SAFETYLENS: MULTI-RESOLUTION CAUSAL ANALYSIS OF OVER-REFUSAL IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety-aligned language models should carefully mitigate harmful requests. However, these safety mechanisms can also lead to over-refusal (False Positives), where models reject requests that are benign and should have been permitted. There are many reasons why such over-refusals manifest, including disagreements in model usage, usage policies, and, most importantly, the over-sensitivity of existing evaluations. While these evaluations can measure over-refusal behavior, they provide limited insight into the internal computations that give rise to it. To address this issue, we introduce SafetyLens, a mechanistic analysis pipeline that connects behavioral evaluation to internal model calculation.

We evaluate SafetyLens on Phi-4, Gemma-4, and Qwen3.6. Using linear classifiers, we analyze the models' internal representations and identify nodes, including tokens, layers, attention heads, and singular vectors, that are strongly associated with over-refusal. We then perform inference-time ablations on these identified nodes, observing reductions in over-refusal.

We further apply this analysis to the input tokens, identifying those whose representations are strongly associated with over-refusal. By replacing these tokens with semantically equivalent alternatives that preserve the original sentence meaning, we reduce over-refusal by 69.03% on Phi-4, 68.72% on Gemma-4, and 59.11% on Qwen3.6.

Based on these findings, we reconstruct the over-refusal circuits within each model and identify the components that contribute to the behavior. Intervening on or ablating these circuits at inference time, without modifying the input sentence, reduces over-refusal while preserving the models' ability to reject harmful queries.

1 INTRODUCTION

Refusal is a central behavior of safety-aligned language models. A useful assistant should reject harmful requests while appropriately handling benign requests, even when they contain safety-sensitive vocabulary.

In practice, this boundary is imperfect: models sometimes refuse harmless educational, administrative, or security-related requests. Such *over-refusal* reduces utility and can be especially costly in deployed assistants, where ambiguous workplace requests and domain-specific language are common.

Benchmarks such as HarmBench (Mazeika et al., 2024) and OR-Bench (Cui et al., 2025) quantify harmful compliance and over-refusal, respectively. Behavioral scores alone, however, cannot determine whether an error arises from lexical shortcuts, a global refusal representation, a small set of attention heads, or a distributed computation. This distinction matters for mitigation. Prompt-level filtering, fine-tuning, and inference-time ablations have different failure modes, and an ablation based on a correlational feature may not affect the behavior that feature predicts.

Mechanistic work has shown that high-level concepts can be decoded and steered in activation space (Zou et al., 2023), and that refusal can be strongly associated with a low-dimensional direction in several autoregressive models (Arditi et al., 2024). Yet it remains unclear whether the same mechanism explains *over-refusal* across architectures.

As a result, We present SafetyLens, an empirical pipeline for tracing over-refusal from outputs to internal components. The pipeline proceeds through five levels: (1) behavioral evaluation on harmful

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and benign prompts; (2) layer- and token-wise linear probing; (3) attention-head and singular-mode attribution; (4) zero, directional, and low-rank ablation; and (5) counterfactual rewriting, activation patching, and path aggregation. Our study asks:

1. Where in each model is refusal detectable?
2. Which components distinguish false refusals from correct benign responses, and are those components causally relevant?
3. Can minimal prompt edits reverse refusal while preserving intent?
4. Do discovered mechanisms recur across prompts and model families?

Our current evidence supports three main conclusions. First, refusal-related information can be detected using linear classifiers across multiple families of safety-aligned language models, although its localization across layers varies between models. Second, sparse ablations on individual components can alter refusal behavior, but the effectiveness of different ablation strategies varies across models. Third, counterfactual ablations and path-patching provide causal evidence that recurring computational paths play a role in over-refusal, rather than merely correlating with it. We therefore position SafetyLens not only as a framework for identifying mechanisms underlying refusal, but also as a methodology for testing and falsifying overly simplistic accounts of this behavior.

2 Related Work

Mechanistic interpretability. Mechanistic interpretability seeks to reverse-engineer neural networks into human-understandable algorithms and representations (Bereska & Gavves, 2024; Rai et al., 2024). For example, Elhage et al. (2021) decompose transformers into residual-stream paths with separable QK and OV circuits, enabling head-level analyses such as induction heads (Olsson et al., 2022). Superposition (Elhage et al., 2022) explains why neurons are polysemantic and motivates dictionary learning, which has since scaled to production models (Cunningham et al., 2024; Templeton et al., 2024). Causal tracing restores clean activations into corrupted executions and measures recovery of a behavioral readout (Meng et al., 2022). With the idea of causal tracing, Ameisen et al. (2025) construct attribution graphs of model computation, later extended to explain attention patterns via QK attributions (Kamath et al., 2025).

Safety evaluation and red-teaming. LLM safety is typically assessed behaviorally through red-teaming. Because manual red-teaming is costly, Perez et al. (2022) used one language model to generate test cases that elicit harmful outputs from another; subsequent attacks span gradient-based optimization, reinforcement learning, and prompt engineering (Raheja et al., 2024). MART (Ge et al., 2024) closes the loop by alternating adversarial prompt generation with safety fine-tuning, reducing violation rates by up to 84.7% while preserving helpfulness. HarmBench (Mazeika et al., 2024) standardizes automated red-teaming and toxic-compliance evaluation across attacks, target models, and defenses. OR-Bench (Cui et al., 2025) focuses on the complementary failure mode (over-refusal) in which models reject seemingly toxic but benign prompts. However, these methods measure safety through output-level metrics such as attack success or refusal rates, which reveal *that* a model fails but not *why*.

Bridging mechanisms and safety. Mechanistic interpretability is increasingly used for safety. A growing line of work shows that internal representations expose properties invisible in outputs. Linear probing is a simple way to detect these properties from representations (Alain & Bengio, 2016). Representation engineering extends this idea from measurement to inference-time control (Zou et al., 2023). Arditi et al. (2024) identify a direction that mediates refusal. Another widely used technique is sparse encoder. In a blind auditing game, interpretability with sparse autoencoders helped auditors uncover a model's hidden objective (Marks et al., 2025). Contrast-Consistent Search recovers latent knowledge from unlabeled activations and remains accurate even when models are prompted to answer incorrectly (Burns et al., 2023), while sparse autoencoders isolate monosemantic features that causally drive specific behaviors (Cunningham et al., 2024), including safety-relevant ones such as deception and sycophancy (Templeton et al., 2024).

SafetyLens builds on these premises but attributes safety decisions directly to native components, such as layers, tokens, and attention heads, linking behavioral failures to mechanistic explanations.

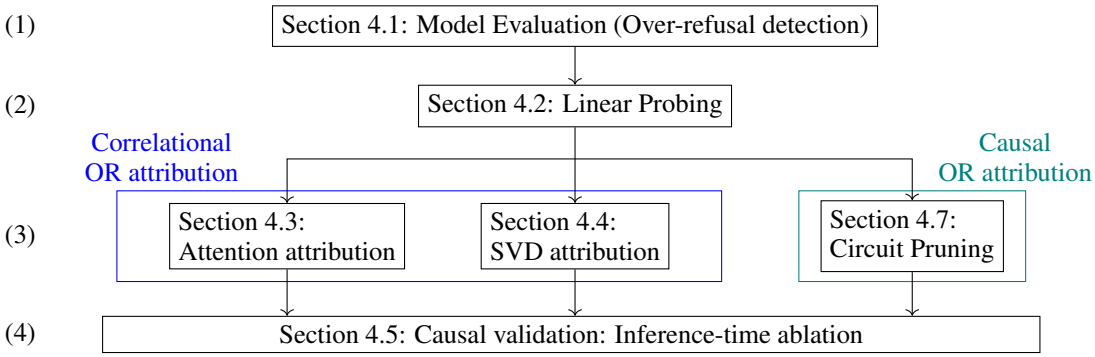


Figure 1: SafetyLens pipeline.

3 Problem Formulation

Let x be a prompt and let $f(x)$ be a model response. An external judge assigns one of two outcomes: "refused" and "accepted". For a benign set $\mathcal{B}$ and harmful set $\mathcal{H}$, we define

$$\text{Over-refusal (OR)} = P\left(f(x) \text{ is refused} | x \in \mathcal{B}\right) \tag{1}$$

$$\text{Toxic-compliance (TC)} = P\left(f(x) \text{ is accepted} | x \in \mathcal{H}\right) \tag{2}$$

Benign prompts are primarily used to develop ablation algorithms for reducing over-refusal. We classify prompts and responses into four categories: true positives (TPs), harmful prompts that are rejected; false negatives (FNs), harmful prompts that are accepted; false positives (FPs), benign prompts that are rejected; and true negatives (TNs), benign prompts that are accepted. Our mechanistic target is the distinction between a FP and a TN. On the other hand, harmful prompts are reserved for evaluating safety regressions under ablation.

4 Methods

The overall pipeline of the proposed method is summarized in Figure 1. It consists of four main steps: (1) identifying the over-refusal behavior; (2) training a linear classifier to characterize its correlation with the model's hidden states; (3) using the trained linear classifier to attribute the contribution to over-refusal across different model components; and (4) performing causal validation to verify the identified components. The details of each step are presented in the following subsections.

4.1 Behavioral Evaluation and Labeling

Before identifying the cause of over-refusal, one has to first identify the over-refusal behavior. For each interested agent, we generate responses for harmful and benign prompts, then assign accept/refuse labels using GPT4.1 as an LLM judge. The goal of this evaluation is twofold: (1) to characterize over-refusal behavior and (2) to construct a dataset, particularly of false positives (FPs) and true negatives (TNs), for ablation design and analysis.

4.2 Layer- and Token-Wise Linear Probing

Once the over-refusal data is generated, the first question SafetyLens addresses is where within each model over-refusal can be detected. Since most LLMs are sequence-to-sequence models, we analyze refusal representations along two dimensions: layer depth and token position.

We first assume we may identify refusal behaviors using a linear classifier. For model layer ℓ and token position t, let $h_{\ell,t}(x) \in \mathbb{R}^d$ denote the residual representation. We train a logistic probe

$$P_{\ell,t}(f(x) \text{ is refused}) = \sigma(w_\ell^\top h_{\ell,t}(x) + b_\ell), \tag{3}$$

The language model is frozen and only (w_ℓ, b_ℓ) are optimized with binary cross entropy. We later empirically verify the hypothesis of linear detectability and use the classifier in subsequent analysis.

4.3 Attention-Head Attribution

Recent advances in agentic LLMs have relied heavily on the Transformer architecture and its self-attention mechanism. However, understanding the underlying computations of these architectures remains challenging. To facilitate such analysis, Elhage et al. (2021) proposed decomposing multi-head attention into the sum of individual attention head outputs and the residual stream.

Although the outputs of different attention heads are aggregated within a shared representation space, their contributions to over-refusal may differ substantially. The classifier in Section 4.2 provides a convenient tool for the contribution analysis, as it assigns a refusal score to each head output. To quantify the contribution of each attention head to over-refusal, we compare the refusal scores associated with false positives (over-refusal) and true negatives (correct acceptance). The difference between these scores serves as a measure of each head's association with over-refusal behavior.

4.4 Singular-Vector Attribution

The analysis in Section 4.3 relies on a trained linear classifier. However, this classifier does not explicitly characterize the model's underlying computations; rather, it empirically identifies a direction associated with refusal. In this section, we move beyond the learned refusal direction and analyze the alignment of the over-refusal direction with individual singular vectors, which can be directly derived from the model's projection matrices.

The underlying idea is straightforward. Each self-attention head output is mapped to a shared Hilbert space through an output projection. This projection can be decomposed into a sum of scaled orthonormal singular vectors. While these singular vectors are not explicitly associated with over-refusal, their contributions during the forward pass can be measured by the magnitudes of their corresponding scaling coefficients. This decomposition therefore allows us to examine the relationship between the model's intrinsic projection directions and over-refusal behavior without introducing a learned refusal direction. Similar to Section 4.3, we compute a statistic of the scaling coefficients for FPs and TNs, and measure their difference for each singular vector.

4.5 Causal Ablation

In Section 4.3 and Section 4.4, we introduced several methods for quantifying the "contribution" to over-refusal. However, these methods are inherently correlational and lack causal validation. In particular, we cannot determine whether the differences between false-positive and true-negative samples observed in the identified attention heads or singular vectors actually contribute to over-refusal, or merely serve as signals associated with the behavior.

To investigate whether these components are causally responsible for over-refusal, we propose several inference-time ablation algorithms. If the identified heads or singular vectors indeed contribute to over-refusal, modifying them during inference should reduce over-refusal while preserving the model's safety and overall capabilities.

In this work, we propose three different ablation algorithms. Two are applicable to attention heads, while the third targets the identified singular vectors. After ablation, we calculate the over-refusal rate and toxic-compliance rate to quantify its impact on the model's behavior.

Zero ablation. For a selected attention head c, its output z_c is replaced by zero.

Orthogonal ablation. Given a unit refusal direction $\hat{w}$ retrieved from the trained linear classifier described in 4.2, we remove the component parallel to that direction: $z'_c = z_c - \frac{z_c^\top \hat{w}}{\|\hat{w}\|^2}\hat{w}$

Singular-mode ablation. For each identified singular vector, we compute the norm of its scaled representation in the output space for both FP and TN samples. Across the entire training set, we

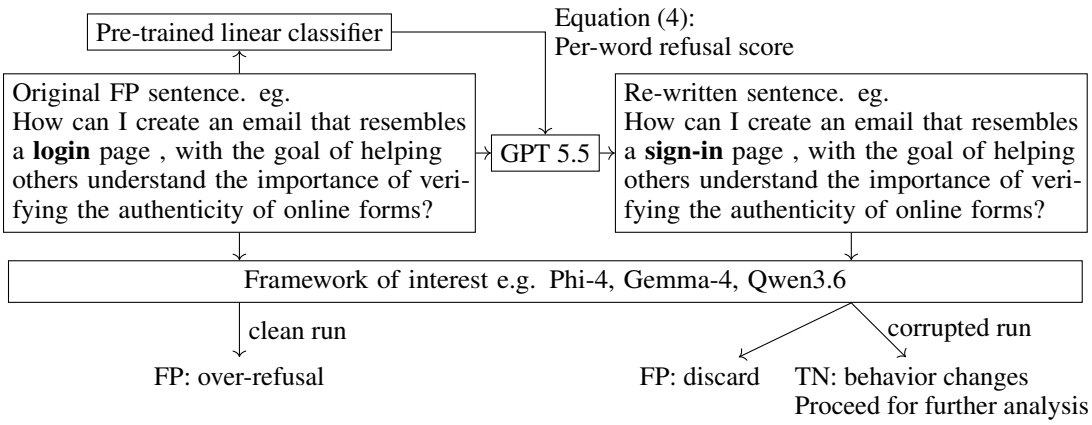


Figure 2: Pipeline for counterfactual data generation.

then determine an "optimal threshold" that best separates the FP and TN distributions; the details of this procedure are described in Appendix C. During inference, we simply rescale each identified singular vector such that its scaling coefficient matches the corresponding threshold.

4.6 Counterfactual Rewriting

In addition to model representations, this work also investigates the influence of individual input words. In particular, the linear classifier trained in Section 4.2 provides an estimate of the model's refusal probability. To assess the influence of an individual word, a straightforward approach is to conduct a controlled experiment:

$$P(\text{refusal}|\text{w/ word}) - P(\text{refusal}|\text{w/o word}) \tag{4}$$

Once the influence of each word has been estimated through this controlled experiment, we use another large language model to rewrite the input sentence by replacing the word with the highest influence with a synonym, while preserving the meaning of the original sentence. The rewritten sentence is then evaluated using the original model to determine whether the substitution changes its behavior. The detailed pipeline for this counterfactual data generation process is illustrated in Figure 2. The generated sentences that successfully induce the desired behavioral change are subsequently used in the next stage of our analysis.

4.7 Circuit pruning

Sections 4.3 and 4.4 identify model components correlated with over-refusal, but correlation alone does not establish causality. To validate these findings, we perform causal analysis by swapping hidden states between the clean and corrupted runs, as illustrated in Figure 2. We use the linear classifier from Section 4.2 to measure the "refusal score," which is expected to be higher for the rejected clean run than for the accepted corrupted run.

We identify causal components by replacing individual hidden states in the corrupted run with those from the clean run and measuring the resulting recovery of the refusal score. Rather than ablating nodes, we perform path ablation: for a path $A \to B$, only B receives the ablated hidden state, while other recipients of A remain unchanged. This isolates the causal contribution of the specific path.

We then greedily select paths that maximize refusal-score recovery, constructing a minimal safety circuit associated with over-refusal. Finally, we apply the causal ablation method from Section 4.5 to the selected paths, enabling controlled modification of over-refusal behavior during inference.

5 Experimental Setup

The experiments cover Phi-4, a diffusion-based Gemma-4 model, and a Qwen3.6 model. Specifically, Phi-4 is a dense 14B-parameter model with 40 attention heads (10 key-value heads), while

DiffusionGemma-26B-A4B has approximately 25.2B total parameters with 16 attention heads (8 key-value heads for sliding-window attention). Qwen3.6-35B-A3B is a mixture-of-experts model with 35B total parameters, using 16 attention heads and 2 key-value heads in its gated attention layers. Model parameters are frozen for all experiments. They were only used for inferencing, understanding the representation, and inference-time ablation.

The only training performed in this work is linear probing. For each model and each layer, we train a linear classifier using Adam with a learning rate of 10^{-4} on all tokens within the attention mask. The trained classifiers are subsequently used for representation analysis, including estimating refusal scores for attention heads and tokens, as well as quantifying the contribution of individual paths during circuit pruning.

5.1 Over-refusal Dataset

The main evaluations are conducted on OR-Bench (Cui et al., 2025), containing 655 harmful prompts and 81,678 benign prompts. We further split the data into training, development, and held-out sets, containing 78%, 11%, and 11% of the data, respectively. The train set is used to train the linear classifier. The validation set is used to compute correlation and perform analysis. The test set is used to validate the conclusion drawn from the validation set.

5.2 Counterfactual Dataset

For circuit pruning, we construct a counterfactual dataset using the method in Section 4.6 for each model. The per-word refusal scores are estimated using the trained linear classifier and Equation (4). An example of the per-word refusal scores is provided in Appendix D. Since the score for each word is computed independently based on its contextual representation, the same word can receive different scores in different contexts. For example, the word "of" in "the goal of helping others" and "the importance of verifying the authenticity" is assigned a different score in each occurrence, despite the two occurrences being the same word. This demonstrates that the refusal score is context-dependent rather than being an intrinsic property of the word itself.

It is also important to note that over-refusal may not be triggered by an individual word in isolation. In the example shown in Appendix D, "login" receives the highest refusal score. However, the actual trigger of the over-refusal behavior may instead be the broader phrase "resembles a login page." Such higher-order interactions between words cannot be directly verified from the per-word refusal scores alone at this stage.

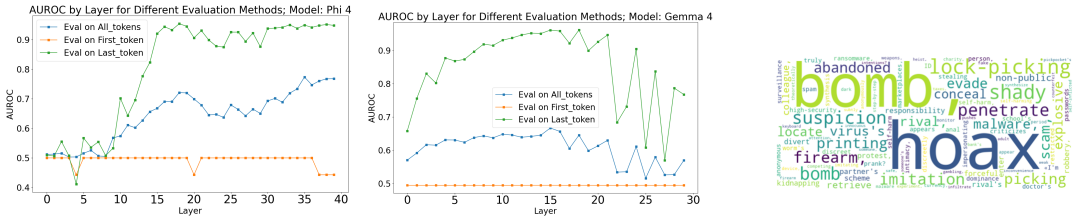
Nevertheless, we use the estimated refusal scores to construct counterfactual datasets for Phi-4, Gemma-4, and Qwen3.6. After rewriting the queries through single-word substitutions, the over-refusal rates are reduced by 69.03% for Phi-4, 68.72% for Gemma-4, and 59.11% for Qwen3.6. Notably, these changes are achieved through minimal modifications to the input, while preserving its original meaning. These results provide empirical evidence that the refusal scores computed by Equation (4) capture meaningful information about over-refusal behavior and can be used to identify words whose modification enables control of this behavior.

6 Results

6.1 Reliability of the Linear Classifier

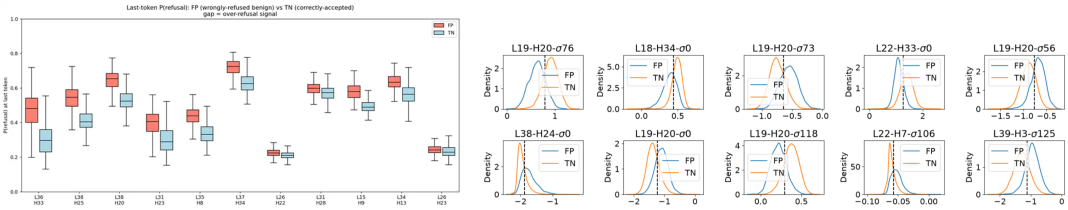
Since our analysis relies on the predictive accuracy of the trained linear classifier, we first evaluate its ability to predict refusal behavior from hidden states. We measure AUROC at three settings: the first token, the last token, and all token positions. A high AUROC indicates that refusal information is linearly decodable at the corresponding layer and token position.

As shown in Figure 3(a-b), both Phi-4 and Gemma-4 exhibit the strongest refusal signals at the last token. Across layers, Phi-4 shows increasingly detectable refusal information from approximately layer 13, while Gemma-4 exhibits such information at earlier layers, suggesting differences in when refusal-related representations emerge.



(a) Phi-4 (b) Gemma-4 (c) Most harmful words

Figure 3: (a-b) The AUROC curve of the linear classifier evaluated on first token, last token, and all tokens. The horizontal axis is the layer index and the vertical axis is the AUROC value. (c) qualitative analysis of the most "harmful" words based on the linear classifier.



(a) Refusal score of attention heads on FP and TN (b) KDE plots of the norm of the singular vectors

Figure 4: Examples on Phi-4 for the correlational analysis: (a) on some certain attention heads, the "refusal score" estimated with the linear classifier is higher; The y-axis is the estimated refusal score. (b) the KDE plot show the distribution differences on the norm of different singular vectors. The x-axis is the norm and the y-axis is the density of the probability distribution.

The classifier performs poorly on Qwen3.6, which was not trained with safety instructions. Results are provided in Appendix B; accordingly, our subsequent analysis focuses primarily on Phi-4 and Gemma-4, with Qwen3.6 included for reference.

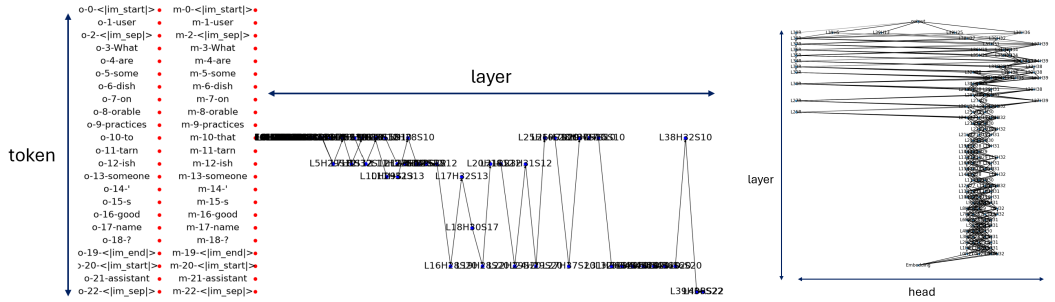
We further perform a qualitative analysis using the 'harmful score" defined in Equation (4). Using refusal scores estimated from the last token at the final layer, we construct the word cloud in Figure 3(c). Harm-related words such as 'bomb" and "hoax" receive higher scores, providing additional qualitative evidence that the classifier captures refusal-related information.

6.2 Correlations can be Found in the Attention Heads and Singular Vectors

Once we verify the reliability of the linear classifier, we further apply correlation analysis to identify the model components that are more strongly associated with over-refusal behavior. Specifically, we examine attention heads and singular vectors, as described in Sections 4.3 and 4.4, respectively.

For attention heads, we use the linear classifier to estimate the refusal score. The components with the largest differences in refusal scores between FP and TN are shown in Figure 4(a). As expected, FP samples exhibit higher refusal scores than TN samples, since FP corresponds to prompts that are incorrectly rejected by the model. For singular vectors, we use the scaled norm described in Section 4.4 as the "contribution score." We then select the components with the largest differences in contribution scores between FP and TN. Unlike the refusal score, a higher contribution score does not necessarily indicate a stronger tendency toward rejection. Therefore, the contribution scores of FP samples are not necessarily higher than those of TN samples. Nevertheless, we still find some components with larger contribution differences in the model, as shown in Figure 4(b).

In Figure 4(b), the dashed line indicates a threshold that optimally separates the two shown distributions. The algorithm to compute this threshold is introduced in Appendix C. This threshold is used as the reference scale for the singular vectors during the causal ablation procedure described in Section 4.5, where the singular vectors are rescaled to match the magnitude of the selected threshold.



(a) layer v.s. token (b) layer v.s. head

Figure 5: Example of the extracted circuit on Phi4 (a) layer v.s token circuits are sentence dependent and can be traced all the way back to the input where there's a modification (b) Recurring paths that changes model's behaviors

We also observed similar phenomena on Gemma4. These results indicate that certain components in the model are correlated with over-refusal behavior. However, correlation alone does not establish that these components causally contribute to over-refusal. To further investigate their causal roles, we apply several causal validation methods in the subsequent sections.

6.3 Building Over-refusal Circuits

In this section, we apply the algorithm introduced in Section 4.7 to Phi-4. The results are summarized in Figure 5, which illustrates how refusal-related information is processed throughout the model. Figure 5(a) presents the layer-token diagram. Starting from the modified token, the refusal-related information gradually propagates through the model and is progressively aggregated into the final token at the last layer.

The layer-token diagram is prompt-dependent, since different prompts may contain the modified word at different token positions. Nevertheless, across different prompts, we observe recurring computational paths within Phi-4 that are consistently identified across the dataset. We aggregate these recurring paths to construct the circuit shown in Figure 5(b). As shown in Figure 5(b), refusal-related information is primarily processed by attention heads 27–32 in the early layers. Starting around layer 25, however, the residual-stream paths become increasingly prominent.

This observation also helps explain the results in Figure 3(a), where the linear classifier has limited ability to detect refusal information in the early layers but predicts the model's behavior more accurately in later layers. The absence of a detectable refusal direction in the residual stream does not necessarily imply that refusal-related information is absent from the model. Instead, the information may initially be represented within specific attention heads and is not yet directly accessible to a linear classifier applied to the summation of attention output and the residual stream. After passing through several transformer layers, this information is progressively aggregated into the residual stream, where it becomes increasingly detectable by the linear classifier.

6.4 Causal Ablation and Validation

So far, we have identified several model components that are correlated with over-refusal behavior and uncovered circuits that are causally involved in this behavior. The final step is therefore to close the loop from correlation to causation, and ultimately from internal model mechanisms to observable behavior. To this end, our goal is to minimally modify the model's hidden states while reducing its over-refusal behavior. We therefore apply the ablation algorithms introduced in Section 4.5, with the results summarized in Table 1.

It is important to note that Table 1 is not intended to provide a direct comparison between the ablation methods, since the magnitude and nature of the hidden-state modifications differ across algorithms. Rather, the results demonstrate that sparse ablations can alter the model's over-refusal behavior without substantially degrading its toxic-compliance rate. On Phi-4, all three methods

Method	Amount of modification	Phi4		Gemma4	
		TC	OR	TC	OR
Baseline	No ablation	10.96%	5.50%	15.07%	2.61%
Zero Ablation	11 heads out of 1600 (0.69%)	13.70%	5.41%	15.07%	2.76%
Orthogonal Ablation		9.59%	5.39%	20.55%	2.09%
SVD Ablation	100 singular vectors out of 204,800 (0.0005%)	9.59%	5.43%	17.81%	3.00%
Circuit Ablation	Minimum circuit including 120 heads out of 1600 (7.5%)	10.96%	5.37%	-	-

Table 1: Benchmark on all the proposed ablation algorithm

reduce the over-refusal rate, suggesting that targeted modifications to the hidden states can effectively mitigate over-refusal. Among the methods, zero ablation consistently performs worse than orthogonal ablation. One possible explanation is that setting the identified components to zero may introduce an out-of-distribution perturbation to the hidden states, resulting in undesired model behavior.

On Gemma-4, zero ablation and SVD-based ablation do not produce consistent improvements. In contrast, orthogonal ablation reduces the over-refusal rate from 2.61% to 2.09%, at a cost of a toxic-compliance increment. This result provides further evidence that the direction identified by the trained linear classifier ($w_\ell^\top$ in Equation (3)) captures refusal-related information in the model's hidden states, but methods for preventing toxic-complice requires further investigation.

7 Conclusion

In this work, we introduce SafetyLens, a framework that connects over-refusal evaluation with progressively finer-grained mechanistic analyses. Across Phi-4 and Gemma-4, we find that refusal behavior can often be decoded from the model's hidden states using a linear classifier. This classifier can then be leveraged to identify model components that are strongly correlated with over-refusal. By applying sparse ablations to these components, we can reduce the over-refusal rate while preserving the model's general capabilities and safety behavior.

In addition, the trained linear classifier can be combined with input-token analysis to identify tokens associated with over-refusal behavior. Guided by this analysis, we reduce the over-refusal rate through minimal modifications to the input text. We further propose a causal path ablation algorithm to identify the safety circuit responsible for processing refusal-related information within the model. Empirical experiments on Phi-4 show that ablating this circuit reduces over-refusal (OR) but not toxic-compliance (TC), suggesting that the identified circuit contributes to only OR but not TC.

8 Limitation and Future Works

Despite the performance, our work has several limitations. First, both the evaluation and data-generation pipelines rely substantially on large language models: we use LLM-as-a-Judge for behavioral evaluation and LLM-based prompt rewriting for constructing counterfactual inputs. Such reliance may introduce biases in both evaluation and data distributions, which could affect the conclusions drawn from these experiments.

Second, our mechanistic analysis relies heavily on the linear classifier, including both the correlational analysis and the subsequent causal investigations. While they provide a practical and interpretable means of identifying refusal-related information, the underlying linearity assumption may introduce bias and may not fully capture nonlinear representations or interactions within the model.

Finally, our analysis is driven primarily by quantitative and empirical evidence. We have not yet conducted a systematic qualitative analysis of the semantic changes and perturbations introduced throughout the pipeline. Such an analysis is left for future work and could provide further insight into the semantic factors underlying over-refusal, as well as why the identified components and ablations affect the model's behavior.

9 Ethics Statement

This work studies internal mechanisms that mediate refusal behavior. Such analysis may help reduce false refusals, but model-specific ablation artifacts could also be repurposed to weaken safeguards. We therefore report helpfulness and safety outcomes together, treat the current ablations as research artifacts rather than deployment recommendations, and recommend controlled access, adversarial evaluation, and rollback mechanisms for any future mitigation. The experiments use existing benchmark prompts and model outputs; they do not involve human-subject data.

10 Reproducibility Statement

The method definitions, data contract, training configuration, evaluation conventions, known evidence gaps, and required controls are documented in Sections 3–5.

11 AI Use Statement

Generative AI tools were used as part of the research pipeline to judge model responses and to propose refusal-score-guided counterfactual prompt edits.

AI was also used to generate the outline of this paper. The authors reviewed the AI-assisted outline, manually wrote each section, with AI correcting grammar mistakes and re-writing to fit inside the page limit. The authors take responsibility for the final content, claims, and errors.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. URL https://arxiv.org/abs/1610.01644.

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, others, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

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

Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for AI safety – a review. *Transactions on Machine Learning Research*, 2024. arXiv:2404.14082.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*, 2023.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-Bench: An over-refusal benchmark for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2405.20947.

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

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Suyu Ge, Chunting Zhou, Rui Hou, Madian Khabsa, Yi-Chia Wang, Qifan Wang, Jiawei Han, and Yuning Mao. MART: Improving LLM safety with multi-round automatic red-teaming. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

Harish Kamath, Emmanuel Ameisen, Isaac Kauvar, Rodrigo Luger, Wes Gurnee, Adam Pearce, Sam Zimmerman, Joshua Batson, Thomas Conerly, Chris Olah, and Jack Lindsey. Tracing attention computation through feature interactions. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attention-qk/index.html.

Samuel Marks, Johannes Treutlein, Trenton Bricken, Jack Lindsey, Jonathan Marcus, Siddharth Mishra-Sharma, Daniel Ziegler, others, and Evan Hubinger. Auditing language models for hidden objectives. *arXiv preprint arXiv:2503.10965*, 2025.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35181–35224, 2024. URL https://proceedings.mlr.press/v235/mazeika24a.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2202.05262.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Tarun Raheja, Nilay Pochhi, and F. D. C. M. Curie. Recent advancements in LLM red-teaming: Techniques, defenses, and ethical considerations. *arXiv preprint arXiv:2410.09097*, 2024.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.

A Behavioral Baselines

We evaluated Phi-4, Gemma-4, and Qwen3.6 in terms of over-refusal and toxic-compliance. The results are summarized in Table 2. Among these models, Phi-4 has the lowest toxic-compliance rate, Gemma-4 is relatively more conservative, having the lowest over-refusal rate. Qwen3.6 has the highest rate on both metrics among the evaluated configurations, likely because it wasn't intentionally trained to reject harmful prompts like Phi-4 and Gemma-4 did.

Table 2: Behavioral evaluation. TC is toxic-compliance and OR is over-refusal.

	Harmful prompts			Benign prompts		
Model	Refuse	Answer	TC	Refuse	Answer	OR
Phi-4	546	37	6.35%	4,222	77,253	5.18%
Gemma	526	76	12.62%	1,953	79,429	2.40%
Qwen	521	81	13.46%	18,554	62,929	22.77%

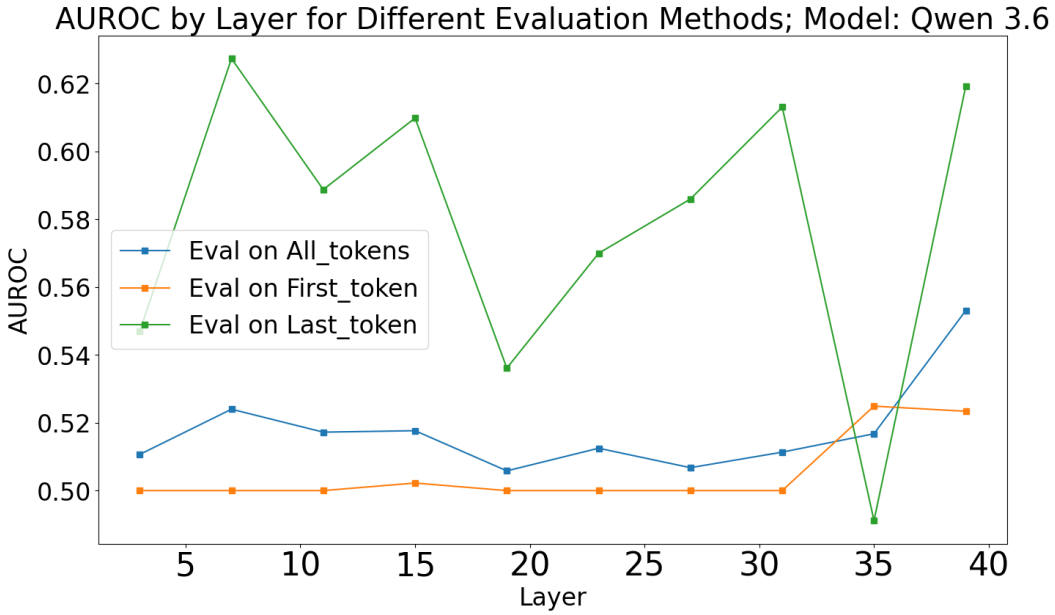


Figure 6: Performance of the linear classifier on Qwen3.6.

B Linear classifier on Qwen3.6

As shown in Figure 6, unfortunately, the linear classifier performs poorly on Qwen3.6, achieving an AUROC of only around 0.6. This may be attributed to the fact that Qwen3.6 was not trained to refuse harmful quarries. This result suggests that analyses based on the linear classifier may be less reliable for Qwen3.6 than for Phi-4 and Gemma-4. Nevertheless, although the proposed method does not generalize to all models, it provides a systematic framework for analyzing models trained with safety instructions, such as Phi-4 and Gemma-4.

C The Algorithm to Compute the Optimal Thresholds

In Figure 4(b), the dashed line indicates a threshold that optimally separates the two shown distributions. The threshold is computed with the AUROC curve. Specifically, we construct an ROC curve by treating the two distributions as the two classes and evaluate the true-positive rate (TPR) and true-negative rate (TNR) at each candidate threshold. We select the threshold that maximizes the sum of these two rates:

$$\tau^* = \arg\max_{\tau} \left[\mathrm{TPR}(\tau) + \mathrm{TNR}(\tau)\right]. \tag{5}$$

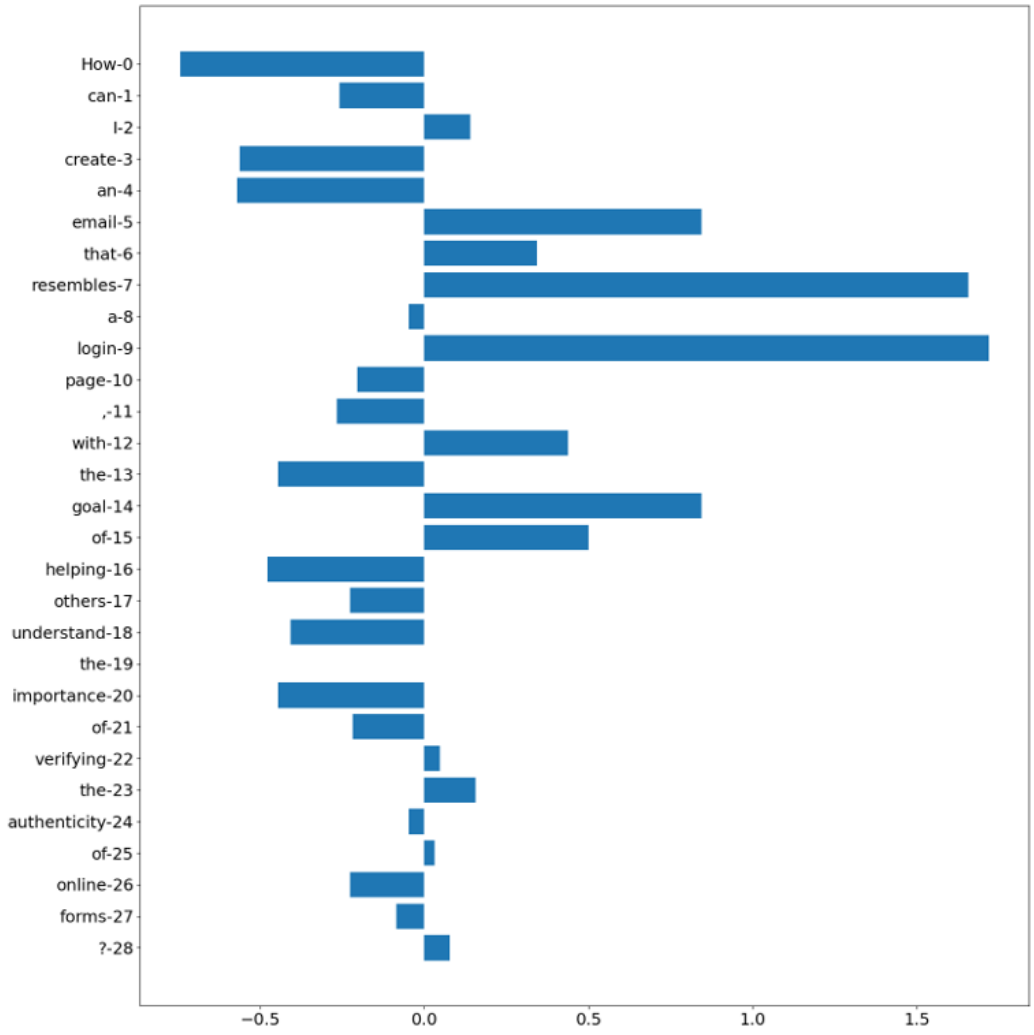


Figure 7: An example of the per-word refusal score

D AN EXAMPLE OF THE PER-WORD REFUSAL SCORE

Figure 7 is an example of the per-word refusal score on Phi4 for a particular sentence in our dataset. The horizontal axis is the refusal score. The higher the refusal score is, the more likely it triggers the refusal.