Improving Explanation Faithfulness via Counterfactual Tests and Activation-Level Steering

Anonymous ACL submission

Abstract

Recent advances in explainable artificial intelligence have emphasized generating natural language explanations (NLE) to justify model predictions. However, NLE often fail to faithfully reflect a model's underlying decision process, potentially misleading users and undermining trust in deployed systems. In this work, we aim to improve explanation faithfulness in the natural language inference (NLI) setting by automatically constructing a dataset of unfaithful explanations using counterfactual tests and leveraging it for activation-level steering. Starting from the e-SNLI dataset, we apply rule-based counterfactual edits that locally modify hypotheses and regenerate NLI labels and explanations for the edited premise–hypothesis pairs. Among cases where the predicted label changes, we identify unfaithful explanations as those that completely ignore the attribute introduced by the counterfactual edit. To reduce false positives from surface-level matching, we further introduce attribute-based semantic filtering. Using the resulting high-confidence unfaithful explanations, we compute steering vectors via Contrastive Activation Addition (CAA) and apply them during decoding to adjust the model's internal representations toward greater causal alignment between predictions and explanations. Experimental results show consistent improvements in explanation faithfulness not only under the Adding Modifier (AM) rule but also across multiple counterfactual rules. Importantly, NLI prediction accuracy on in-distribution evaluation sets remains largely unchanged, indicating that the proposed method enhances explanation faithfulness without degrading predictive performance.

1 Introduction

Explainable Artificial Intelligence (XAI) aims to make model decisions transparent and trustworthy for human users, and its importance has grown alongside the rapid adoption of large language models (LLM). Among various explanation paradigms, NLE are particularly appealing because they present model predictions and their justifications in an intuitive, human-readable form. However, recent studies have shown that NLE often fail to faithfully reflect a model's actual reasoning process, producing explanations that are plausible yet causally unrelated to the underlying decision logic (Atanasova et al., 2020). Such mismatches can mislead users into trusting incorrect rationales and may obscure model biases or failure modes.

In this context, faithfulness has emerged as a central criterion for evaluating explanation quality, referring to how accurately an explanation reflects the factors that truly influence a model's prediction. Prior work has proposed several methods for assessing faithfulness, including counterfactual tests and input reconstruction tests (Atanasova et al., 2023). In particular, counterfactual tests evaluate explanations by applying controlled edits to the input that induce prediction changes and verifying whether the causal source of the change is reflected in the explanation. Despite their effectiveness as evaluation tools, existing approaches largely stop at diagnosis and do not provide mechanisms for improving the faithfulness of generated explanations.

To address this limitation, we reinterpret counterfactual tests not merely as evaluation procedures but as a data construction framework for learning faithful explanations. Specifically, in an e-SNLI–based NLI setting, we apply rule-based counterfactual edits that locally modify the hypothesis and regenerate NLI labels and explanations for the edited premise–hypothesis pairs. Among cases where the predicted label changes, we automatically collect high-confidence unfaithful explanations—those in which the attribute introduced by the counterfactual edit is completely ignored in the explanation. To mitigate false positives arising from surface-level lexical matching, we further introduce attribute-based semantic filtering, which determines faithfulness at the level of semantic at-

tributes rather than exact word overlap.

Building on this automatically constructed unfaithful explanation dataset, we propose an explanation alignment method based on activation-level steering using CAA (Rimsky et al., 2024). Without modifying model parameters, we compute steering vectors from contrastive representations of faithful and unfaithful explanations and inject them into the decoding process to adjust internal activations toward greater causal alignment. We evaluate our approach beyond a single counterfactual rule by extending experiments to multiple counterfactual rules. Additionally, we analyze NLI prediction performance to verify that improvements in explanation faithfulness do not come at the cost of degraded reasoning accuracy. Our contributions are summarized as follows:

1. We propose an automatic framework for constructing high-confidence unfaithful NLE datasets using counterfactual tests, enhanced with attribute-based semantic filtering to improve data quality.

2. We extend counterfactual faithfulness testing beyond a single rule by systematically applying multiple counterfactual generation rules, demonstrating that the proposed framework is rule-agnostic and robust across diverse counterfactual perturbations.

3. We introduce a CAA-based activation-level steering method that leverages contrastive signals between faithful and unfaithful explanations to improve explanation faithfulness without retraining.

4. We demonstrate consistent faithfulness improvements across multiple counterfactual rules, while maintaining NLI performance on in-distribution evaluation data.

2 Related Work

NLE have become a central component of explainable artificial intelligence, as they provide intuitive, human-readable justifications for model predictions. Despite their appeal, numerous studies have reported that NLE often fail to faithfully reflect a model's actual decision-making process, raising concerns that users may be misled into trusting incorrect or irrelevant rationales (He et al., 2024). This issue is commonly framed as a problem of faithfulness, which has been recognized as a key challenge in explainable AI.

Faithfulness evaluation for NLE is generally based on two criteria: (i) whether the explanation reflects factors that truly contribute to the model's decision, and (ii) whether the explanation allows one to reconstruct the model's reasoning. To this end, prior work has proposed rationale-based methods that compare token-level importance, as well as input perturbation approaches such as sufficiency, comprehensiveness, and ROAR (Rawat, 2016). However, these techniques largely rely on token attribution scores or visual indicators, making them difficult to apply directly to freely generated natural language explanations (Atanasova et al., 2020).

As an alternative, counterfactual tests have recently gained attention as a direct means of assessing the faithfulness of natural language explanations (Atanasova et al., 2023). Counterfactual testing evaluates explanations by applying controlled edits to the input that induce changes in model predictions and examining whether the causal source of the change is reflected in the generated explanation. While effective as an evaluation tool, prior work has primarily treated counterfactual signals as diagnostic metrics and has not fully explored their potential as supervision for improving explanation generation itself.

In parallel, a growing body of work has investigated LLM steering, which aims to control model behavior by directly manipulating internal representations. Steering methods typically compute direction vectors that correspond to specific behaviors or traits and inject them into the model's hidden states during inference (Chen et al., 2025). Among these approaches, CAA has demonstrated that high-level model behaviors can be effectively modulated by adding contrastive activation differences—derived from paired positive and negative prompts—at the residual stream level, without modifying model parameters (Rimsky et al., 2024).

Our work extends CAA-based activation-level steering to the domain of natural language explanation generation. While prior CAA studies have focused on general behavioral alignment objectives such as honesty, sycophancy, refusal behavior, or improvements in plausibility and fluency, we instead target explanation faithfulness using data in which unfaithfulness is explicitly defined through counterfactual reasoning. By leveraging counterfactual-based unfaithful explanations as a

contrastive signal, we propose a steering framework that aligns generated explanations more directly with the model's true causal reasoning. In this sense, our work introduces a task-specific application of activation-level steering that is explicitly tailored to improving explanation faithfulness.

3 Methodology

We propose a data-centric framework to improve the faithfulness of NLE generated by models in a NLI setting. Our approach consists of three stages: (1) automatic collection of unfaithful explanations using counterfactual tests, (2) data refinement via attribute-based semantic filtering, and (3) alignment of the explanation generation process through activation-level LLM steering based on CAA.

3.1 Counterfactual Editing with the Adding Modifier (AM) Rule

Our base data source is the e-SNLI dataset, where each sample is defined as a tuple

$$(p, h, y, e), \tag{1}$$

with premise p, hypothesis h, NLI label y, and natural language explanation e.

To generate counterfactual inputs, we apply the Adding Modifier (AM) rule (Varshney et al., 2022), which inserts a single adjective W into the hypothesis. Specifically, W is inserted immediately before an existing noun in the hypothesis, while preserving all other tokens and the original sentence structure. The resulting counterfactual hypothesis h' is defined as:

$$h' = \text{Insert}(h, W). \tag{2}$$

Here, W is restricted to a single-token adjective, and no other words are added, removed, or reordered. This constraint ensures that the counterfactual edit is local and interpretable, allowing the inserted adjective to serve as a clear causal factor for potential prediction changes.

Given the counterfactually edited pair (p, h'), we prompt a LLM to regenerate both a new NLI label y' and a corresponding natural language explanation e'. These regenerated outputs are then used in subsequent steps to determine whether the explanation faithfully reflects the causal effect introduced by the counterfactual edit.

3.2 Label-change-based Causality Check

To ensure that a counterfactual edit introduces a genuine causal effect on the model's decision, we first verify whether the predicted NLI label changes after inserting the modifier W. Let $\hat{y}$ and $\hat{y}'$ denote the predicted labels for the original and edited input pairs, respectively. We regard the counterfactual edit as causally valid only if the following condition holds:

$$\hat{y}' \neq \hat{y} \tag{3}$$

This criterion ensures that the inserted modifier W has a non-trivial impact on the model's inference outcome. Only samples satisfying Equation 3 are retained as candidates for further faithfulness analysis, since label-invariant edits do not provide meaningful counterfactual supervision signals.

3.3 First-stage Faithfulness Filtering: Lexical Exclusion Test

Among the causally valid counterfactual samples, we next identify *unfaithful* natural language explanations by checking whether the explanation explicitly mentions the inserted modifier W. If the explanation fails to reference W, we consider it a candidate unfaithful explanation.

Formally, let $\hat{c}$ denote the generated natural language explanation. A sample is marked as unfaithful in the first filtering stage if:

$$W \notin \hat{c} \tag{4}$$

Applying this simple lexical exclusion criterion to 54,921 counterfactual candidates, which were randomly sampled as approximately 10% of the 550k instances in the e-SNLI training set, results in 1,450 unfaithful explanation candidates. This outcome demonstrates that even a lightweight word-level check can effectively surface a substantial number of explanations that ignore the true cause of the prediction change.

3.4 Attribute-based Semantic Filtering

In the first filtering stage, an explanation was preliminarily classified as *unfaithful* if the inserted adjective W did not explicitly appear in the generated NLE. However, relying solely on surface-level word matching can lead to false positives. For example:

- W = *small* → explanation: "its size is not mentioned"

- W = *red* → explanation: "its color is unclear"

In these cases, although the adjective W itself does not occur in the explanation, the semantic *attribute* introduced by W (e.g., size, color, material) is still explicitly discussed. Such explanations should be considered *faithful*, as they reflect the causal factor underlying the label change.

To address this limitation, we introduce an attribute-based semantic filtering step using GPT-4o-mini as a scalable proxy for human semantic judgment. We emphasize that this component is used solely for data refinement rather than evaluation, and does not directly influence the steering or decoding process. Given the inserted adjective W and an explanation $\hat{e}$, the model determines whether the explanation is semantically grounded in the attribute introduced by W, even if W itself is not mentioned. We define the attribute-level relevance function as follows:

$$\mathrm{Rel}(W, \hat{e}) = \begin{cases} 1, & \text{if attribute-relevant} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Here, $\mathrm{Rel}(W, \hat{e}) = 1$ indicates that the explanation refers to the attribute of W through direct mention, synonyms, antonyms, abstract properties, or statements of irrelevance (e.g "size does not matter").

An explanation is finally classified as *unfaithful* only if it satisfies both conditions: (i) the inserted adjective W is not explicitly mentioned, and (ii) the explanation does not semantically ground the attribute introduced by W. Formally, the final unfaithfulness criterion is defined as:

$$\mathrm{Rel}(W, \hat{e}) = 0 \ \wedge \ W \notin \hat{e} \tag{6}$$

This semantic filtering step removes false positives caused by purely lexical matching and ensures that the resulting dataset consists of high-confidence unfaithful explanations, where the causal factor of the label change is entirely ignored in the explanation.

3.5 CAA-based Faithfulness Steering

We adopt an activation-level steering strategy based on Contrastive Activation Addition (CAA) to encourage the model to generate explanations that faithfully reflect the true inference evidence. The key idea is to identify a representation direction that separates faithful explanations from unfaithful ones and to steer the model's internal activations along this direction during generation.

To this end, we collect two sets of explanations: a *faithful* set (from e-SNLI) and an *unfaithful* set (constructed via counterfactual filtering). For each sample, we extract the hidden states corresponding to the output tokens of the predicted label and the natural language explanation.

We define $\mu^{(\ell)}_{\text{faithful}}$ and $\mu^{(\ell)}_{\text{unfaithful}}$ as the mean hidden representations of faithful and unfaithful explanations at layer ℓ, respectively. The faithfulness steering vector is then defined as the contrastive difference between these two means:

$$v^{(\ell)} = \mu^{(\ell)}_{\text{faithful}} - \mu^{(\ell)}_{\text{unfaithful}}. \tag{7}$$

During generation, we apply an additive adjustment to the hidden state at layer ℓ:

$$h^{(\ell)\prime} = h^{(\ell)} + \mathit{coef} \cdot v^{(\ell)} \tag{8}$$

Here, ℓ denotes the steering layer and *coef* controls the steering strength. Based on validation experiments, we select the steering layer ℓ separately for each experimental setting and fix the steering coefficient to $\mathit{coef} = 1$ in all reported results. Detailed ablation results for layer and coefficient selection are provided in Appendix B. By steering the model's internal representations toward the faithful direction, this approach aligns explanation generation with the causal evidence underlying the model's predictions, without modifying model parameters or degrading inference performance.

3.6 Extension to Multiple Counterfactual Rules and Generability Analysis

So far, our methodology has focused on the Adding Modifier (AM) rule to construct counterfactual examples and collect unfaithful natural language explanations. In this section, we extend the scope of our analysis to examine whether the proposed counterfactual-based framework generalizes beyond a single editing rule, and whether unfaithful explanations can be automatically generated under a broader set of counterfactual transformations.

Specifically, we consider a total of 19 counterfactual rules drawn from two prior lines of work:

- 15 rules, including AM, proposed in Varshney et al. (2022)
- 4 rules (CA, CV, EI, VS) introduced in Han et al. (2025)

Each rule aims to induce a change in the NLI label by counterfactually editing the hypothesis while keeping the premise fixed. Following the same protocol as the AM-based setup, we randomly sampled approximately 10% of the e-SNLI training set (54,921 instances) and provided rule-specific prompts to a Qwen-family LLM to automatically generate counterfactual hypotheses.

This multi-rule setting allows us to analyze the generability and stability of counterfactual edits across heterogeneous transformation types, and to assess which rules yield reliable label changes and meaningful faithfulness signals for explanation alignment.

4 Experiments and Results

4.1 Experimental Setup

To evaluate the faithfulness of natural language explanations, we conduct experiments on the e-SNLI dataset (Camburu et al., 2018). Following the methodology described in Section 3, we apply the Adding Modifier (AM) rule to Qwen2.5-7B-Instruct (Bai et al., 2025) in order to generate counterfactual candidates.

Specifically, we randomly sample approximately 10% of the e-SNLI training set, resulting in 54,921 instances, and perform counterfactual edits by inserting a single adjective into the hypothesis while keeping the premise unchanged. For each edited $(premise, hypothesis)$ pair, the model is prompted to generate a new NLI label and a natural language explanation.

We then apply a two-stage filtering process to identify high-confidence unfaithful explanations. In the first stage, we retain only samples where the predicted NLI label changes and the inserted modifier W is not explicitly mentioned in the explanation. In the second stage, we apply attribute-based semantic filtering to remove false positives where the explanation implicitly refers to the attribute introduced by W (e.g., size, color, or material).

After applying both filtering stages, we obtain a final set of 589 high-confidence unfaithful NLE instances. Table 1 summarizes the data reduction process.

Stage	Description	Count	Remaining (%)
Stage 0	Initial random e-SNLI samples	54,921	–
Stage 1	Label-change & lexical filtering	1,450	2.64
Stage 2	Semantic attribute filtering	589	1.07

Table 1: Automatic construction of the unfaithful NLE dataset.

4.2 Faithfulness Evaluation of Natural Language Explanations

To quantitatively evaluate the faithfulness of NLEs, we employ an automatic evaluation protocol based on GPT-4o-mini. While LLM-based evaluation may introduce model-specific biases, we adopt it as a consistent and scalable proxy to compare relative faithfulness across experimental conditions. Importantly, all conditions are evaluated using the identical prompt and evaluator, ensuring that observed differences reflect steering effects rather than evaluation artifacts. Given a premise–hypothesis pair, Qwen2.5-14B-Instruct generates an NLI label along with a concise natural language explanation. The evaluator model then assesses whether the generated explanation accurately reflects the semantic factors that led to the prediction, particularly the meaning-changing elements introduced by counterfactual edits.

Faithfulness is scored on a continuous scale according to the following criteria:

- **0**: The explanation is unrelated to the input or semantically incorrect.
- **100**: The explanation accurately and explicitly reflects the true reasoning behind the prediction.
- **REFUSAL**: The model explicitly refuses to provide an answer.

All faithfulness scores are real-valued in the range $[0, 100]$. Importantly, the evaluation prompt and input format are kept *identical* across all experimental conditions to ensure comparability.

Concretely, the model is instructed to perform the NLI task by predicting a label from the given premise and hypothesis and providing a brief natural language justification. To standardize the explanation style, we provide six few-shot examples in total—two each for entailment, neutral, and contradiction. The output format is strictly constrained to two lines: the first line contains the predicted label, and the second line contains a short explanation.

Finally, each evaluation instance is presented using the same prompt structure, and the evaluator judges only the faithfulness of the generated explanation with respect to the input semantics. The

average faithfulness scores reported in Table 2 are computed based on this automated evaluation protocol.

4.3 CAA-based Steering for Improving Explanation Faithfulness

We evaluate the effectiveness of CAA as an activation-level steering method for improving the faithfulness of natural language explanations. Following the formulation in Section 3, we apply CAA-based steering during the decoding process of Qwen2.5-14B-Instruct, without modifying model parameters.

The steering vector is computed as the difference between the mean hidden representations of faithful and unfaithful explanations, and is interpreted as a direction that increases explanation faithfulness. During generation, this vector is additively injected into the hidden states of a selected transformer layer. Based on a hyperparameter search, we fix the steering configuration to layer $= 24$ and coefficient $= 1$, which yields the most stable improvements. Detailed experimental results for different steering layers and coefficient values are provided in Appendix B.

Data splits We consider two datasets constructed via the filtering pipeline described in Section 3: (1) a first-stage filtered dataset consisting of 1,450 instances, and (2) a second-stage semantically filtered dataset consisting of 589 high-confidence unfaithful explanations. For each dataset, we exclude six examples used in few-shot prompting and split the remaining data into a 9:1 ratio. In each case, 90% of the data is used to compute the CAA steering vector, while the remaining 10% is reserved exclusively for faithfulness evaluation. Importantly, steering vectors for the **Baseline** and **Steering** conditions are computed using the first-stage filtered dataset, whereas steering vectors for the **Baseline + Semantic Filtering** and **Steering + Semantic Filtering** conditions are computed using the semantically filtered dataset.

Evaluation conditions To analyze the effects of steering and data quality, we define four experimental conditions:

- **Baseline**: No steering is applied; faithfulness is measured on the evaluation split of the first-stage filtered dataset.
- **Steering**: CAA-based steering is applied during decoding, using a steering vector computed from the training split of the first-stage filtered dataset and evaluated on the same data as the Baseline.
- **Baseline + Semantic Filtering**: Faithfulness is measured on the evaluation split of the second-stage semantically filtered dataset, without steering.
- **Steering + Semantic Filtering**: Steering is applied using a steering vector computed from the training split of the semantically filtered dataset and evaluated on the same data as **Baseline + Semantic Filtering**.

Dataset	Condition	Faithfulness	Unfaithful Reduction
Stage-1 filtered	Baseline	94.58 ± 6.61	12.55%
	Steering	95.26 ± 6.20	
Stage-2 filtered	Baseline	95.89 ± 6.55	20.68%
	Steering	96.74 ± 5.84	

Table 2: Faithfulness improvements achieved by CAA-based steering. Steering consistently improves explanation faithfulness across both filtering stages, with larger gains observed on the high-confidence semantically filtered dataset.

In all cases, the evaluation data remain identical between Baseline and Steering conditions, ensuring that any observed differences are attributable solely to the steering mechanism.

As shown in Table 2, applying CAA-based steering consistently improves average faithfulness scores across both datasets. Notably, larger gains are observed on the second-stage semantically filtered dataset, indicating that higher-quality unfaithful supervision signals lead to more effective steering directions.

These results demonstrate that adjusting internal hidden representations alone—without retraining or modifying model parameters—can reliably encourage the model to generate explanations that better reflect its true reasoning process. Overall, the combination of semantic data filtering and activation-level steering provides complementary benefits, improving explanation faithfulness from both data-centric and representation-centric perspectives.

4.4 Qualitative Analysis of Explanation Faithfulness

To illustrate how CAA-based steering improves explanation faithfulness, we present a representative qualitative example from the Adding Modi-

Premise: There are cheerleaders tossing a member in the air.

Original Hypothesis: The cheerleaders are cheering.

Edited Hypothesis(AM): The **energetic** cheerleaders are cheering.

Label Change: entailment → neutral

Original Explanation:
Cheerleaders tossing a member in the air means that they are cheering.

Unfaithful Explanation (before steering):
Tossing and cheering are different actions.

Steered Explanation (after steering):
Their energy level isn't stated, even if they are cheering while tossing someone.

Figure 1: An AM-based counterfactual example

fier (AM) setting. This example demonstrates how steering aligns the generated explanation with the attribute introduced by the counterfactual edit, without altering the predicted NLI label.

In this example, the counterfactual edit introduces the adjective *energetic*, which changes the hypothesis from being entailed to neutral. However, the unfaithful explanation ignores this attribute and instead provides a generic distinction between actions, failing to justify the label change. After applying activation-level steering, the model generates an explanation that explicitly acknowledges the missing information about energy level, directly grounding the neutral prediction in the introduced attribute. This demonstrates that steering improves explanation faithfulness by aligning explanations with the true causal factor of the prediction change, rather than altering the prediction itself.

4.5 Multi-rule Counterfactual Unfaithful NLE Construction and Faithfulness Evaluation

In this section, we examine whether the proposed counterfactual-based approach generalizes beyond a single rule such as AM, and whether explanation faithfulness can be consistently improved under diverse counterfactual transformation settings. This analysis is critical for distinguishing whether the observed faithfulness gains arise from rule-specific heuristics or from a more general alignment signal based on causal attribution.

To this end, we conduct multi-rule experiments using only counterfactual rules that satisfy two criteria: (i) stable automatic generation via prompting, and (ii) clear causal relationships between edits and NLI label changes. Among 19 candidate rules, we exclude those that tend to preserve semantic meaning and thus do not induce label changes (e.g., PS, HS, PA, ES, SOS, IrH), as well as rules that require multi-token or phrase-level restructuring and exhibit unstable generation behavior (e.g., SSNCV, Con, CT).

As a result, we select 10 counterfactual rules that involve localized edits and yield clear causal effects on predictions: *NS, AM, CW-adj, CV, NI, CW-noun, RG, VS, CA, EI*. Detailed descriptions and examples of each rule are provided in Appendix A.

Using these rules, we randomly sample approximately 10% of the e-SNLI training set (about 54,000 instances) and generate counterfactual samples with a Qwen-family LLM. For each instance, the premise is kept fixed while the hypothesis is edited according to a given rule. The edited $(premise, hypothesis)$ pair is then used to prompt the model to generate a new NLI label and a natural language explanation.

We apply the same three-stage filtering pipeline described in Section 3: (1) label-change-based causality filtering, (2) exclusion of explanations that do not reference the edited element, and (3) attribute-based semantic filtering. Through this process, we obtain 10,536 high-confidence unfaithful NLE instances. All of these instances are used exclusively for steering vector computation, while evaluation is conducted on independent test sets.

Evaluation datasets Faithfulness is evaluated on two datasets: (i) **original 1k**, consisting of 1,000 randomly sampled instances from the original e-SNLI test set, preserving the original data distribution; and (ii) **10-per-rule 100**, a counterfactual evaluation set constructed by generating 10 counterfactually edited (p, h') pairs per rule from the e-SNLI test set, totaling 100 instances. These (p, h') pairs are used as inputs to generate natural language explanations during evaluation, including unfaithful explanations.

Table 3 reports the faithfulness evaluation results using steering vectors computed on the multi-rule unfaithful dataset. Hyperparameter search results for the steering layer and coefficient are provided in Appendix B. As shown in Table 3, applying steering improves the average explanation faithfulness on both evaluation sets. Notably, a larger gain is observed on the counterfactually generated test set,

Evaluation Set	Baseline	Steering	Diff.	Unfaithful Reduction
Original 1k	97.53 ($\pm$6.22)	97.73 ($\pm$5.60)	+0.20	8.1%
10-per-rule 100	96.70 ($\pm$5.82)	97.16 ($\pm$5.54)	+0.46	13.9%

Table 3: Faithfulness evaluation results using steering vectors computed on multi-rule counterfactual unfaithful data. Steering improves faithfulness on both original and counterfactual test sets, with larger gains observed under counterfactual perturbations.

Evaluation Set	Baseline	Steering
testset 1k	89.20	89.10
testset 10-per-rule 100	57	57

Table 4: NLI accuracy before and after applying steering vectors computed from Qwen-based counterfactual data.

suggesting that the proposed steering method more effectively mitigates explanation mismatches that arise under counterfactual perturbations.

These results indicate that our approach does not rely on rule-specific heuristics, but instead leverages a general counterfactual faithfulness signal—whether the cause of a label change is reflected in the explanation—across diverse transformation types. This demonstrates that counterfactual-based unfaithfulness serves as a robust and transferable alignment signal for improving explanation faithfulness.

4.6 Analysis of Steering Effects on NLI Performance

When improving the faithfulness of natural language explanations, it is essential to ensure that the model's core inference performance is not degraded, particularly in our setting where activation-level steering modifies internal representations during decoding. To assess this potential trade-off, we evaluate NLI prediction accuracy under the same experimental settings used in Sections 3.6 and 4.5.

Specifically, we compare NLI accuracy before and after applying steering vectors computed from Qwen-based counterfactual datasets. As shown in Table 4, NLI accuracy on the original distribution evaluation set (testset 1k) remains nearly unchanged after steering, indicating that the proposed method preserves the model's core inference behavior while selectively aligning explanation generation.

On the counterfactually generated evaluation set (testset 10-per-rule 100), NLI accuracy is substantially lower than on the original distribution. However, this degradation is consistent across both Baseline and Steering conditions, indicating that it is not a side effect of steering. Instead, it reflects the distribution shift introduced by rule-based counterfactual generation, which increases semantic ambiguity and blurs the boundary between neutral and contradiction labels.

Taken together, these results show that activation-level steering improves explanation faithfulness without trading off NLI prediction performance. Importantly, steering operates by aligning explanations with existing decision pathways, rather than modifying the model's decision boundaries.

5 Conclusion

In this work, we address the problem of improving the faithfulness of NLE in the context of NLI. We propose a data-centric framework that automatically constructs high-confidence *unfaithful* explanation datasets using counterfactual tests, together with an activation-level steering method based on CAA.

Specifically, we generate counterfactual samples by locally editing hypotheses in e-SNLI and identify cases where prediction labels change but the causal editing factors are not reflected in the generated explanations. To reduce false positives arising from surface-level lexical matching, we introduce attribute-based semantic filtering, which ensures that only explanations that completely ignore the causal attribute introduced by counterfactual edits are retained. Using the resulting unfaithful explanations, we compute steering vectors that are applied to hidden representations during decoding, enabling selective alignment of the explanation generation process without modifying model parameters or prediction outcomes.

Experimental results show that the proposed approach consistently improves explanation faithfulness not only under a single counterfactual rule but also across multiple counterfactual rules. Importantly, NLI prediction accuracy on the original in-distribution evaluation set remains largely unchanged, demonstrating that faithfulness gains are achieved without degrading core inference performance. These findings indicate that counterfactual-based faithfulness signals can serve as an effective and generalizable alignment mechanism for improving the causal consistency between model predictions and natural language explanations.

Limitations

Despite the effectiveness of the proposed counterfactual-based unfaithful explanation construction and activation-level steering approach, our work has several limitations. First, the counterfactual rules used in this study are restricted to *local edits* that can be reliably generated via prompt-based methods. More complex transformations involving phrase-level or multi-token edits, such as subject–object swaps or aggregation-based substitutions, were excluded due to current generation stability constraints, rather than conceptual limitations of the proposed framework. Future work may address this limitation by incorporating syntactic editing frameworks or constraint-based decoding to enable more structured and diverse counterfactual transformations. Second, the identification of unfaithful explanations relies on LLM-based semantic judgments over attribute, which may still introduce residual false positives. To further improve reliability, future work could adopt hybrid verification strategies that combine multiple evaluation models or integrate limited human annotation for validation. Third, our experiments focus exclusively on the NLI setting using e-SNLI. While counterfactual-based faithfulness signals are conceptually task-agnostic, their effectiveness and generalizability to other reasoning tasks remain to be empirically validated. Finally, the effectiveness of activation-level steering depends on specific choices of the steering layer and scaling coefficient. Although we observe stable improvements under selected hyperparameter settings, optimal configurations may vary across model architectures. Automated or adaptive hyperparameter selection for steering thus remains an important direction for future research.

References

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: natural language inference with natural language explanations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9560–9572, Red Hook, NY, USA. Curran Associates Inc.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*.

Juyoung Han, Hyunsun Hwang, and Changki Lee. 2025. Rule discovery for natural language inference data generation using out-of-distribution detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25971–25991, Suzhou, China. Association for Computational Linguistics.

Xuanli He, Yuxiang Wu, Oana-Maria Camburu, Pasquale Minervini, and Pontus Stenetorp. 2024. Using natural language explanations to improve robustness of in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13477–13499, Bangkok, Thailand. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.

Sawan Kumar and Partha Talukdar. 2020. Nile: Natural language inference with faithful natural language explanations. *arXiv preprint arXiv:2005.12116*.

David R Large, Leigh Clark, Annie Quandt, Gary Burnett, and Lee Skrypchuk. 2017. Steering the conversation: A linguistic exploration of natural language interactions with a digital assistant during simulated driving. *Applied ergonomics*, 63:53–61.

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

Danda B. Rawat. 2016. Roar: An architecture for real-time opportunistic spectrum access in cloud-assisted cognitive radio networks. In *2016 13th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, page 936–941. IEEE Press.

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

Neeraj Varshney, Pratyay Banerjee, Tejas Gokhale, and Chitta Baral. 2022. Unsupervised natural language inference using PHL triplet generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2003–2016, Dublin, Ireland. Association for Computational Linguistics.

Wei Jie Yeo, Ranjan Satapathy, and Erik Cambria. 2025. Towards faithful natural language explanations: A study using activation patching in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10436–10458.

A Definitions of Counterfactual NLI Generation Rules

This appendix provides formal definitions of the counterfactual hypothesis generation rules used in our multi-rule experiments (Section 3.6 and Section 4.4). Each rule specifies a localized transformation applied to the hypothesis while keeping the premise fixed, and is designed to induce a predictable change in the NLI label. Table 5 summarizes the rules grouped by their typical target label (entailment, neutral, or contradiction), along with brief descriptions of the corresponding hypothesis modifications.

B Ablation Studies on Steering Layer and Coefficient

The steering vectors used in our experiments are computed following the Contrastive Activation Addition (CAA) formulation, as described in Section 3. To determine an effective and stable steering configuration, we conduct a limited hyperparameter search over the choice of steering layer ℓ and steering coefficient *coef*. This appendix presents detailed ablation results for these two hyperparameters under two experimental settings: (i) steering based on unfaithful explanations constructed using the Adding Modifier (AM) rule only, and (ii) steering based on unfaithful explanations constructed using multiple counterfactual rules.

B.1 Ablation Results under the Adding Modifier (AM) Rule

This section reports ablation results for different steering layers and coefficient values when the steering vectors are computed using unfaithful explanations constructed solely with the Adding Modifier (AM) rule. Following the setup described in Section 4.3, we use the first-stage(Table 6) and second-stage(Table 7) filtered AM datasets to compute steering vectors, and evaluate explanation faithfulness on the corresponding evaluation splits. Layer $\ell = 24$ achieves the highest average faithfulness score in both the Stage-1 and Stage-2 filtered settings when the steering coefficient is fixed to $coef = 1$. Based on this consistent performance across datasets, we adopt $\ell = 24$ as the steering layer in our main AM-based experiments reported in Table 2.

B.2 Ablation Results under Multi-rule Counterfactual Steering

This section presents ablation results for steering layer and coefficient selection when steering vectors are computed using unfaithful explanations constructed from multiple counterfactual rules. As described in Section 4.5, we use a dataset of 10,536 high-confidence unfaithful explanations generated from ten counterfactual rules to compute steering vectors, and evaluate faithfulness on independent evaluation sets.

We further analyze the sensitivity of steering performance to the choice of steering layer under the multi-rule setting, evaluating both explanation faithfulness and NLI prediction accuracy. Results are reported on two evaluation sets: the original testset 1k(Table 8) and the counterfactually generated testset 10-per-rule 100(Table 9).

On the original testset 1k, steering at layer $\ell = 37$ yields the highest average faithfulness score while preserving NLI prediction accuracy. On the counterfactually generated testset, the strongest faithfulness improvement is observed at layer $\ell = 33$. Across both evaluation settings, steering does not substantially degrade NLI accuracy, indicating that activation-level steering selectively improves explanation faithfulness without altering the underlying inference behavior.

B.3 Ablation Study on Steering Coefficient

In addition to the steering layer, we analyze the effect of the steering coefficient *coef*, which controls the strength of the activation-level intervention. Following the multi-rule experimental setup described in Section 4.5, we vary the coefficient value from 0.5 to 2.5 while fixing the steering layer to $\ell = 33$, which yields the strongest faithfulness im-

Label	Rule-Name	Explanation
Entailment	Role Generalization (RG)	Generate a hypothesis by replacing specific roles with general categories.
Neutral	Adding Modifiers (AM)	Generate a hypothesis by adding modifiers to nouns.
	Visual Specification (VS)	Generate a hypothesis by adding visual characteristics.
	Contextual Augmentation (CA)	Generate a hypothesis by adding implicit purposes or background.
	Emotion Inference (EI)	Generate a hypothesis by inferring emotions or states from actions.
Contradiction	Number Substitution (NS)	Generate a hypothesis by replacing numbers with different numbers.
	Contradictory Words-adj (CW-adj)	Generate a hypothesis by replacing adjectives with their antonyms.
	Contradictory Verb (CV)	Generate a hypothesis by replacing verbs with their antonyms.
	Negation Introduction (NI)	Generate a hypothesis by introducing negation.
	Contradictory Words-noun (CW-noun)	Generate a hypothesis by replacing nouns with contradictory nouns.

Table 5: Definitions of Counterfactual NLI Generation Rules

Coef	Layer	Faithfulness
1	5	94.62 ± 6.47
	10	94.79 ± 6.23
	15	95.00 ± 6.31
	20	94.30 ± 6.81
	24	**95.26 ± 6.20**
	26	94.58 ± 6.76
	28	94.39 ± 6.90
	30	94.49 ± 7.36
	32	94.12 ± 7.85
	34	94.82 ± 6.97
	36	95.21 ± 6.47
	38	94.80 ± 6.48
	40	95.20 ± 5.79
	45	95.03 ± 5.68

Table 6: Ablation results over steering layers (ℓ) on the Stage-1 filtered dataset under the Adding Modifier (AM) rule with $coef = 1$. Layer 24 achieves the highest average faithfulness score.

provement on the counterfactually generated evaluation set. As shown in Table 10, a coefficient of $coef = 1$ achieves the highest explanation faithfulness while maintaining stable NLI prediction accuracy. Based on this result, we adopt $coef = 1$ in all multi-rule steering experiments.

Coef	Layer	Faithfulness
1	5	96.52 ± 5.88
	10	96.25 ± 5.90
	15	96.42 ± 5.77
	20	96.26 ± 5.83
	24	**96.74 ± 5.84**
	26	96.35 ± 5.90
	28	95.97 ± 7.25
	30	96.16 ± 7.22
	32	96.64 ± 7.22
	34	96.43 ± 7.09
	36	96.56 ± 6.96
	38	96.30 ± 7.12
	40	95.35 ± 7.59
	45	96.03 ± 6.42

Table 7: Ablation results over steering layers (ℓ) on the Stage-2 semantically filtered dataset under the Adding Modifier (AM) rule with $coef = 1$. Layer 24 yields the strongest faithfulness improvement.

Coef	Layer	Baseline		Steering	
		Faithfulness	NLI Acc.	Faithfulness	NLI Acc.
1	24	97.53 ± 6.22	89.20	97.35 ± 6.08	89.00
	25			97.30 ± 6.42	89.20
	26			97.41 ± 5.72	88.90
	27			97.27 ± 6.73	89.00
	28			97.42 ± 6.12	88.90
	29			97.44 ± 6.09	89.00
	30			97.26 ± 6.69	88.40
	31			97.33 ± 6.64	87.80
	32			97.30 ± 6.15	87.90
	33			97.31 ± 6.17	88.70
	34			97.22 ± 6.59	88.80
	35			97.37 ± 6.13	88.90
	36			97.48 ± 5.72	88.90
	37			**97.73 ± 5.60**	89.10
	38			97.69 ± 5.61	89.10
	39			97.68 ± 5.63	89.10
	40			97.63 ± 5.69	89.10
	41			97.31 ± 5.82	89.10
	42			97.40 ± 5.57	89.10
	43			97.32 ± 5.81	89.10
	44			97.33 ± 5.86	89.30
	45			97.46 ± 5.44	89.00

Table 8: Ablation results over steering layers under the multi-rule setting on the original testset 1k with $coef = 1$. The highest faithfulness score is achieved at layer 37, while NLI accuracy remains largely stable.

Coef	Layer	Baseline		Steering	
		Faithfulness	NLI Acc.	Faithfulness	NLI Acc.
1	24	96.70 ± 5.82	57	96.38 ± 7.35	57
	25			96.30 ± 7.27	57
	26			96.60 ± 5.82	58
	27			95.69 ± 8.64	56
	28			96.47 ± 7.19	57
	29			97.00 ± 6.00	56
	30			96.13 ± 10.21	56
	31			96.35 ± 9.71	57
	32			96.35 ± 9.27	57
	33			**97.16 ± 5.54**	57
	34			96.45 ± 9.27	57
	35			96.95 ± 5.71	57
	36			96.58 ± 6.93	57
	37			96.90 ± 5.90	57
	38			96.79 ± 5.83	57
	39			97.07 ± 5.45	58
	40			96.35 ± 6.99	56
	41			96.85 ± 5.58	57
	42			96.77 ± 5.67	56
	43			96.77 ± 5.64	59
	44			96.72 ± 5.64	56
	45			97.02 ± 5.37	57

Table 9: Ablation results over steering layers under the multi-rule setting on the counterfactually generated testset 10-per-rule 100 with $coef = 1$. Layer 33 achieves the highest faithfulness score, while NLI accuracy remains stable.

Layer	Coef	Faithfulness	NLI Acc.
33	0.5	96.71 ± 5.89	57
	1.0	**97.16 ± 5.54**	57
	1.5	96.31 ± 10.50	57
	2.0	96.03 ± 10.71	57
	2.5	95.80 ± 10.99	57

Table 10: Ablation results over steering coefficient values under the multi-rule setting on the counterfactually generated testset 10-per-rule 100. The steering layer is fixed to $\ell = 33$. A coefficient of $coef = 1$ yields the highest faithfulness score while preserving NLI accuracy.