. Overall, our findings indicate that plausibility and faithfulness capture complementary but distinct aspects of model behavior and thus should be assessed as non-interchangeable notions of explanation quality. Because explanations in media bias detection may inform fairness judgments, journalistic analysis, and model auditing, we argue that trustworthy systems should be evaluated along both token-level plausibility and component-level faithfulness dimensions, each on its own terms.

6 Limitations

Dataset scale and diversity. All experiments use the BABE dataset, which is relatively small and topically constrained compared to real-world media environments. The observed attention patterns and circuit structures may therefore reflect dataset-specific artifacts rather than general properties of bias detection. Evaluation on larger and more diverse news corpora is needed.

Statistical power. The number of model configurations evaluated is relatively small, limiting the strength of conclusions regarding relationships among predictive performance, plausibility, and mechanistic faithfulness. Future work should examine a broader range of architectures, datasets, and training interventions.

Approximate circuit discovery. Our mechanistic analysis relies on activation patching with top-k component selection, which provides an approximate view of the model's internal computation. The resulting circuits should not be interpreted as the unique or complete mechanism underlying bias

detection. Encoder-based transformers distribute information bidirectionally across layers, making compact causal decomposition harder than in autoregressive architectures.

Architectural scope. Our experiments are limited to encoder-only models (BERT and RoBERTa). Circuit discovery may behave differently in decoder-only or instruction-tuned language models, where autoregressive token flow introduces different computational structures.

7 Ethical Considerations

Dual-use risk. Automated bias detection could be misused to craft less detectable biased content rather than to identify and mitigate it. We emphasize that our methods are intended to support transparency in media analysis, not to enable adversarial manipulation of news framing.

Subjectivity of bias labels. The BABE dataset relies on expert annotations that reflect the annotators' cultural and political perspectives. Bias is context-dependent and contested; no single annotation scheme captures all forms of framing. Model outputs should be treated as one signal among many rather than as definitive judgments.

Risk of misapplication. Deploying bias classifiers without interpretable explanations risks unjustified censorship or editorial suppression. Our work on explainability is motivated in part by this concern: bias predictions should be accompanied by transparent reasoning to support human decision-making rather than automated content moderation.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. *Preprint*, arXiv:2005.00928.

Dana Arad, Yonatan Belinkov, Hanjie Chen, Najoung Kim, Hosein Mohebbi, Aaron Mueller, Gabriele Sarti, and Martin Tutek. 2025. Findings of the BlackboxNLP 2025 shared task: Localizing circuits and causal variables in language models. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 543–552, Suzhou, China. Association for Computational Linguistics.

Hakaze Cho, Peng Luo, Mariko Kato, Rin Kaenbyou, and Naoya Inoue. 2025. Mechanistic fine-tuning for in-context learning. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 330–357, Suzhou, China. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does bert look at? an analysis of bert's attention. *Preprint*, arXiv:1906.04341.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *Preprint*, arXiv:2304.14997.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. *Preprint*, arxiv:1911.03429 [cs].

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. *Preprint*, arxiv:2004.11207 [cs].

Raphael Hernandes and Giulio Corsi. 2024. Llms left, right, and center: Assessing gpt's capabilities to label political bias from web domains. *Preprint*, arXiv:2407.14344.

Nannan Huang, Iffat Maab, and Junichi Yamagishi. 2026. When bigger isn't better: A comprehensive fairness evaluation of political bias in multi-news summarisation. *Preprint*, arXiv:2604.21309.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. *Preprint*, arXiv:1902.10186.

Hoang Anh Just, Ming Jin, Anit Sahu, Huy Phan, and Ruoxi Jia. 2025. Data-centric human preference with rationales for direct preference alignment. *Preprint*, arxiv:2407.14477 [cs].

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Haoyan Luo and Lucia Specia. 2024. From understanding to utilization: A survey on explainability for large language models. *Preprint*, arxiv:2401.12874 [cs].

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723. Place: Cambridge, MA Publisher: MIT Press.

Jayawant N. Mandrekar. 2010. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316.

Nikolaos Mylonas, Ioannis Mollas, and Grigorios Tsoumakas. 2022. Improving attention-based interpretability of text classification transformers. In *Artificial Neural Networks and Machine Learning – ICANN 2022*.

Omar Naim and Nicholas Asher. 2024. On explaining with attention matrices.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Preprint*, arXiv:2209.11895.

Timo Spinde, Smi Hinterreiter, Fabian Haak, Terry Ruas, Helge Giese, Norman Meuschke, and Bela Gipp. 2024. The media bias taxonomy: A systematic literature review on the forms and automated detection of media bias. *Preprint*, arxiv:2312.16148 [cs].

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021a. Neural media bias detection using distant supervision with BABE – bias annotations by experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177.

Timo Spinde, Lada Rudnitckaia, Jelena Mitrović, Felix Hamborg, Michael Granitzer, Bela Gipp, and Karsten Donnay. 2021b. Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing Management*, 58(3):102505.

Aaquib Syed, Can Rager, and Arthur Conmy. 2024. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 407–416, Miami, Florida, US. Association for Computational Linguistics.

Ainura Tursunalieva, David L. J. Alexander, Rob Dunne, Jiaming Li, Luis Riera, and Yanchang Zhao. 2024. Making sense of machine learning: A review of interpretation techniques and their applications. *Applied Sciences*, 14(2):496.

Zhiqiang Wang, Yiran Pang, Yanbin Lin, and Xingquan Zhu. 2024. Adaptable and reliable text classification using large language models. *Preprint*, arxiv:2405.10523 [cs]. Version: 3.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arxiv:2201.11903 [cs].

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. *Preprint*, arxiv:1908.04626 [cs].

Wei Jie Yeo, Ranjan Satapathy, Rick Siow Mong Goh, and Erik Cambria. 2024. How interpretable are reasoning explanations from prompting large language models? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2148–2164.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2025. Attention heads of large language models. *Patterns*, 6(2):101176.

A Licenses and Artifact Usage

Our experiments use the BABE dataset (Spinde et al., 2021a), which is distributed under the CC BY-NC-SA 4.0 license for non-commercial research use. We additionally use pretrained BERT models released under the Apache License 2.0 and RoBERTa models released under the MIT License through the HuggingFace Transformers library. All experiments comply with the respective licenses and intended research usage terms of these artifacts.

B Dataset Statistics

Table B1 summarizes the BABE dataset used in our experiments after preprocessing and class balancing. Notably, expert-annotated rationales cover only 6.18% of tokens on average, with a mean of 1.72 rationale tokens per biased sentence. This extreme sparsity makes plausibility evaluation particularly challenging, as attribution methods must identify a small number of signal-bearing tokens against a large background of uninformative context.

C Hyperparameter Configuration

Table C1 lists the hyperparameters used across all experiments. All models were trained using the same configuration unless otherwise noted.

Table C2 reports validation-set performance across the full λ sweep used for hyperparameter selection on BERT-base. The results reveal

Statistic	Value
Total sentences	2762
Biased sentences	1381
Unbiased sentences	1381
Avg. sentence length (tokens)	31.38
Avg. rationale tokens per biased sentence	1.72
Avg. rationale coverage (%)	6.18
Train / Val / Test split	70 / 15 / 15
Class balancing	Yes (downsampled)

Table B1: BABE dataset statistics after preprocessing and class balancing.

Hyperparameter	Value
Optimizer	AdamW
Learning rate	2×10^{-5}
Learning rate schedule	Linear decay
Batch size	16
Max sequence length	128
Number of epochs	4
Early stopping patience	N/A
Attention loss weight λ	0.5 (selected from sweep)
Random seeds	5

Table C1: Hyperparameter configuration for all finetuning experiments.

a non-monotonic relationship: moderate values ($\lambda \in [0.2, 0.7]$) improve both Macro F1 and rationale AUROC over the unsupervised baseline ($\lambda = 0$), while stronger supervision ($\lambda \geq 1.5$) degrades both metrics. At $\lambda = 5.0$, Macro F1 drops by over five points, indicating that excessive attention pressure distorts the learned representations and harms the primary classification objective. The selected value of $\lambda = 0.5$ achieves the best joint performance.

λ	Macro F1	AUROC (Grad×Attn)
0.0	0.818	0.623
0.2	0.822	0.641
0.5	**0.824**	**0.658**
0.7	0.822	0.651
1.0	0.817	0.644
1.5	0.819	0.628
2.0	0.801	0.611
5.0	0.764	0.573

Table C2: Attention supervision weight (λ) sweep on the validation set used for hyperparameter selection. Moderate supervision strengths improve both classification performance and rationale plausibility, while overly strong supervision degrades task performance and explanation quality. $\lambda = 0.5$ provides the best tradeoff between predictive accuracy and rationale alignment.

All experiments were executed using a SLURM-managed computing cluster. Each run requested a GPU with 32GB of memory, while the specific GPU model varied depending on resource availability across compute nodes. Table C3 reports the approximate training time and total compute for each model configuration.

Model	Parameters	Train Time
BERT-base	110M	∼3 min
BERT-large	340M	∼9 min
RoBERTa-base	125M	∼4 min
RoBERTa-large	355M	∼10 min
λ sweep (BERT-base only)		8 values × 5 seeds
Total GPU hours (all experiments)		∼3.5 GPU-hours

Table C3: Computational budget. Training times are per single run (one seed, one λ setting) on a single GPU. Times are estimated based on dataset size (1,933 training samples), batch size 16, sequence length 128, and 4 epochs.

D Per-Class and Per-Seed Results

Table D1 disaggregates classification performance by class and finetuning condition, averaged across 5 sampling seeds. Across most architectures, attention-supervised models tend to increase precision on the biased class while reducing recall, suggesting that rationale supervision encourages more conservative bias predictions. For example, BERT-base attention-supervised achieves 0.907 biased precision (vs. 0.880 standard) but only 0.732 biased recall (vs. 0.789). RoBERTa-base exhibits the opposite behavior, trading a small reduction in biased precision (0.901 to 0.889) for a substantial increase in biased recall (0.803 to 0.841), resulting in improved biased-class F1. Standard deviations remain small across seeds (≤ 0.038), indicating stable training dynamics.

Model	FT	Biased			Unbiased		
		P	R	F1	P	R	F1
BERT-b	Std	$.880_{\pm 15}$	$.789_{\pm 22}$	$.832_{\pm 15}$	$.766_{\pm 19}$	$.865_{\pm 19}$	$.813_{\pm 14}$
	Attn	$.907_{\pm 11}$	$.732_{\pm 23}$	$.810_{\pm 14}$	$.729_{\pm 16}$	$.906_{\pm 13}$	$.808_{\pm 10}$
BERT-l	Std	$.868_{\pm 13}$	$.801_{\pm 38}$	$.832_{\pm 20}$	$.773_{\pm 32}$	$.846_{\pm 21}$	$.808_{\pm 16}$
	Attn	$.877_{\pm 18}$	$.800_{\pm 31}$	$.837_{\pm 24}$	$.775_{\pm 30}$	$.860_{\pm 20}$	$.815_{\pm 24}$
RoB-b	Std	$.901_{\pm 20}$	$.803_{\pm 25}$	$.849_{\pm 12}$	$.783_{\pm 18}$	$.888_{\pm 26}$	$.832_{\pm 11}$
	Attn	$.889_{\pm 10}$	$.841_{\pm 22}$	$.864_{\pm 10}$	$.813_{\pm 19}$	$.868_{\pm 16}$	$.840_{\pm 08}$
RoB-l	Std	$.908_{\pm 14}$	$.831_{\pm 34}$	$.868_{\pm 20}$	$.810_{\pm 29}$	$.894_{\pm 19}$	$.849_{\pm 18}$
	Attn	$.918_{\pm 18}$	$.798_{\pm 20}$	$.854_{\pm 18}$	$.783_{\pm 20}$	$.911_{\pm 20}$	$.842_{\pm 19}$

Table D1: Per-class precision (P), recall (R), and F1. Subscripts denote std. dev. ($\times 10^{-3}$) across 5 seeds. BERT-b/l = BERT base/large; RoB-b/l = RoBERTa base/large; Std = standard; Attn = attention-supervised.

E Full Circuit Discovery Results

Table E1 presents the complete activation patching results across all models, finetuning conditions, and circuit sizes ($k = 1$ to 30). Across nearly all configurations, retention increases with circuit size, indicating that predictive information is distributed across multiple attention heads rather than concentrated in a single component. Rescue scores are typically negative for very small circuits and become positive only as additional heads are incorporated, suggesting that meaningful predictive behavior emerges from coordinated computation across multiple components.

Circuit behavior varies substantially across architectures. BERT-base and RoBERTa-base exhibit steadily increasing retention and rescue as circuit size grows, indicating progressively more recoverable internal computation. RoBERTa-large, by contrast, maintains weak retention and rescue scores even for larger circuits, suggesting more distributed predictive representations. Because retention is normalized by the base predictive gap, we additionally report Rescue as a denominator-free measure of recovered predictive signal.

Model	FT	Base	k	None	Circ	Faithfulness	
						Ret.	Resc.
BERT-b	Std	0.3059	1	0.3045	0.0560	0.1830	−0.2485
			3	0.2958	0.0633	0.2071	−0.2325
			5	0.2712	0.0778	0.2542	−0.1935
			10	0.2476	0.0887	0.2900	−0.1589
			20	0.1748	0.1387	0.4533	−0.0362
			30	0.0825	0.1729	0.5653	0.0905
	Attn	0.1200	1	0.0990	0.0009	0.0076	−0.0981
			3	0.0885	0.0122	0.1015	−0.0763
			5	0.0667	0.0149	0.1238	−0.0518
			10	0.0560	0.0285	0.2372	−0.0275
			20	0.0379	0.0445	0.3710	0.0067
			30	0.0300	0.0680	0.5663	0.0380
BERT-l	Std	0.0163	1	0.0177	−0.0001	−0.0061	−0.0178
			3	0.0165	0.0007	0.0415	−0.0158
			5	0.0154	0.0015	0.0911	−0.0139
			10	0.0234	0.0048	0.2922	−0.0187
			20	0.0199	0.0094	0.5754	−0.0106
			30	0.0159	0.0150	0.9210	−0.0009
	Attn	0.0627	1	0.0564	−0.0023	−0.0371	−0.0588
			3	0.0481	−0.0037	−0.0588	−0.0518
			5	0.0412	−0.0009	−0.0141	−0.0421
			10	0.0272	0.0052	0.0836	−0.0219
			20	0.0234	0.0138	0.2202	−0.0096
			30	0.0201	0.0260	0.4137	0.0058
RoB-b	Std	0.1259	1	0.1040	−0.0006	−0.0045	−0.1046
			3	0.0876	−0.0011	−0.0086	−0.0886
			5	0.0495	−0.0001	−0.0011	−0.0497
			10	0.0088	0.0004	0.0035	−0.0083
			20	0.0049	0.0116	0.0918	0.0066
			30	0.0014	0.0362	0.2873	0.0347
	Attn	0.0869	1	0.0638	−0.0000	0.0000	−0.0638
			3	0.0418	0.0029	0.0338	−0.0388
			5	0.0230	0.0063	0.0721	−0.0167
			10	0.0064	0.0164	0.1882	0.0100
			20	−0.0028	0.0213	0.2446	0.0240
			30	−0.0037	0.0423	0.4871	0.0460
RoB-l	Std	0.0468	1	0.0429	−0.0001	−0.0024	−0.0430
			3	0.0095	0.0009	0.0183	−0.0086
			5	0.0110	0.0011	0.0244	−0.0098
			10	−0.0104	0.0009	0.0195	0.0113
			20	−0.0110	0.0021	0.0445	0.0131
			30	−0.0156	0.0033	0.0712	0.0189
	Attn	0.0707	1	0.0698	−0.0002	−0.0030	−0.0700
			3	0.0402	−0.0009	−0.0131	−0.0411
			5	0.0524	0.0000	0.0004	−0.0524
			10	0.0487	0.0004	0.0054	−0.0483
			20	0.0461	−0.0004	−0.0050	−0.0465
			30	0.0374	0.0022	0.0318	−0.0351

Table E1: Full circuit discovery results via activation patching. **Base**: original logit gap; **None**: gap after ablating the circuit; **Circ**: gap using only top-k components. **Ret.** = Circ/|Base| (sufficiency); **Resc.** = Circ − None (recovered signal). All values rounded to 4 decimal places.