*i*Flip: Iterative Feedback-driven Counterfactual Example Refinement

Anonymous ACL submission

Abstract

Counterfactual examples are minimal edits to an input that alter a model's prediction. They are widely employed in explainable AI to probe model behavior and in natural language processing (NLP) to augment training data. However, generating valid counterfactuals with large language models (LLMs) remains challenging, as existing single-pass methods often fail to induce reliable label changes, neglecting LLMs' self-correction capabilities. To explore this untapped potential, we propose *i*Flip[1], an iterative refinement approach that leverages three types of feedback, including model confidence, feature attribution, and natural language. Our results show that *i*Flip achieves an average 57.8% higher validity than the five state-of-the-art baselines, as measured by the label flipping rate. The user study further corroborates that *i*Flip outperforms baselines in completeness, overall satisfaction, and feasibility. In addition, ablation studies demonstrate that three components are paramount for *i*Flip to generate valid counterfactuals: leveraging an appropriate number of iterations, pointing to highly attributed words, and early stopping. Finally, counterfactuals generated by *i*Flip enable effective counterfactual data augmentation, substantially improving model performance and robustness.

1 Introduction

Counterfactual examples are minimally modified inputs that cause a model to alter its prediction (Madsen et al., 2022; Wang et al., 2024; Zhao et al., 2024). Applications include elucidating opaque LLMs through contrastive causal analysis (Ross et al., 2021; Treviso et al., 2023; Nguyen et al., 2024b) and enhancing model performance and robustness (Kaushik et al., 2020; Dixit et al., 2022; Qiu et al., 2024). However, most recent approaches (Bhan et al., 2023; Bhattacharjee et al., 2024a; Nguyen et al., 2025) rely on single-pass generation, often yielding invalid counterfactuals that are insufficient to shift the model's original prediction. This neglects LLMs' inherent self-correction capabilities, despite their proven efficacy in other domains and downstream tasks (Madaan et al., 2023; Gou et al., 2024; Rahmani et al., 2025).

To overcome these limitations, we propose *i*Flip (Figure 1), a framework that iteratively refines counterfactuals via diverse feedback signals to enhance validity. **First**, we evaluate *i*Flip across three datasets and three LLMs using confidence, feature attribution, and natural language feedback. Compared to five baselines – BAE (Garg and Ramakrishnan, 2020), Polyjuice (Wu et al., 2021), CGG (Nguyen et al., 2025), FIZLE (Bhattacharjee et al., 2024a), and Causal What-Ifs (Narendra and Chatterjee, 2025) – *i*Flip improves validity by 57.8% at the cost of 8.3% lower similarity, with natural language feedback proving most effective.

Second, a user study confirms *i*Flip with natural language feedback outperforms FIZLE and CGG in *completeness*, *overall satisfaction*, and *feasibility*, achieving the most substantial improvement in *overall satisfaction* (83.21%). Error analysis reveals that the iterative refinement process effectively addresses both incomplete and contextually inappropriate edits in counterfactuals (§6.2).

Third, we perform systematic ablation studies to assess the contribution of each component in *i*Flip: (1) the number of refinement iterations, (2) the choice of feedback signal, and (3) of early stopping criteria. We find that counterfactual validity can be noticeably improved across iterations, even within a limited number of rounds (e.g., $\mathcal{K} = 5$). Feedback-guided generation considerably boosts performance over the without-feedback baseline. Moreover, we establish that early stopping upon achieving validity is imperative to prevent valid counterfactuals from being erroneously overturned in subsequent iterations.

[1]The code is available anonymously at https://anonymous.4open.science/r/iFlip.

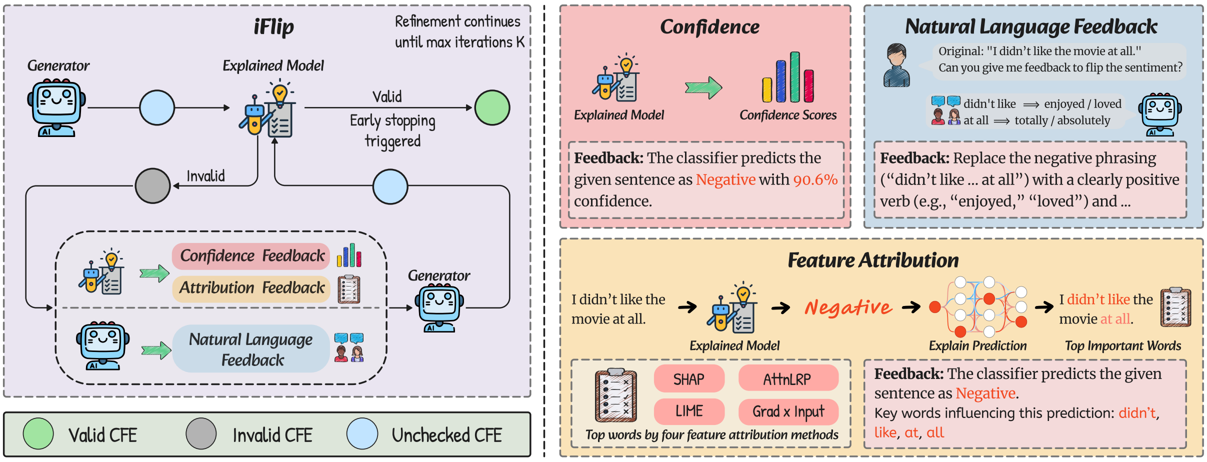

Figure 1: iFlip framework overview. ***Left***: The core iterative loop of our method, which includes a generator $\mathcal{G}$ and an explained model $\mathcal{M}$. $\mathcal{G}$ generates and refines counterfactuals. $\mathcal{M}$ evaluates counterfactuals under our validity criterion: a counterfactual is considered *valid* if $\mathcal{M}$'s predicted label is equal to the target label $\tilde{y}$. Otherwise, *invalid* counterfactuals are iteratively refined until the validity criterion is met to avoid unnecessary iterations or the maximal number of refinement steps is reached. ***Right***: Examples of the three types of feedback we employed for iFlip: *Confidence*, *Feature Attribution*, and *Natural Language Feedback*.

Lastly, we show that counterfactuals generated by iFlip enable effective counterfactual data augmentation (CDA), improving model performance and robustness, measured by accuracy on both the original and the human-annotated counterfactual test sets. Additionally, our analysis confirms that the quality of the generated counterfactuals positively correlates with the CDA performance gains.

2 Related Work

LLM-based Counterfactual Generation. FIZLE employs LLMs for zero-shot counterfactual generation to explain and evaluate black-box classifiers (Bhattacharjee et al., 2024a). Bhattacharjee et al. (2024b) propose a three-step pipeline that leverages instruction-tuned LLMs to extract latent features and their associated input words. Several recent methods incorporate feature attribution signals to guide counterfactual generation. ZeroCF improves zero-shot counterfactual generation by leveraging feature attribution methods to guide text edits (Wang et al., 2025b). Similarly, CGG inserts token attributions into few-shot prompts to guide LLMs in generating high-validity CFEs (Nguyen et al., 2025). Narendra and Chatterjee (2025) propose an agentic framework that iteratively refines counterfactuals through self-reflection but does not incorporate feedback or early stopping, which we find crucial for enhancing counterfactual validity (§6.3.2, §6.3.3). Unlike these approaches, iFlip explicitly incorporates three types of feedback, including feature attribution signals, and iteratively refines counterfactuals based on this feedback.

Iterative Refinement with Feedback. Xi et al. (2023) and Liu et al. (2024) empirically observe that model performance tends to improve over successive rounds of self-correction, wherein LLMs refine their answers using either intrinsic (Madaan et al., 2023; Xu et al., 2024) or external (Welleck et al., 2023) feedback. In the field of explainability, SR-NLE improves the faithfulness of post-hoc explanations through an iterative critique and refinement process (Wang and Atanasova, 2025). CROSS-REFINE enhances the quality of natural language explanations via tandem learning through feedback and critiques in an iterative manner (Wang et al., 2025a). In this work, we employ three types of feedback signals to guide iterative refinement of generated counterfactuals across multiple rounds.

3 Methodology

3.1 iFlip

As illustrated in Figure 1, iFlip employs iterative refinement driven by the interaction of two components: the counterfactual generator $\mathcal{G}$ and the explained model $\mathcal{M}$. The explained model validates the counterfactual candidates produced by the generator. When a candidate is unsuccessful in changing the model's prediction, it is sent back to the generator, accompanied by specific feedback signals (§3.2), to guide the next generation step.

i`Flip` is structured into three primary stages:

Step 1: Generation. Given the original input x and the explained model $\mathcal{M}$, we first obtain the original model prediction $y = \mathcal{M}(x)$, a label in a classification task, then define a target label $\tilde{y} \neq y$. Generator $\mathcal{G}$ produces an initial counterfactual candidate:

$$\tilde{x}_0 \leftarrow \mathcal{G}(x, y, \tilde{y}) \quad (1)$$

Step 2: Verification. Explained model $\mathcal{M}$ validates the candidate $\tilde{x}_0$. If $\mathcal{M}(\tilde{x}_0) = \tilde{y}$, i.e., the desired target label is achieved, we return $\tilde{x}_0$ (early stopping); otherwise, we proceed to **refinement**.

Step 3: Refinement. If the initial counterfactual candidate $\tilde{x}_0$ fails to induce the model prediction to $\tilde{y}$, we perform up to $\mathcal{K}$ iterative refinement steps, indexed by $k \in \{1, 2, \cdots, \mathcal{K}\}$, each guided by feedback signals f_k:

$$\tilde{x}_k \leftarrow \mathcal{G}(x, \tilde{x}_{k-1}, y, \tilde{y}, f_k) \quad (2)$$

Upon generating a new counterfactual candidate $\tilde{x}_k$, we return to the validation phrase (**Step 2**). This iterative process continues until either the validation succeeds ($\mathcal{M}(\tilde{x}_k) = \tilde{y}$) or the maximum iteration count $\mathcal{K}$ is attained.

3.2 Feedback Signals

In our experiments, we cover three feedback signals $F = \{f_{\text{Conf}}, f_{\text{Attr}}, f_{\text{NL}}\}$ that are prevalently employed in the literature: *confidence*, *feature attribution*, and *natural language feedback* (Figure 1).

Confidence. The self-correction capability of language models comprises two components: confidence, defined as the ability to retain correct answers, and critique, defined as the ability to revise incorrect answers (Yang et al., 2025b). Consequently, following Hernandez et al. (2025) and Mavi et al. (2025), we incorporate the confidence score as one of feedback signals.

Feature Attribution. Following prior work that established guiding LLMs for counterfactual generation using important words (Bhan et al., 2023; Wang et al., 2025b; Nguyen et al., 2025), we apply established four feature attribution methods: SHAP (Lundberg and Lee, 2017), AttnLRP (Achtibat et al., 2024), LIME (Ribeiro et al., 2016), and Gradient × Input (Shrikumar et al., 2017), which determine token importance in an input text.

Natural Language Feedback. LLM-generated feedback in free-text form has been shown to be effective for improving both downstream task performance and explanation generation (Madaan et al., 2023; Wang et al., 2025a). Accordingly, we instruct the LLMs to provide natural language feedback on their own counterfactuals.

4 Experimental Setup

4.1 Datasets & Models

Datasets. We test *i*`Flip` on three NLP datasets that are well-studied in the counterfactual generation literature.[2] IMDb comprises diverse movie reviews labeled as *positive* or *negative* sentiment (Maas et al., 2011). AG News contains news articles categorized into four topics: *World*, *Sports*, Business, and *Sci/Tech* (Zhang et al., 2015). Each article includes a title and a description. SNLI is a dataset for natural language inference (NLI) tasks (Bowman et al., 2015). It consists of premise-hypothesis pairs labeled as *entailment*, *contradiction*, or *neutral*.

Models. We employ three state-of-the-art, open-source LLMs from different model families and with increasing parameter sizes: `OLMo2-7B` (OLMo et al., 2025), `Qwen3-32B` (Yang et al., 2025a), `LLaMA3.3-70B` (Grattafiori et al., 2024).[3]

4.2 Baselines

We compare *i*`Flip` against two widely adopted baselines and three state-of-the-art LLM-based counterfactual generation methods. ① BAE is an adversarial approach and perturbs inputs using a `BERT` masked language model to replace masked tokens based on semantic similarity (Garg and Ramakrishnan, 2020). ② Polyjuice leverages a fine-tuned `GPT-2` to generate counterfactuals by framing the task as a conditional text generation problem, using control codes to produce diverse and fluent counterfactuals (Wu et al., 2021). ③ CGG uses feature attributions to select the top 25% most important words and incorporates them into a predefined prompt with factual-counterfactual examples, explicitly directing the LLM which words to modify for the target label flip (Nguyen et al., 2025). ④ FIZLE$_{\text{naive}}$ directly prompts an LLM to minimally edit the input to achieve a label flip, without prior feature identification or guidance (Bhattacharjee et al., 2024a). ⑤ Causal What-Ifs is an agentic approach that iteratively improves counterfactual generation through self-reflection, enabling the identifi-

[2] Dataset examples, label distributions, and dataset sources are provided in Appendix A.

[3] Information about inference time and generation parameters is provided in Appendix B.

cation of causally consistent edits *without external feedback* (Narendra and Chatterjee, 2025).

5 Evaluation

5.1 Automatic Evaluation

To assess the quality of generated counterfactuals, we adopt three commonly used automatic evaluation metrics following prior work (Ross et al., 2021; Wang et al., 2024; Nguyen et al., 2024a; Bhattacharjee et al., 2024a): *validity*, *similarity*, and *fluency*.

Label Flipping Rate (LFR) evaluates the *validity* of counterfactuals by measuring the proportion of instances whose predicted label changes after modification. Given the explained model $\mathcal{M}$[4], the original inputs x, and counterfactuals $\tilde{x}$, LFR is calculated by:

$$\mathrm{LFR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[\mathcal{M}(x_i) \neq \mathcal{M}(\tilde{x}_i)]$$

where $\mathbb{I}[\cdot]$ denotes the indicator function.

Semantic Similarity (SS) We compute semantic similarity between the original input x and its counterfactual $\tilde{x}$ using cosine similarity between their sentence embeddings $\mathbf{e}(\cdot)$, where each embedding is produced by a pretrained encoder[5]:

$$\mathrm{SS}(x_i, \tilde{x}_i) = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbf{e}(x_i)^\top \mathbf{e}(\tilde{x}_i)}{\|\mathbf{e}(x_i)\|_2 \, \|\mathbf{e}(\tilde{x}_i)\|_2}$$

Perplexity (PPL) To assess *fluency*, we compute the perplexity of generated counterfactuals using `GPT-2`. Given a counterfactual $\tilde{x} = (t_1, \ldots, t_n)$ consisting of n tokens and a model parameterized by θ, the perplexity of $\tilde{x}$ is defined as:

$$\mathrm{PPL}(\tilde{x}) = \exp\left(-\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(t_i \mid t_{<i})\right)$$

5.2 User Study & LLM-as-a-Judge

We further assess the quality of generated counterfactuals by conducting a *user study* and a complementary *LLM-as-a-Judge evaluation*. Participants and judge models, respectively, evaluate the counterfactuals along three dimensions (§5.2.1) on a 6-point Likert scale.

5.2.1 Subjective Ratings

Following Nguyen et al. (2024a); Domnich et al. (2025), we ask human annotators to evaluate counterfactuals across three subjective dimensions:

- **Completeness**: The counterfactual sufficiently explains the model's decision.
- **Overall Satisfaction**: The counterfactual effectively shows how to reach a different outcome.
- **Feasibility**: The counterfactual provides actions that are practical, realistic, and actionable.

5.2.2 User Study Setup

We conduct a user study on AG News, from which we randomly sample 10 examples, with ($N = 3$) participants, all proficient in English.[6] We evaluate two LLM-based baselines (FIZLE and CGG) alongside *i*`Flip` configured with natural language feedback, which demonstrates the best performance among all feedback types (§6.1). In Table 2, We report inter-annotator and inter-judge-model agreements with Krippendorff's α for each dimension.

5.2.3 LLM-as-a-Judge Setup

LLM-as-a-Judge (LaaJ) has become a widely adopted approach for conducting evaluations by assigning quality scores that align with human judgment (Huang et al., 2024; Li et al., 2025). Following Gu et al. (2025), we employ three commonly used open-source LLMs of varying sizes: `Gemma3-27B` (Team et al., 2025), `GPT-OSS-120B` (OpenAI et al., 2025), and `DeepSeek-R1` (DeepSeek-AI et al., 2025). Judge models are instructed to assess the quality of generated counterfactuals based on three subjective dimensions, as in the user study (§5.2.2).[7]

5.3 Ablation Study

As outlined in §3.1, our framework incorporates three essential components: (1) the number of refinement iteration steps, (2) feedback signals, and (3) early stopping. We conduct ablation studies to assess the contribution of each component.

The Number of Refinement Iterations. We aim to assess the contribution of iterative refinement to counterfactual generation effectiveness and quantify its advantages over single-pass methods. Our analysis adopts the pass@k metric (Brown et al., 2024), which measures the fraction of counterfactuals achieving successful label flips within k refinement attempts. This metric tests whether multiple refinement steps considerably enhance the LFR.

Feedback Signal. We analyze the role of different feedback signals in counterfactual generation by

[4]Model $\mathcal{M}$ is `BERT` in Table 1 and `RoBERTa` in App. C.1.
[5]`sentence-transformers/all-MiniLM-L6-v2`

[6]The annotation guideline is detailed in Appendix L.
[7]The full judge prompt is provided in Appendix M.1.

	Method	IMDb			AG News			SNLI (Premise)			SNLI (Hypothesis)		
		LFR↑	SS↑	PPL↓	LFR↑	SS↑	PPL↓	LFR↑	SS↑	PPL↓	LFR↑	SS↑	PPL↓
	Polyjuice	0.275	0.736	60.81	0.148	0.744	93.22	0.281	0.772	96.95	0.350	0.733	114.59
	BAE	0.553	0.976	63.15	0.227	0.764	295.96	0.135	0.966	82.98	0.115	0.979	81.00
OLMo2-7B	CGG	0.885	0.837	44.67	0.510	0.380	208.56	0.169	0.882	57.03	0.253	0.948	53.92
	FIZLE	0.585	0.584	38.71	0.308	0.584	77.38	0.305	0.730	173.94	0.404	0.882	74.15
	Causal What-Ifs	0.890	0.801	37.97	0.400	0.607	37.54	0.413	0.854	47.27	0.487	0.911	35.56
	*i*Flip-Conf	0.964	0.809	41.19	0.784	0.503	**41.41**	**0.660**	0.854	**53.72**	**0.826**	0.870	40.65
	*i*Flip-SHAP	0.968	0.805	41.87	**0.802**	0.488	42.95	0.587	0.847	56.81	0.760	0.867	42.39
	*i*Flip-AttnLRP	0.978	0.803	42.61	0.768	0.500	46.61	0.596	0.851	55.30	0.810	0.869	**40.31**
	*i*Flip-Grad×Input	0.976	0.801	43.30	0.773	0.498	50.32	0.624	0.854	56.02	0.782	0.871	41.81
	*i*Flip-LIME	0.968	0.805	52.82	0.730	0.504	44.32	0.558	0.866	58.18	0.736	0.875	41.02
	*i*Flip-NL	**0.984**	**0.822**	**36.47**	0.735	**0.515**	44.73	0.638	**0.879**	148.41	0.770	**0.901**	43.81
Qwen3-32B	CGG	0.863	0.877	62.47	0.364	0.697	97.10	0.200	0.881	50.31	0.283	0.938	47.94
	FIZLE	0.684	0.900	36.00	0.164	0.864	61.46	0.454	0.901	38.66	0.524	0.918	35.97
	Causal What-Ifs	0.848	0.821	42.03	0.436	0.570	53.50	0.482	0.842	36.14	0.518	0.920	34.18
	*i*Flip-Conf	0.996	0.855	33.30	0.726	0.521	**41.59**	0.466	**0.895**	62.69	0.562	0.871	47.03
	*i*Flip-SHAP	0.994	0.857	33.70	0.736	0.519	44.02	0.452	0.878	58.60	0.586	0.866	45.04
	*i*Flip-AttnLRP	**1.000**	0.858	33.19	**0.746**	0.513	44.15	0.448	0.882	**58.41**	0.610	0.867	44.77
	*i*Flip-Grad×Input	0.996	0.865	33.65	0.730	0.522	44.33	0.440	0.881	58.66	0.542	0.865	45.60
	*i*Flip-LIME	0.998	0.863	**32.57**	0.742	0.527	44.52	0.458	0.891	62.98	0.570	0.870	**44.22**
	*i*Flip-NL	0.980	**0.878**	32.99	0.732	**0.536**	49.68	**0.596**	**0.895**	68.16	**0.754**	**0.906**	47.20
LLaMA3.3-70B	CGG	0.887	0.869	52.81	0.526	0.694	122.40	0.307	0.846	51.62	0.271	0.937	46.58
	FIZLE	0.924	0.868	34.88	0.416	0.707	56.59	0.402	0.903	41.54	0.534	0.929	42.39
	Causal What-Ifs	0.922	0.788	31.93	0.636	0.461	43.56	0.372	0.816	38.97	0.420	0.904	41.29
	*i*Flip-Conf	**0.996**	0.877	34.74	0.890	0.468	44.38	**0.580**	0.763	58.65	0.642	0.843	43.88
	*i*Flip-SHAP	**0.996**	0.878	**34.32**	0.860	0.481	46.07	0.522	0.743	57.62	0.606	0.828	41.09
	*i*Flip-AttnLRP	0.992	0.879	34.58	0.858	0.477	45.78	0.528	0.743	56.42	0.602	0.822	39.92
	*i*Flip-Grad×Input	**0.996**	0.878	34.80	0.852	0.479	46.12	**0.580**	0.751	**55.38**	0.603	0.825	**39.59**
	*i*Flip-LIME	**0.996**	0.877	34.54	0.852	0.475	**44.20**	0.558	0.759	57.87	0.608	0.825	42.91
	*i*Flip-NL	**0.996**	**0.895**	36.90	**0.900**	**0.527**	63.86	0.566	**0.821**	69.45	**0.680**	**0.861**	69.45

Table 1: Automatic evaluation results of counterfactuals generated by baselines (Polyjuice, BAE, CGG, and FIZLE) and our *i*Flip methods on the BERT models with different feedback: ❶ confidence (Conf), ❷ feature attribution (SHAP, AttnLRP, Grad×Input, LIME), ❸ natural language (NL). Results are reported on IMDb, AG News, and SNLI using Label Flipping Rate (LFR), Semantic Similarity (SS), and Perplexity (PPL). **Boldface** indicates the best feedback type within *i*Flip. Wavy underline indicates the best result across methods.

considering three feedback settings: (*i*) the baseline *no feedback*, where the model refines without any external signal; (*ii*) a *random feedback*, where the important features are randomly selected; (*iii*) *least-attributed feedback*, where feedback is derived from the least important words, identified by the corresponding attribution methods.

Without Early Stopping. We investigate the impact of early stopping on the counterfactual refinement process. Early stopping halts generation upon a successful label flip; without it, the model continues to edit valid counterfactuals until $\mathcal{K}$ iterations. The core objective is to determine if eliminating early stopping can maintain the achieved label flip while simultaneously improving the overall quality of the generated counterfactuals.

5.4 Counterfactual Data Augmentation

We examine to what extent counterfactuals generated by *i*Flip enhance both model performance and robustness through counterfactual data augmentation (Kaushik et al., 2020; Dixit et al., 2022; Wang et al., 2025c). Our baseline is a model $\mathcal{M}_{\text{base}}$ which is trained only on the original dataset $\mathcal{D}_{\text{base}} = \{(x_i, y_i)\}_{i=1}^{N}$, while the CDA-enhanced model $\mathcal{M}_c$ is trained on both the original data and the corresponding counterfactuals (either generated by *i*Flip or human-annotated) $\mathcal{D}_c = \{(x_i, y_i), (\tilde{x}_i, \hat{y}_i)\}_{i=1}^{N}$. The CDA evaluation involves assessment using both the test set data and human-annotated counterfactuals.

6 Results

6.1 Automatic Evaluation

Comparison among the baselines. Table 1 shows that perturbation-based baselines (Polyjuice, BAE) tend to make fewer edits but often struggle with the model prediction flipping, whereas LLM-based baselines (CGG, FIZLE and Causal What-Ifs) can more reliably generate counterfactuals achieving the target labels while being more natural. Among the LLM-based baselines, CGG, which leverages feature importance methods to identify the most important words, demonstrates higher LFR on IMDb and AG News than FIZLE, where such word-level guidance facilitates more effective label flips. In contrast, FIZLE, which generates counterfactuals through direct prompting, performs better on SNLI, likely because for NLI tasks, modifying important words identified by feature importance methods often fails to fully capture the underlying logical relations between the premise and the hypothesis. Notably, Causal What-Ifs attains the highest validity among baselines on NLI tasks, though it struggles to outperform attribution-guided methods (CGG) on news topic classification (AG News). However, Causal What-Ifs requires noticeably more edits compared to FIZLE and CGG, suggesting a trade-off between validity gains and input modifications.

***i*`Flip` outperforms all baselines in terms of LFR across datasets and models.** Table 1 illustrates that *i*`Flip` consistently achieves the highest LFR across all datasets and models while maintaining competitive fluency.[8] Averaged across all experimental settings, *i*`Flip` improves LFR by 71.4% and fluency by 13.8% relative to single-pass LLM-based baselines. These gains are particularly pronounced on AG News, where *i*`Flip` achieves a 107% higher LFR than those baselines. Compared with self-reflection-based Causal What-Ifs, *i*`Flip` further improves validity by 35.7% on average. This additional gain further validates the effectiveness of the feedback mechanism (§6.3.2) and the early stopping strategy (§6.3.3).

***i*`Flip` gains involve a trade-off between validity and similarity.** While *i*`Flip` improves the validity of generated counterfactuals, this comes at the cost of more extensive edits (Table 1)[9], indicating a trade-off between validity and the extent of input modifications. Specifically, `OLMo2-7B`, despite being the smallest model employed, attains comparable or even higher LFR than larger models by incorporating more edits (Table 1). Particularly, on AG News, topic transitions typically require more intensive multi-entity and event-level modifications. Counterintuitively, we find that counterfactuals generated by larger LLMs do not necessarily achieve higher LFR, consistent with prior observations (Nguyen et al., 2024b; Wang et al., 2025b).

Comparison of different feedback signals. As shown in Figure 6, natural language feedback generally achieves the most (consistent) enhancement across all experimental setups in terms of LFR. However, the improvement comes at the cost of substantially higher inference time due to the LLM-generated natural language feedback (Madaan et al., 2023) (Table 7). Confidence-based feedback ranks second, achieving a balance between effectiveness and computational efficiency. In contrast, attribution-based feedback generally performs poorly, likely because LLMs neither strictly adhere to the identified important words nor limit themselves to targeted modifications, instead often rephrasing entire sentences.[10] Among attribution approaches, propagation- and gradient-based methods (AttnLRP and Grad×Input) consistently outperform perturbation-based ones (SHAP and LIME), as further detailed in Appendix E.2.

Case Study. We examine SNLI more closely, where generating valid counterfactuals is particularly challenging due to the inherent logical relationships between premises and hypotheses, making it an outlier relative to IMDb and AG News. Following prior work on counterfactual generation for NLI (Nguyen et al., 2024b, 2025), we generate counterfactuals by modifying either the premise or hypothesis. For SNLI, editing hypotheses is substantially more effective, as their shorter length allows *i*`Flip` to induce label flips more easily. Empirically, however, we observe that modifying important words is less effective in improving LFR compared to other ways of generating counterfactuals. Among *i*`Flip` variants, attribution-based feedback yields lower LFR than confidence-based or natural language feedback (Figure 7,Appendix D), suggesting that targeting only salient words is insufficient to reverse entailment relations in NLI tasks. Moreover, SNLI exhibits both substantially lower feature attribution faithfulness than IMDb and AG News, and weaker alignment between faithfulness and counterfactual quality.[11]

6.2 User Study & LLM-as-a-Judge

User Study. Table 2a **quantitatively** confirms that *i*`Flip`-NL consistently outperforms CGG and FIZLE across all three subjective evaluation di-

[8] App. C.2 reports the transferability of generated counterfactuals across architectures and the transferability of feedback extracted from `BERT`.

[9] See App. E.1.3 for a visualization of SS-LFR trade-off. App. E.1.1 and E.1.2 report the average number of refinements and the average length of counterfactuals at each iteration.

[10] Representative examples are presented in Appendix J.

[11] App. G reports the faithfulness of the employed attribution methods, while App. H.2 details the correlation between the feature attribution faithfulness and counterfactual quality.

Method	Complet.	Overall Sat.	Feasib.
CGG	2.83 ± 1.69	2.67 ± 1.49	2.80 ± 1.60
FIZLE$_{naive}$	3.27 ± 2.00	2.93 ± 1.71	3.40 ± 1.80
*i*Flip-NL	**5.10** $\pm$ **1.19**	**5.13** $\pm$ **1.12**	**4.83** $\pm$ **1.27**
Krippendorff's α	0.6467	0.8690	0.8694

(a) Human evaluation.

Method	Complet.	Overall Sat.	Feasib.
CGG	3.20 ± 1.94	3.10 ± 2.06	3.93 ± 2.03
FIZLE$_{naive}$	3.10 ± 2.12	3.07 ± 2.13	3.80 ± 2.17
*i*Flip-NL	**4.53** $\pm$ **1.78**	**4.47** $\pm$ **1.86**	**4.40** $\pm$ **1.80**
Krippendorff's α	0.7615	0.8215	0.4005

(b) LLM-as-a-judge evaluation.

Table 2: Evaluation results (**mean** $\pm$ **std**) on three subjective dimensions: *Completeness*, *Overall Satisfaction* and *Feasibility*. Best results are **bolded**.

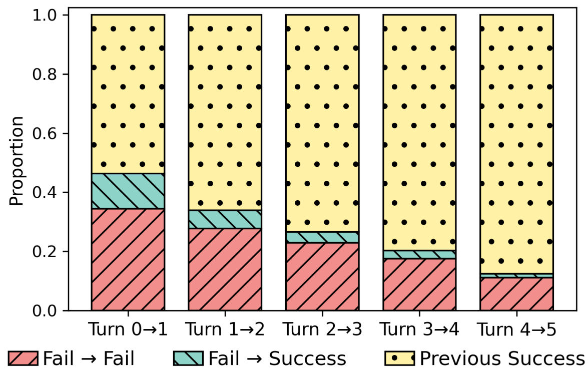

Figure 3: Iterative refinement with OLMo2-7B under early stopping, averaged across all datasets and feedback signals. *Fail* $\rightarrow$ *Fail* indicates no label change, *Fail* $\rightarrow$ *Success* denotes a successful label flip in the current turn, and *Previous Success* marks instances flipped in earlier turns.

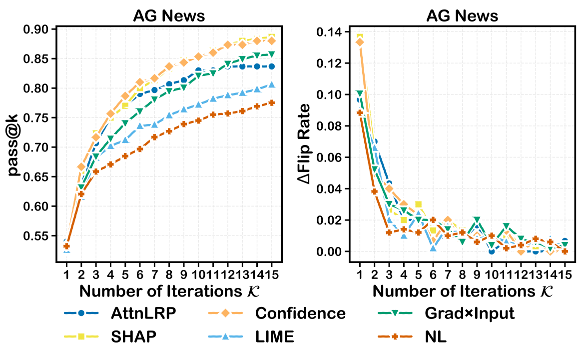

Figure 2: Iterative refinement effectiveness on OLMo2-7B (AG News) across feedback signals. Left: pass@k curves up to $\mathcal{K} = 15$. Right: per-iteration improvement in flip rate (ΔFlip Rate).

mensions (§5.2.1), achieving an average relative gain of 68.27%. Notably, the most substantial improvement occurs in *Overall Satisfaction*, with an 83.21% relative gain. To understand the source of these improvements, we **qualitatively** analyze representative failure cases in Appendix N. We observe two recurring issues in baseline methods: (i) they frequently edit only partial entity mentions, resulting in *incomplete* changes that fail to flip the model prediction; and (ii) they sometimes produce contextually inappropriate edits, yielding *unnatural* counterfactuals. In contrast, *i*Flip-NL mitigates these issues through iterative refinement, improving both *Completeness* and *Feasibility* of generated counterfactuals.[12]

LLM-as-a-Judge. Table 2b reveals that LaaJ results exhibit a trend consistent with human evaluation: *i*Flip-NL invariably outperforms selected baselines across all three dimensions.[13] Notably, the LLM judgments demonstrate larger standard deviations than human ratings, indicating that LLMs tend to assign more extreme scores. Moreover, the human-LLM alignment, as measured by Krippendorff's α 0.72, indicates relatively strong agreement between human and LaaJ evaluation.

6.3 Ablation Study

6.3.1 Iterative Refinement

Figure 2 illustrates that pass@k increases steadily with the number of refinement iterations, though the improvement per round gradually decreases as refinement progresses.[14] This trend demonstrates LLMs' feedback-driven self-correction capacity while revealing single-pass generation's substantial performance gap relative to iterative refinement, especially when initial counterfactuals exhibit low validity (e.g., SNLI premise editing achieves less than 30% LFR in the first iteration). Even with a small number of iterations ($\mathcal{K} = 5$), iterative refinement yields markedly higher-validity counterfactuals than single-pass generation. Figure 3 further provides these dynamics under early stopping, where *Fail* $\rightarrow$ *Fail* transitions diminish and *Previous Success* cases dominate over time.[15]

6.3.2 Feedback Signal

We examine whether incorporating feedback is intrinsically beneficial for iterative counterfactual re-

[12] See Appendix N.3 and N.4 for per-annotator human/LLM ratings and examples showing how iterative refinement fixes *incomplete* and *unnatural* edits

[13] Evaluation across judge models are reported in App. M.2.

[14] Full pass@k curves and flip-rate improvements with OLMo2-7B across all datasets are provided in Appendix E.1.4.

[15] *Success* $\rightarrow$ *Fail* are not observed due to early stopping.

Feedback of *i*Flip	IMDb	AG News	SNLI (Premise)	SNLI (Hypothesis)
Without Feedback	0.964	0.742	0.508	0.778
Random Feedback	0.970(↑0.006)	0.794(↑0.052)	0.672(↑0.164)	0.764(↓0.014)
Least SHAP	0.970(↑0.006)	0.770(↑0.028)	0.612(↑0.104)	0.780(↑0.002)
Least AttnLRP	0.980(↑0.016)	0.776(↑0.034)	0.562(↑0.054)	0.764(↓0.014)
Least Grad×Input	0.968(↑0.004)	0.758(↑0.016)	0.593(↑0.085)	0.760(↓0.018)
Least LIME	0.968(↑0.004)	0.725(↓0.017)	0.556(↑0.048)	0.754(↓0.024)
Conf	0.964(=)	0.784(↑0.042)	0.660(↑0.152)	0.826(↑0.048)
SHAP	0.968(↑0.004)	0.802(↑0.060)	0.587(↑0.079)	0.760(↓0.018)
AttnLRP	0.978(↑0.014)	0.768(↑0.026)	0.596(↑0.088)	0.810(↑0.032)
Grad×Input	0.976(↑0.012)	0.773(↑0.031)	0.624(↑0.116)	0.782(↑0.004)
LIME	0.968(↑0.004)	0.730(↓0.012)	0.558(↑0.050)	0.736(↓0.042)
NL	0.984(↑0.020)	0.735(↓0.007)	0.638(↑0.130)	0.770(↓0.008)

Table 3: Automatic evaluation results in terms of LFR of `OLMo2-7B`. The tables present an ablation study of different feedback settings, including *Without Feedback*, *Random Feedback*, *Least-Attributed Feedback*, and the full variants of *i*`Flip`. Colored values indicate relative changes compared to the *Without Feedback* baseline (↑ = increase, ↓ = decrease).

finement. Table 3 illustrates that incorporating feedback generally improves LFR compared to the *without feedback* baseline (Appendix E.2). However, the contribution of feedback signals is not uniform; certain signals may exhibit differential utility, potentially introducing misguidance into the editing process (§6.1). We further assess the effectiveness of *i*`Flip` with *random feedback*, which offers only modest improvements relative to *without feedback* in terms of LFR. Among attribution-based signals, directing edits towards the most important words consistently outperforms targeting the least important words, yielding a mean LFR gain of 1.09% alongside an 11.02% reduction in edits, indicating more effective yet less disruptive edits.

6.3.3 Early Stopping

Tables 17 and 18 show that the removal of early stopping induces a decline in LFR, which is attributed to valid counterfactuals being overturned in successive iterations, as illustrated by the proportion of *(Success → Fail)* cases (Figure 13).[16] This demonstrates the crucial role of early stopping in maintaining reliable counterfactual generation. Conversely, although additional iterations enable LLMs to enhance text fluency, they typically involve more extensive modifications. Our extended analysis in Appendix C.2 further reveals that early stopping is the primary driver of cross-architecture transferability, as it prevents the generator from overfitting to the specific decision boundary of the explained model.

[16]Detailed examples appear in Appendix J.3 and J.4.

Method	IMDb	CFs	AG News	SNLI (P)	CFs	SNLI (H)	CFs
Baseline	77.93	50.09	66.87	40.87	33.43	40.87	33.85
Human	**92.73**	**91.26**	–	**42.27**	**50.90**	**58.40**	**56.95**
Conf	87.93	71.09	78.60	38.47	**32.98**	**44.13**	**36.88**
SHAP	91.33	90.30	80.27	37.40	27.82	40.33	35.83
AttenLRP	89.27	90.28	77.87	38.13	27.97	40.13	35.53
Grad×Input	**93.07**	90.07	79.07	39.00	26.98	39.40	35.98
LIME	91.40	**90.91**	78.80	38.33	28.93	38.93	35.74
NL	90.73	71.97	**81.80**	**40.40**	30.91	37.47	33.79

Table 4: Accuracy (%) on test sets and human-annotated counterfactuals (CFs). Counterfactuals used for CDA are generated by `OLMo2-7B`. Best results are **bolded**.

6.4 Counterfactual Data Augmentation

Although human-annotated counterfactuals provide marginally stronger performance gains than *i*`Flip`-generated counterfactuals (Table 4), their prohibitive cost and time requirements motivate automated approaches. Noticeably, models trained with *i*`Flip`-based counterfactuals achieve substantial performance and robustness improvements over the baseline, except on SNLI (Premise). This exception may be attributed to imperfect counterfactuals, as indicated by the relatively lower LFR compared to other datasets (Table 1), which introduce ambiguous and noisy labels that adversely affect model robustness (Zhu et al., 2022; Wang et al., 2025c).[17] We further observe a moderate correlation between CDA performance and counterfactual quality (Appendix H.1).

7 Conclusion

In this work, we propose *i*`Flip`, a feedback-driven framework that iteratively refines counterfactuals by incorporating multiple feedback signals to enhance their validity. Empirical results demonstrate that *i*`Flip` consistently surpasses state-of-the-art baselines, with natural language feedback yielding the strongest and most consistent improvements. In addition, a user study shows that *i*`Flip`-NL exceeds baseline ratings across all subjective dimensions and that its iterative refinement mechanism substantially improves the completeness and realism of generated counterfactuals. Moreover, our ablation studies reveal three key components of *i*`Flip`: iterative refinement, feedback signals, and early stopping, with early stopping emerging as the simplest and most cost-effective. Finally, we extend the study to counterfactual data augmentation and show that counterfactuals generated by *i*`Flip` effectively enhance model performance and robustness.

[17]Additional OOD augmentation results are provided in Appendix F. For the News Classification task, human-annotated counterfactuals are not available.

Limitations

In this work, we propose a feedback-driven framework `iFlip` for iterative counterfactual refinement. While `iFlip` noticeably improves the validity of generated counterfactuals compared to five selected baselines, we acknowledge several limitations.

Usage of limited Feature Attribution Methods. We do not exhaustively explore all feature attribution-based feedback signals; investigating additional feature attribution methods remains future work.

Computational Costs. Despite the effectiveness of `iFlip` in producing high-quality counterfactuals, it incurs relatively high computational costs, particularly when using natural language feedback, as the iterative refinement process requires multiple rounds of generation and validation compared to single-pass methods.

English-Centric Evaluation. In addition, all experiments in this work were carried out exclusively on English datasets, and it remains unclear how well the approach would generalize to other languages. We plan to extend our evaluation to a multilingual setting in future work to assess the generalizability of our findings across languages.

Limited Scale of the User Study. Following prior work (Nguyen et al., 2024a; McAleese and Keane, 2024), we predominantly assess the quality of generated counterfactuals by three widely used automatic metrics regarding *validity*, *fluency*, and *minimality*, given that user studies for counterfactuals are uncommon in the literature. Nevertheless, to further assess textual quality beyond automatic metrics, we conduct a small-scale user study with $(N = 3)$ participants evaluating *completeness*, *understandability*, and *cohesiveness*, following Domnich et al. (2025). As future work, we consider to include a comprehensive and large-scale user study to assess textual quality, e.g., coherence, feasibility, or complexity (Domnich et al., 2025; Wang et al., 2025d), and the effectiveness of counterfactuals for model predictions.

Ethics Statement

The participants in our user studies were compensated at or above the minimum wage in accordance with the standards of our host institutions' regions. The annotation took each annotator 30 minutes on average.

References

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. 2024. AttnLRP: Attention-aware layer-wise relevance propagation for transformers. In *Forty-first International Conference on Machine Learning*.

Giuseppe Attanasio, Eliana Pastor, Chiara Di Bonaventura, and Debora Nozza. 2023. ferret: a framework for benchmarking explainers on transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, page 256–266. Association for Computational Linguistics.

Milan Bhan, Jean-Noël Vittaut, Nicolas Chesneau, and Marie-Jeanne Lesot. 2023. Tigtec: Token importance guided text counterfactuals. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, pages 496–512, Cham. Springer Nature Switzerland.

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2024a. Zero-shot LLM-guided Counterfactual Generation: A Case Study on NLP Model Evaluation . In *2024 IEEE International Conference on Big Data (BigData)*, pages 1243–1248, Los Alamitos, CA, USA. IEEE Computer Society.

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2024b. Towards llm-guided causal explainability for black-box text classifiers. In *AAAI 2024 Workshop on Responsible Language Models, Vancouver, BC, Canada*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *Preprint*, arXiv:2407.21787.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Tanay Dixit, Bhargavi Paranjape, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. CORE: A retrieve-then-edit framework for counterfactual data generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2964–2984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marharyta Domnich, Julius Välja, Rasmus Moorits Veski, Giacomo Magnifico, Kadi Tulver, Eduard Barbu, and Raul Vicente. 2025. Towards unifying evaluation of counterfactual explanations: Leveraging large language models for human-centric assessments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(15):16308–16316.

Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Abner Hernandez, Tomás Arias Vergara, Andreas Maier, and Paula Andrea Pérez-Toro. 2025. Confidence-guided error correction for disordered speech recognition. *Preprint*, arXiv:2509.25048.

Fan Huang, Haewoon Kwak, Kunwoo Park, and Jisun An. 2024. ChatGPT rates natural language explanation quality like humans: But on which scales? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3111–3132, Torino, Italia. ELRA and ICCL.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

Guangliang Liu, Haitao Mao, Bochuan Cao, Zhiyu Xue, Xitong Zhang, Rongrong Wang, Jiliang Tang, and Kristen Johnson. 2024. On the intrinsic self-correction capability of llms: Uncertainty and latent concept. *Preprint*, arXiv:2406.02378.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in nlp: A survey. *Computational Linguistics*, 50(2):657–723.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural nlp: A survey. *ACM Comput. Surv.*, 55(8).

Vaibhav Mavi, Shubh Jaroria, and Weiqi Sun. 2025. Self-evaluating LLMs for multi-step tasks: Stepwise confidence estimation for failure detection. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.

Stephen McAleese and Mark Keane. 2024. A comparative analysis of counterfactual explanation methods for text classifiers. *Preprint*, arXiv:2411.02643.

Yash Narendra and Niladri Chatterjee. 2025. Causal what-ifs: Rethinking counterfactuals with llm agents. In *Neural Information Processing*, pages 184–198, Singapore. Springer Nature Singapore.

Van Bach Nguyen, Christin Seifert, and Jörg Schlötterer. 2024a. CEval: A benchmark for evaluating counterfactual text generation. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 55–69, Tokyo, Japan. Association for Computational Linguistics.

Van Bach Nguyen, Christin Seifert, and Jörg Schlötterer. 2025. Guiding llms to generate high-fidelity and high-quality counterfactual explanations for text classification. In *Explainable Artificial Intelligence*, pages 158–176, Cham. Springer Nature Switzerland.

Van Bach Nguyen, Paul Youssef, Christin Seifert, and Jörg Schlötterer. 2024b. LLMs for generating and evaluating counterfactuals: A comprehensive study. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14809–14824, Miami, Florida, USA. Association for Computational Linguistics.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Xiaoqi Qiu, Yongjie Wang, Xu Guo, Zhiwei Zeng, Yu Yue, Yuhong Feng, and Chunyan Miao. 2024. PairCFR: Enhancing model training on paired counterfactually augmented data through contrastive learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11955–11971, Bangkok, Thailand. Association for Computational Linguistics.

Hossein A. Rahmani, Satyapriya Krishna, Xi Wang, Mohammadmehdi Naghiaei, and Emine Yilmaz. 2025. Self-correcting large language models: Generation vs. multiple choice. *Preprint*, arXiv:2511.09381.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Alexis Ross, Ana Marasović, and Matthew Peters. 2021. Explaining NLP models via minimal contrastive editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3840–3852, Online. Association for Computational Linguistics.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2017. Not just a black box: Learning important features through propagating activation differences. *Preprint*, arXiv:1605.01713.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Marcos Treviso, Alexis Ross, Nuno M. Guerreiro, and André Martins. 2023. CREST: A joint framework for rationalization and counterfactual text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15109–15126, Toronto, Canada. Association for Computational Linguistics.

Qianli Wang, Tatiana Anikina, Nils Feldhus, Simon Ostermann, Sebastian Möller, and Vera Schmitt. 2025a. Cross-refine: Improving natural language explanation generation by learning in tandem. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1150–1167, Abu Dhabi, UAE. Association for Computational Linguistics.

Qianli Wang, Nils Feldhus, Simon Ostermann, Luis Felipe Villa-Arenas, Sebastian Möller, and Vera Schmitt. 2025b. FitCF: A framework for automatic feature importance-guided counterfactual example generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1176–1191, Vienna, Austria. Association for Computational Linguistics.

Qianli Wang, Van Bach Nguyen, Nils Feldhus, Luis Felipe Villa-Arenas, Christin Seifert, Sebastian Möller, and Vera Schmitt. 2025c. Truth or twist? optimal model selection for reliable label flipping evaluation in llm-based counterfactuals. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 80–97, Hanoi, Vietnam. Association for Computational Linguistics.

Qianli Wang, Mingyang Wang, Nils Feldhus, Simon Ostermann, Yuan Cao, Hinrich Schütze, Sebastian Möller, and Vera Schmitt. 2025d. Through a compressed lens: Investigating the impact of quantization on llm explainability and interpretability. *Preprint*, arXiv:2505.13963.

Yingming Wang and Pepa Atanasova. 2025. Self-critique and refinement for faithful natural language explanations. In *Proceedings of the 2025 Conference* *on Empirical Methods in Natural Language Processing*, pages 8492–8518, Suzhou, China. Association for Computational Linguistics.

Yongjie Wang, Xiaoqi Qiu, Yu Yue, Xu Guo, Zhiwei Zeng, Yuhong Feng, and Zhiqi Shen. 2024. A survey on natural language counterfactual generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4798–4818, Miami, Florida, USA. Association for Computational Linguistics.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2023. Generating sequences by learning to self-correct. In *The Eleventh International Conference on Learning Representations*.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online. Association for Computational Linguistics.

Zhiheng Xi, Senjie Jin, Yuhao Zhou, Rui Zheng, Songyang Gao, Jia Liu, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. Self-Polish: Enhance reasoning in large language models via problem refinement. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11383–11406, Singapore. Association for Computational Linguistics.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. 2024. SaySelf: Teaching LLMs to express confidence with self-reflective rationales. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5985–5998, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Zhe Yang, Yichang Zhang, Yudong Wang, Ziyao Xu, Junyang Lin, and Zhifang Sui. 2025b. Confidence v.s. critique: A decomposition of self-correction capability for LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3998–4014, Vienna, Austria. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2).

Dawei Zhu, Michael A. Hedderich, Fangzhou Zhai, David Ifeoluwa Adelani, and Dietrich Klakow. 2022. Is BERT robust to label noise? a study on learning with noisy labels in text classification. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 62–67, Dublin, Ireland. Association for Computational Linguistics.

A Dataset

A.1 Dataset Example

Figure 4 presents a representative example and its corresponding gold label for each dataset.

Figure 4: Examples from IMDb, AG News, and SNLI datasets.

A.2 Label Distributions

Figure 5 summarizes the label distributions across three datasets.

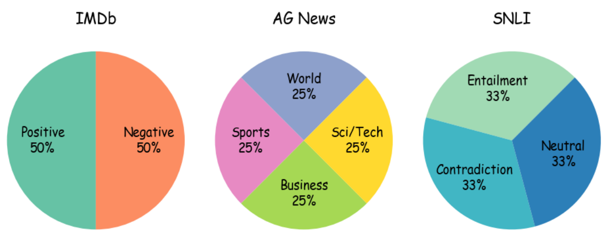

Figure 5: Label distributions across datasets.

A.3 Dataset Sources

Table 5 illustrates the sources of the datasets used in our experiments.

B Experimental Settings

B.1 Experimental Parameters

Table 6 summarizes the key parameters used in our experiments. These values were kept consistent across datasets unless stated otherwise.

B.2 Inference Time

Table 7 reports the inference time required for the `OLMo2-7B` generator model across all tasks.

B.3 Generator Models

We employ three LLMs as generator models to produce counterfactuals, representing diverse scales, parameter sizes, and model families. Table 8 summarizes their details.

B.4 Explained Models

Table 9 summarizes the `BERT` and `RoBERTa` models fine-tuned on the selected datasets. These models serve as the classifier that determines whether the counterfactuals truly change the previous model prediction.

C Extended Evaluation Results

C.1 Evaluation Results on `RoBERTa-base`

Table 10 presents the automatic evaluation results for counterfactuals generated by *i*`Flip` on the `RoBERTa-base` model, using `LLaMA3.3-70B` as the generator. While agentic baselines like Causal What-Ifs remain competitive on NLI tasks, *i*`Flip` demonstrates superior consistency, particularly on sentiment analysis and news topic classification where it frequently outperforms baselines.

C.2 Transferability

C.2.1 Transferability of the counterfactuals across architectures

Table 11 reports the *transferability* of the generated counterfactuals across classifier architectures. Specifically, all counterfactuals are generated for `BERT` classifiers, while the automatic evaluation is conducted by replacing the original `BERT` classifier with a `RoBERTa` classifier.

C.2.2 Transferability of the feedback extracted from `BERT`

Table 12 evaluates the *transferability* of feedback signals extracted from `BERT`. We use these signals to guide *i*`Flip` in generating counterfactuals, and then evaluate the resulting counterfactuals with `RoBERTa`.

C.2.3 Discussion

Early stopping is the key factor for transferability. Regarding transferability, early stopping emerges as the primary driver of robustness across architectures. In Table 11, when counterfactuals generated for `BERT` are evaluated on `RoBERTa`, methods employing early stopping (e.g., `iFlip` and Causal What-Ifs) exhibit substantial validity reductions compared to counterfactuals generated directly for `RoBERTa` (Table 10). This pattern indicates that early stopping enhances transferability by providing the LLM with direct validity feedback on whether candidate counterfactuals successfully flip classifier predictions.

The source of the feedback has limited impact on transferability. Table 12 suggests that the *source* of the feedback (extracted from `BERT` or `RoBERTa`) is comparatively less influential on transferability: using feedback from different models yields only small performance differences, and the resulting counterfactuals achieve similar quality to those guided by `RoBERTa` feedback (Table 10). Nevertheless, feedback is still beneficial for counterfactual generation, where our ablation results showing that removing feedback leads to worse performance than using feedback (see §6.3.2).

C.2.4 Summary of experimental settings

Table 13 summarizes the experimental settings used in Tables 10, 11, and 12, including (i) the *feedback source* model, (ii) the model used for *early stopping* (validity check during generation), and (iii) the *evaluation* classifier.

D Feedback Signal Comparison

We compare different feedback signals in terms of their label flipping rate (LFR) across models and tasks. Figure 6 shows the results across models, and Figure 7 shows the results across tasks.

Dataset	Citation	Source
IMDb	Maas et al. (2011)	https://huggingface.co/datasets/stanfordnlp/imdb
AG News	Zhang et al. (2015)	https://huggingface.co/datasets/sentence-transformers/agnews
SNLI	Bowman et al. (2015)	https://huggingface.co/datasets/stanfordnlp/snli

Table 5: Datasets employed in our experiments with their citations and sources.

Parameter	Value
Generator Parameters	
Temperature	0.9
Top-p	0.95
Top-k	50
Max new tokens	4096
Number of iterations	5
Top-k important words	$\max(10, \lfloor 0.10 \times \lvert\text{orig\_words}\rvert \rfloor)$
Explained Model Settings	
Voting Strategy	Majority voting
Window Size	512
Stride	256
Hardware	
`OLMo2-7B`	1× V100 (32 GB)
`Qwen3-32B`	1× H100 (80 GB)
`LLaMA3.3-70B`	2× H100 (80 GB)

Table 6: Experimental settings.

Task	Method	Duration (hours)
IMDb	Conf	3.40
	AttnLRP	3.22
	SHAP	3.75
	Grad×Input	3.45
	LIME	3.72
	NL	4.55
AG News	Conf	7.30
	AttnLRP	7.22
	SHAP	9.73
	Grad×Input	7.37
	LIME	9.86
	NL	17.90
SNLI-Premise	Conf	1.39
	AttnLRP	1.64
	SHAP	2.10
	Grad×Input	1.55
	LIME	4.80
	NL	7.57
SNLI-Hypothesis	Conf	0.93
	AttnLRP	1.13
	SHAP	1.37
	Grad×Input	1.14
	LIME	3.45
	NL	5.26

Table 7: Inference time (in hours) for `OLMo2-7B` across SNLI, IMDb, and AG News.

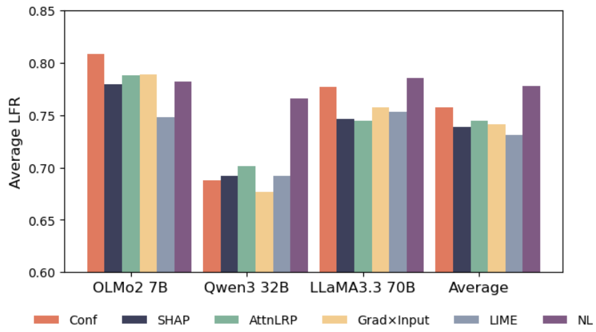

Figure 6: Averaged label flipping rate (LFR) across feedback signals and models on the IMDb, AG News, and SNLI datasets.

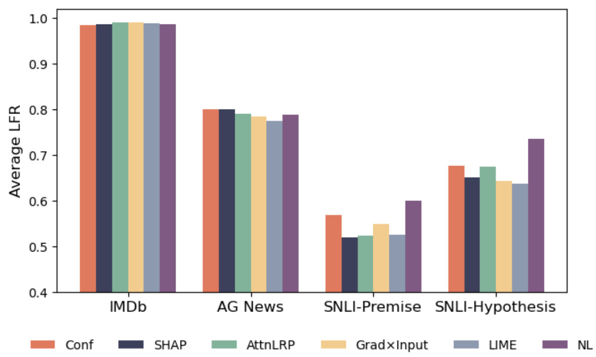

Figure 7: Average label flipping rate (LFR) across feedback signals and tasks on the IMDb, AG News, and SNLI datasets.

E Extended Ablation Results

E.1 Iterative Refinement

E.1.1 Average Early Stop Rounds

Table 14 reports the average number of refinement rounds before early stopping across datasets and models.

E.1.2 Counterfactual Length across Iteration Rounds

Figure 8 illustrates how the average length of generated counterfactuals evolves over refinement iterations.

Name	Citation	Size	Source
`OLMo2-7B`	OLMo et al. (2025)	7B	`https://huggingface.co/allenai/OLMo-2-1124-7B-Instruct`
`Qwen3-30B`	Yang et al. (2025a)	32B	`https://huggingface.co/Qwen/Qwen3-32B`
`Llama3.3-70B`	Grattafiori et al. (2024)	70B	`https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct`

Table 8: Generator models employed in our experiments.

Task	Model	Source
IMDb	`textattack/bert-base-uncased-imdb`	`https://huggingface.co/textattack/bert-base-uncased-imdb`
	`textattack/roberta-base-imdb`	`https://huggingface.co/textattack/roberta-base-imdb`
AG News	`textattack/bert-base-uncased-ag-news`	`https://huggingface.co/textattack/bert-base-uncased-ag-news`
	`textattack/roberta-base-ag-news`	`https://huggingface.co/textattack/roberta-base-ag-news`
SNLI	`textattack/bert-base-uncased-snli`	`https://huggingface.co/textattack/bert-base-uncased-snli`
	`utahnlp/snli_roberta-base_seed-2`	`https://huggingface.co/utahnlp/snli_roberta-base_seed-2`

Table 9: Explained models (BERT and RoBERTa) used for different datasets.

	Method	IMDb			AG News			SNLI (Premise)			SNLI (Hypothesis)		
		LFR↑	SS↑	PPL↓	LFR↑	SS↑	PPL↓	LFR↑	SS↑	PPL↓	LFR↑	SS↑	PPL↓
	`RoBERTa-base`												
	Polyjuice	0.274	0.736	60.81	0.165	0.744	93.22	0.212	0.772	96.95	0.260	0.733	114.59
	BAE	0.637	0.977	62.26	0.443	0.813	246.61	0.518	0.894	103.32	0.665	0.914	92.10
`LLaMA 3.3-70B`	CGG	0.930	0.892	54.57	0.748	0.542	130.31	0.635	0.893	59.23	0.680	0.944	54.23
	FIZLE	0.956	0.868	34.88	0.345	0.707	56.59	0.759	0.908	41.54	0.821	0.929	42.39
	Causal What-Ifs	0.908	0.788	32.78	0.656	0.442	46.18	0.750	0.864	41.07	0.740	0.892	43.63
	i`Flip`-Conf	**1.000**	**0.895**	37.48	0.880	0.575	**55.55**	0.665	0.688	62.94	**0.830**	0.799	47.43
	i`Flip`-SHAP	**1.000**	**0.895**	**36.19**	0.885	0.542	57.69	0.680	0.700	**57.99**	0.800	0.795	**44.69**
	i`Flip`-NL	**1.000**	0.893	36.48	**0.915**	**0.583**	57.66	**0.730**	**0.789**	77.00	0.815	**0.815**	52.91

Table 10: Automatic evaluation results of counterfactuals on the `RoBERTa-base` model using *i*`Flip` methods with different feedback types: ❶ confidence (Conf), ❷ feature attribution (SHAP), and ❸ natural language (NL). **Boldface** indicates the best feedback type within *i*`Flip`. Wavy underline indicates the best result across methods.

E.1.3 Trade-off between Validity and Similarity

Figure 9 plots the trade-off between semantic similarity (SS) and label flipping rate (LFR). Figure 10 further breaks down this trade-off by showing how the SS-LFR trajectory evolves across *i*`Flip`-NL iteration steps k on each task. After $\mathcal{K} = 4$, the degradation in semantic similarity accelerates, so for practical applications requiring high fidelity, $\mathcal{K} = 3$ or $\mathcal{K} = 4$ may offer a better balance.

E.1.4 Iterative Refinement Effectiveness

Figures 11 and 12 show that iterative refinement with `OLMo2-7B` yields steady pass@k gains across IMDb, AG News, and SNLI up to $k = 15$, while the largest marginal improvements in flip rate occur in the early rounds and taper off thereafter depending on the feedback signal.

E.2 Feedback Signals

Table 15 reports the evaluation results of our feedback signal ablations on IMDb and AG News, while Table 16 presents the corresponding results on SNLI.

E.3 Without Early Stopping

Table 17 reports the evaluation results of our without early stopping ablations on IMDb and AG News, while Table 18 presents the corresponding results on SNLI. Figure 13 presents the distribution of state transitions during iterative refinement when the early-stopping mechanism is disabled.

F Counterfactual Data Augmentation

Tables 19 and 20 report the results of data augmentation on in-domain (ID), out-of-domain (OOD), and human-annotated counterfactual (Human CFs)

	Method	IMDb			AG News			SNLI (Premise)			SNLI (Hypothesis)		
		LFR↑	SS↑	PPL↓	LFR↑	SS↑	PPL↓	LFR↑	SS↑	PPL↓	LFR↑	SS↑	PPL↓
	Polyjuice	0.274	0.736	60.81	0.165	0.744	93.22	0.212	0.772	96.95	0.260	0.733	114.59
	BAE	0.090	0.976	63.15	0.157	0.764	295.96	0.068	0.966	82.98	0.095	0.979	81.00
OLMo2-7B	CGG	0.891	0.837	44.67	0.560	0.380	208.56	0.257	0.882	57.03	0.323	0.948	53.92
	FIZLE	0.585	0.584	38.71	0.236	0.584	77.38	0.586	0.730	173.94	0.634	0.882	74.15
	Causal What-Ifs	0.850	0.801	37.97	0.273	0.607	37.54	0.443	0.854	47.27	0.627	0.911	35.56
	*i*Flip-Conf	0.895	0.809	41.19	**0.552**	0.504	**41.41**	0.440	0.854	53.52	0.620	0.870	40.65
	*i*Flip-SHAP	0.906	0.805	41.87	**0.552**	0.489	42.95	0.441	0.847	52.76	0.594	0.867	42.38
	*i*Flip-AttnLRP	0.923	0.803	42.61	0.536	0.501	46.61	**0.458**	0.851	**51.89**	0.624	0.869	**40.31**
	*i*Flip-Grad×Input	0.912	0.801	43.30	0.516	0.499	50.32	0.432	0.854	52.93	0.634	0.871	41.81
	*i*Flip-LIME	0.911	0.805	52.82	0.522	0.506	44.32	0.420	0.866	52.33	0.606	0.875	41.02
	*i*Flip-NL	**0.931**	**0.822**	**36.47**	0.532	**0.516**	44.73	0.444	**0.879**	148.41	**0.636**	**0.901**	43.81
Qwen3-32B	CGG	0.890	0.877	62.47	0.364	0.697	97.10	0.273	0.881	50.31	0.343	0.938	47.94
	FIZLE	0.686	0.900	36.00	0.140	0.864	61.46	0.689	0.901	38.66	0.810	0.918	35.97
	Causal What-Ifs	0.798	0.821	42.03	0.436	0.570	53.50	0.580	0.842	36.14	0.680	0.920	34.18
	*i*Flip-Conf	**0.940**	0.855	33.30	0.622	0.521	**41.59**	0.490	**0.895**	50.44	0.640	0.871	47.03
	*i*Flip-SHAP	0.936	0.857	33.70	0.604	0.520	44.02	0.542	0.878	47.16	0.634	0.866	45.04
	*i*Flip-AttnLRP	0.934	0.858	33.19	**0.626**	0.513	44.15	0.532	0.882	**46.93**	0.648	0.867	44.77
	*i*Flip-Grad×Input	0.934	0.865	33.65	0.592	0.522	44.33	0.528	0.881	47.40	0.658	0.865	45.60
	*i*Flip-LIME	0.938	0.863	**32.57**	0.624	0.527	44.52	0.498	0.891	50.60	0.654	0.870	**44.22**
	*i*Flip-NL	0.922	**0.878**	32.99	0.618	**0.537**	49.68	**0.546**	**0.895**	68.16	**0.668**	**0.906**	47.20
LLaMA3.3-70B	CGG	0.851	0.869	52.81	0.516	0.694	122.40	0.357	0.846	51.62	0.412	0.937	46.58
	FIZLE	0.956	0.868	34.88	0.345	0.707	56.59	0.759	0.903	41.54	0.821	0.929	42.39
	Causal What-Ifs	0.852	0.788	31.93	0.628	0.461	43.56	0.528	0.816	38.97	0.560	0.904	41.29
	*i*Flip-Conf	0.944	0.877	34.74	0.804	0.468	44.38	0.654	0.763	45.99	0.678	0.843	43.88
	*i*Flip-SHAP	0.944	0.878	**34.32**	0.794	0.481	46.07	0.652	0.743	45.15	0.678	0.828	41.09
	*i*Flip-AttnLRP	**0.946**	0.879	34.58	**0.816**	0.477	45.78	0.642	0.743	44.66	0.684	0.822	39.92
	*i*Flip-Grad×Input	0.942	0.878	34.80	0.814	0.479	46.12	**0.666**	0.751	**44.14**	0.696	0.825	**39.55**
	*i*Flip-LIME	0.944	0.877	34.54	0.796	0.476	**44.20**	0.654	0.759	45.51	0.680	0.825	42.91
	*i*Flip-NL	0.942	**0.895**	36.90	0.782	**0.527**	63.86	0.624	**0.821**	69.45	**0.726**	**0.861**	49.74

Table 11: Automatic evaluation results of counterfactual generated for BERT models, and evaluated on RoBERTa classifiers. Baselines (Polyjuice, BAE, CGG, and FIZLE) are compared with our *i*Flip variants under different feedback types: ❶ confidence (Conf), ❷ feature attribution (SHAP, AttnLRP, Grad×Input, LIME), and ❸ natural language (NL). Results are reported on IMDb, AG News, and SNLI using Label Flipping Rate (LFR), Semantic Similarity (SS), and Perplexity (PPL). **Boldface** indicates the best feedback type within *i*Flip. Wavy underline indicates the best result across methods.

	Method	IMDb			AG News			SNLI (Premise)			SNLI (Hypothesis)		
		LFR↑	SS↑	PPL↓	LFR↑	SS↑	PPL↓	LFR↑	SS↑	PPL↓	LFR↑	SS↑	PPL↓
LLaMA 3.3-70B	*i*Flip-Conf	0.995	0.897	37.17	0.896	**0.539**	58.65	0.672	0.702	62.81	0.816	0.792	46.84
	*i*Flip-SHAP	**1.000**	**0.897**	37.12	0.905	0.516	57.88	**0.677**	0.697	62.30	0.831	0.790	**44.08**
	*i*Flip-AttnLRP	**1.000**	0.895	**36.66**	0.876	0.524	59.95	**0.677**	**0.706**	56.80	0.841	**0.796**	45.66
	*i*Flip-Grad×Input	**1.000**	**0.897**	36.78	**0.910**	0.518	59.34	0.672	0.699	**55.40**	0.811	0.792	44.98
	*i*Flip-LIME	**1.000**	**0.898**	36.92	0.905	0.509	**57.58**	0.662	0.704	55.88	**0.846**	0.788	45.36

Table 12: Automatic evaluation results of counterfactual generated by *i*Flip using feedback extracted from BERT models, and evaluated on RoBERTa models. We report results on IMDb, AG News, and SNLI using Label Flipping Rate (LFR), Semantic Similarity (SS), and Perplexity (PPL). Feedback types include ❶ confidence (Conf), ❷ feature attribution (SHAP, AttnLRP, Grad×Input and LIME). **Boldface** indicates the best feedback type within *i*Flip.

test data across four tasks: Sentiment Analysis, News Classification, NLI Premise, and NLI Hypothesis. The CFEs for Table 19 are produced with *i*Flip using the OLMo2-7B model, while those for Table 20 are generated using the LLaMA3-70B model.

G Faithfulness Evaluation

Given that feature attribution-based feedback serve as a central component of our framework, it is essential to assess the faithfulness of different attribution methods. Faithfulness measures how accurately explanations reflect the true reasoning of the model (Lyu et al., 2024). We evaluate attribution methods using FERRET (Attanasio et al., 2023) with three metrics: comprehensiveness, sufficiency (DeYoung et al., 2020), and Kendall's τ correlation with Leave-One-Out (Jain and Wallace, 2019). These metrics jointly assess whether highlighted words faithfully capture the model's decision process.

The results in Table 21 show clear differences in

Table	Feedback source	Early stop under	Evaluation model
Table 10	`RoBERTa-base`	`RoBERTa-base`	`RoBERTa-base`
Table 11	`BERT`	`BERT`	`RoBERTa-base`
Table 12	`BERT`	`RoBERTa-base`	`RoBERTa-base`

Table 13: Overview of experimental settings of *i*`Flip` across tables: feedback source, early-stopping model, and evaluation classifier.

Dataset	`OLMo2-7B`	`Qwen3-32B`	`LLaMA3.3-70B`
IMDb	0.13	0.05	0.01
AG News	1.48	1.53	0.80
SNLI (Premise)	2.20	3.20	2.75
SNLI (Hypothesis)	1.83	2.62	2.41

Table 14: Average early stop rounds per dataset and model.

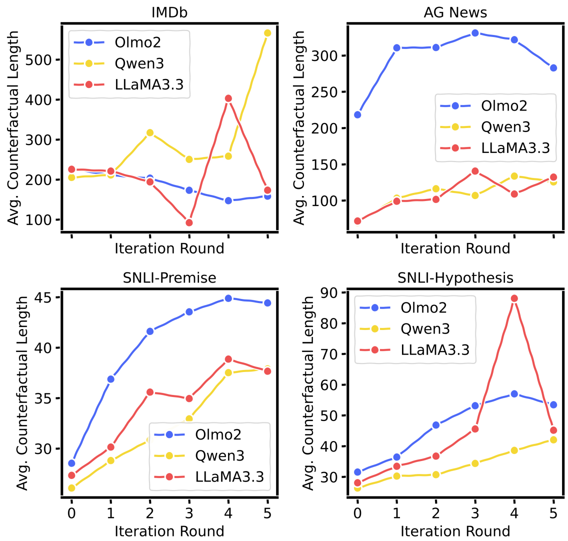

Figure 8: Average counterfactual length vs. iteration round across tasks and models.

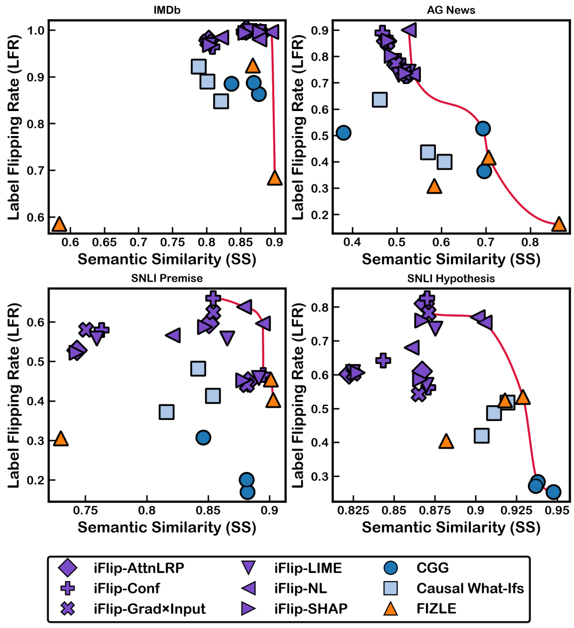

Figure 9: Trade-off between semantic similarity and label flipping rate. The upper-right region indicates the most favorable trade-off. The red curve denotes the Pareto frontier.

how attribution methods capture the model's reasoning process. Among the evaluated approaches, AttnLRP provides the most faithful explanations, showing strong performance across comprehensiveness, sufficiency, and Kendall's τ. Propagation-based methods exhibit solid alignment with the model's decision-making process, which aligns with our earlier observations in Section 6.1. SHAP follows as the next most faithful method across datasets and models, with competitive scores on comprehensiveness and sufficiency, though less consistent on τ.

We further observe that the results on the SNLI dataset show weaker faithfullness compared to sentiment and news classification. In particular, attribution methods on SNLI exhibit lower comprehensiveness, together with less favorable sufficiency values, indicating that the most important words identified by attribution methods capture the model's reasoning less effectively in NLI tasks. This suggests that attribution methods are more challenging when explanations must account for subtle entailment relations rather than more direct sentiment or news topic.

H Correlation Analysis

We further analyze the relationship between counterfactual quality and two key aspects: their performance in Counterfactual Data Augmentation (CDA) and their alignment with attribution-based faithfulness. To assess counterfactual quality, we employ the *label flipping rate* and derive feedback-specific rankings. CDA performance is measured by the ranking of average OOD accuracy on test sets obtained from models trained with different counterfactual sets. Faithfulness is captured through attribution-based metrics – *comprehensiveness* (DeYoung et al., 2020), *sufficiency* (DeYoung et al., 2020), and Kendall's τ correlation with

	Feedback of *iFlip*	IMDb			AG News		
		LFR↑	SS↑	PPL↓	LFR↑	SS↑	PPL↓
	Without Feedback	0.964	0.803	39.12	0.742	0.494	44.71
OLMo2-7B	Random Feedback	0.970(↑0.006)	0.803(—)	44.58(↑5.46)	0.794(↑0.052)	0.498(↑0.004)	43.59(↓1.12)
	Least SHAP	0.970(↑0.006)	0.802(↓0.001)	41.10(↑1.98)	0.770(↑0.028)	0.490(↓0.004)	41.78(↓2.93)
	Least AttnLRP	0.980(↑0.016)	0.805(↑0.002)	41.34(↑2.22)	0.776(↑0.034)	0.495(↑0.001)	40.74(↓3.97)
	Least Grad×Input	0.968(↑0.004)	0.803(—)	43.02(↑3.90)	0.758(↑0.016)	0.495(↑0.001)	49.09(↑4.38)
	Least LIME	0.968(↑0.004)	0.800(↓0.003)	44.99(↑5.87)	0.725(↓0.017)	0.508(↑0.014)	43.53(↓1.18)
	Conf	0.964(—)	0.809(↑0.006)	41.19(↑2.07)	0.784(↑0.042)	0.503(↑0.009)	41.41(↓3.30)
	SHAP	0.968(↑0.004)	0.805(↑0.002)	41.87(↑2.75)	0.802(↑0.060)	0.488(↓0.006)	42.95(↓1.76)
	AttnLRP	0.978(↑0.014)	0.803(—)	42.61(↑3.49)	0.768(↑0.026)	0.500(↑0.006)	46.61(↑1.90)
	Grad×Input	0.976(↑0.012)	0.801(↓0.002)	43.30(↑4.18)	0.773(↑0.031)	0.498(↑0.004)	50.32(↑5.61)
	LIME	0.968(↑0.004)	0.805(↑0.002)	52.82(↑13.70)	0.730(↓0.012)	0.504(↑0.010)	44.32(↓0.39)
	NL	0.984(↑0.020)	0.822(↑0.019)	36.47(↓2.65)	0.735(↓0.007)	0.515(↑0.021)	44.73(↑0.02)

Table 15: Automatic evaluation results of OLMo2-7B on IMDb and AG News. The tables present an ablation study of different feedback settings, including *Without Feedback*, *Random Feedback*, *Least-Attributed Feedback*, and the full variants of *iFlip*. Colored values indicate relative changes compared to the *Without Feedback* baseline (↑ = increase, ↓ = decrease).

	Feedback of *iFlip*	SNLI (Premise)			SNLI (Hypothesis)		
		LFR↑	SS↑	PPL↓	LFR↑	SS↑	PPL↓
	Without Feedback	0.508	0.872	66.42	0.778	0.879	41.61
OLMo2-7B	Random Feedback	0.672(↑0.164)	0.865(↓0.007)	53.64(↓12.78)	0.764(↓0.014)	0.870(↓0.009)	40.25(↓1.36)
	Least SHAP	0.612(↑0.104)	0.854(↓0.018)	54.51(↓11.91)	0.780(↑0.002)	0.874(↓0.005)	41.30(↓0.31)
	Least AttnLRP	0.562(↑0.054)	0.860(↓0.012)	55.31(↓11.11)	0.764(↓0.014)	0.869(↓0.010)	40.58(↓1.03)
	Least Grad×Input	0.593(↑0.085)	0.868(↓0.004)	56.99(↓9.43)	0.760(↓0.018)	0.870(↓0.009)	41.25(↓0.36)
	Least LIME	0.556(↑0.048)	0.868(↓0.004)	57.02(↓9.40)	0.754(↓0.024)	0.870(↓0.009)	41.27(↓0.34)
	Conf	0.660(↑0.152)	0.854(↓0.018)	53.72(↓12.70)	0.826(↑0.048)	0.870(↓0.009)	40.65(↓0.96)
	SHAP	0.587(↑0.079)	0.847(↓0.025)	56.81(↓9.61)	0.760(↓0.018)	0.867(↓0.012)	42.39(↑0.78)
	AttnLRP	0.596(↑0.088)	0.851(↓0.021)	55.30(↓11.12)	0.810(↑0.032)	0.869(↓0.010)	40.31(↓1.30)
	Grad×Input	0.624(↑0.116)	0.854(↓0.018)	56.02(↓10.40)	0.782(↑0.004)	0.871(↓0.008)	41.81(↑0.20)
	LIME	0.558(↑0.050)	0.866(↓0.006)	58.18(↓8.24)	0.736(↓0.042)	0.875(↓0.004)	41.02(↓0.59)
	NL	0.638(↑0.130)	0.879(↑0.007)	148.41(↑81.99)	0.770(↓0.008)	0.901(↑0.022)	43.81(↑2.20)

Table 16: Automatic evaluation results of OLMo2-7B on SNLI. The tables present an ablation study of different feedback settings, including *Without Feedback*, *Random Feedback*, *Least-Attributed Feedback*, and the full variants of *iFlip*. Colored values indicate relative changes compared to the *Without Feedback* baseline (↑ = increase, ↓ = decrease).

Leave-One-Out token removal (Jain and Wallace, 2019).

H.1 Correlation between CDA performance and counterfactual quality

Figure 14 shows the Pearson correlation between CDA performance and counterfactual quality across datasets. Among the datasets examined, AG News exhibits a strong correlation between the counterfactuals quality and CDA effectiveness, whereas IMDb and SNLI demonstrate only moderate correlation. Nonetheless, feedback-based refinement generally yields effective counterfactuals, which in turn lead to enhanced model performance and robustness.

H.2 Correlation between Faithfulness measures and counterfactual quality.

Figure 15 reports the Spearman correlation between faithfulness measures and counterfactual quality. The results reveal heterogeneous patterns across tasks: while datasets such as IMDb and AG News show positive correlations between flip-rate quality and faithfulness metrics, SNLI yields weaker or even negative correlations, especially

	Feedback of *i*`Flip`	IMDb			AG News		
		LFR↑	SS↑	PPL↓	LFR↑	SS↑	PPL↓
`OLMo2-7B`	Conf	0.677 (↓0.287)	0.774 (↓0.035)	41.32 (↑0.13)	0.507 (↓0.277)	0.539 (↑0.036)	40.15 (↓1.26)
	SHAP	0.859 (↓0.109)	0.762 (↓0.043)	41.34 (↓0.53)	0.600 (↓0.202)	0.485 (↓0.003)	40.03 (↓2.92)
	AttnLRP	0.922 (↓0.056)	0.782 (↓0.021)	40.75 (↓1.86)	0.610 (↓0.158)	0.502 (↑0.002)	38.75 (↓7.86)
	Grad×Input	0.916 (↓0.060)	0.778 (↓0.023)	40.11 (↓3.19)	0.559 (↓0.214)	0.499 (↑0.001)	38.61 (↓11.71)
	LIME	0.890 (↓0.078)	0.784 (↓0.021)	38.47 (↓14.35)	0.600 (↓0.130)	0.503 (↓0.001)	37.56 (↓6.76)
	NL	0.980 (↓0.004)	0.747 (↓0.075)	39.87 (↑3.40)	0.548 (↓0.187)	0.499 (↓0.016)	38.52 (↓6.21)

Table 17: Evaluation of `OLMo2-7B` *Without Early Stopping* on IMDb and AG News. Colored values indicate relative changes compared to *Early Stopping* (↑ = increase, ↓ = decrease).

	Feedback of *i*`Flip`	SNLI (Premise)			SNLI (Hypothesis)		
		LFR↑	SS↑	PPL↓	LFR↑	SS↑	PPL↓
`OLMo2-7B`	Conf	0.340 (↓0.320)	0.860 (↑0.006)	52.47 (↓1.25)	0.560 (↓0.266)	0.873 (↑0.003)	39.46 (↓1.19)
	SHAP	0.340 (↓0.247)	0.848 (↑0.001)	57.20 (↑0.39)	0.543 (↓0.217)	0.871 (↑0.004)	41.47 (↓0.92)
	AttnLRP	0.343 (↓0.253)	0.858 (↑0.007)	58.07 (↑2.77)	0.563 (↓0.247)	0.871 (↑0.002)	40.60 (↑0.29)
	Grad×Input	0.347 (↓0.277)	0.854 (—)	55.76 (↓0.26)	0.520 (↓0.262)	0.871 (—)	40.76 (↓1.05)
	LIME	0.323 (↓0.235)	0.875 (↑0.009)	57.04 (↓1.14)	0.497 (↓0.239)	0.872 (↓0.003)	40.38 (↓0.64)
	NL	0.387 (↓0.251)	0.844 (↓0.035)	54.87 (↓93.54)	0.546 (↓0.224)	0.903 (↑0.002)	43.81 (—)

Table 18: Evaluation of `OLMo2-7B` *Without Early Stopping* on SNLI (Premise) and SNLI (Hypothesis). Colored values indicate relative changes compared to *Early Stopping* (↑ = increase, ↓ = decrease).

Methods	**Sentiment Analysis**				**News Classification**				**NLI - Premise**				**NLI - Hypothesis**			
	Test Data				*Test Data*				*Test Data*				*Test Data*			
	ID	**OOD**			**ID**	**OOD**			**ID**	**OOD**			**ID**	**OOD**		
	IMDb	Amazon	SST-2	CFs	AG News	BBC	20NG	CFs	SNLI	MNLI	ANLI	CFs	SNLI	MNLI	ANLI	CFs
Baseline	77.93	75.00	67.43	50.09	66.87	62.80	41.07	–	40.87	34.87	**35.00**	33.43	40.87	34.87	**35.00**	33.85
Human	**92.73**	**89.53**	**85.55**	**91.26**	–	–	–	–	**42.27**	**37.47**	33.93	**50.90**	**58.40**	**52.07**	30.13	**56.95**
Conf	87.93	88.33	83.83	71.09	78.60	45.30	55.80	–	38.47	**36.33**	32.80	**32.98**	**44.13**	33.13	32.87	**36.88**
SHAP	91.33	**89.33**	83.94	90.30	80.27	**48.60**	56.33	–	37.40	34.93	**32.93**	27.82	40.33	**36.87**	33.13	35.83
AttnLRP	89.27	88.00	**84.75**	90.28	77.87	42.50	53.60	–	38.13	35.67	32.87	27.97	40.13	36.33	33.13	35.53
Grad×Input	**93.07**	88.00	82.45	90.07	79.07	43.70	55.80	–	39.00	31.93	**32.93**	26.98	39.40	34.13	**33.20**	35.98
LIME	91.40	88.73	83.37	**90.91**	78.80	42.50	54.53	–	38.33	32.60	**32.93**	28.93	38.93	35.80	32.93	35.74
NL	90.73	88.27	84.40	71.97	**81.80**	46.10	**58.47**	–	**40.40**	33.60	32.73	30.91	37.47	33.27	**33.20**	33.79

Table 19: Accuracy (%) on in-domain (ID), out-of-domain (OOD), and human-annotated counterfactual (Human CFs) test data across four tasks: Sentiment Analysis, News Classification, NLI Premise, and NLI Hypothesis. Best results are bolded. The CFEs used for data augmentation, produced with *i*`Flip`, are generated by an OLMo2-7B model.

Methods	**Sentiment Analysis**				**News Classification**				**NLI - Premise**				**NLI - Hypothesis**			
	Test Data				*Test Data*				*Test Data*				*Test Data*			
	ID	**OOD**			**ID**	**OOD**			**ID**	**OOD**			**ID**	**OOD**		
	IMDb	Amazon	SST-2	CFs	AG News	BBC	20NG	CFs	SNLI	MNLI	ANLI	CFs	SNLI	MNLI	ANLI	CFs
Baseline	77.93	75.00	67.43	50.09	66.87	62.80	41.07	–	40.87	34.87	**35.00**	33.43	40.87	34.87	**35.00**	33.85
Human	**92.73**	**89.53**	**85.55**	**91.26**	–	–	–	–	**42.27**	**37.47**	33.93	**50.90**	**58.40**	**52.07**	30.13	**56.95**
Conf	91.67	88.60	84.29	90.92	**83.07**	66.80	58.33	–	34.53	**33.00**	32.80	33.49	**48.40**	**37.40**	32.00	**41.09**
SHAP	91.53	88.27	82.80	90.71	79.73	59.30	62.40	–	36.20	32.47	32.73	33.16	40.60	34.20	31.40	38.03
AttnLRP	88.80	88.80	**84,40**	**91.21**	80.33	58.50	61.33	–	33.27	31.87	32.80	32.32	39.80	33.40	32.73	36.58
Grad×Input	**92.27**	**88.87**	83.72	90.89	81.60	59.00	62.73	–	34.47	32.27	32.73	31.84	45.47	36.00	32.07	38.71
LIME	91.00	88.80	84.06	90.83	79.80	57.30	62.47	–	36.13	32.87	32.60	33.28	41.47	35.47	**33.40**	35.86
NL	91.67	88.73	83.83	90.89	81.73	**69.50**	**66.27**	–	**37.13**	31.00	**32.87**	**33.94**	39.73	36.33	32.27	36.94

Table 20: Accuracy (%) on in-domain (ID), out-of-domain (OOD), and human-annotated counterfactual (Human CFs) test data across four tasks: Sentiment Analysis, News Classification, NLI Premise, and NLI Hypothesis. Best results are bolded. The CFEs used for data augmentation, produced with *i*`Flip`, are generated by an LLaMA3.3-70B model.

for SNLI Hypothesis. These findings suggest that higher-quality counterfactuals tend to align better

Model	Feedback	IMDb			AG News			SNLI-Premise			SNLI-Hypothesis		
		comp.	suff.	τ(loo)	comp.	suff.	τ(loo)	comp.	suff.	τ(loo)	comp.	suff.	τ(loo)
OLMo2-7B	SHAP	0.28	**-0.01**	0.02	**0.56**	**0.04**	0.03	0.09	-0.19	0.21	0.07	-0.22	0.20
	AttnLRP	**0.76**	**-0.00**	**0.09**	0.53	0.20	**0.18**	**0.12**	-0.23	0.18	**0.12**	-0.25	0.17
	Grad×Input	0.05	0.07	-0.00	0.31	0.27	0.01	0.00	**-0.07**	0.05	-0.02	**-0.11**	0.03
	LIME	0.12	0.03	0.01	0.39	0.21	0.04	0.10	-0.16	**0.23**	0.08	-0.20	**0.23**
Qwen-32B	SHAP	0.34	**-0.01**	0.04	**0.66**	**-0.03**	0.08	0.08	-0.22	0.23	0.06	-0.26	0.22
	AttnLRP	**0.78**	**-0.01**	**0.11**	0.38	0.19	**0.22**	**0.10**	-0.26	0.22	**0.10**	-0.27	0.24
	Grad×Input	0.06	0.08	-0.00	0.24	0.20	0.05	-0.01	**-0.10**	0.06	-0.03	**-0.11**	0.07
	LIME	0.15	0.05	0.02	0.43	0.06	0.08	0.08	-0.19	**0.26**	0.07	-0.24	**0.26**
Llama-70B	SHAP	0.34	**-0.00**	0.03	**0.59**	**-0.00**	0.07	0.06	-0.21	0.19	0.06	-0.25	0.19
	AttnLRP	**0.79**	**-0.00**	**0.11**	0.32	0.19	**0.22**	**0.07**	-0.24	0.20	**0.08**	-0.28	0.21
	Grad×Input	0.05	0.06	-0.00	0.22	0.16	0.06	-0.02	**-0.09**	0.05	-0.04	**-0.12**	0.04
	LIME	0.12	0.04	0.02	0.36	0.08	0.06	**0.07**	-0.19	**0.25**	0.06	-0.24	**0.23**

Table 21: Faithfulness evaluation results: **AOPC Comprehensiveness** (comp., ↑), **AOPC Sufficiency** (suff., 0 best), and **Kendall's τ correlation with Leave-One-Out** (τ(loo), ↑). Higher ↑ and lower ↓ are better. Best results per column are **bold**.

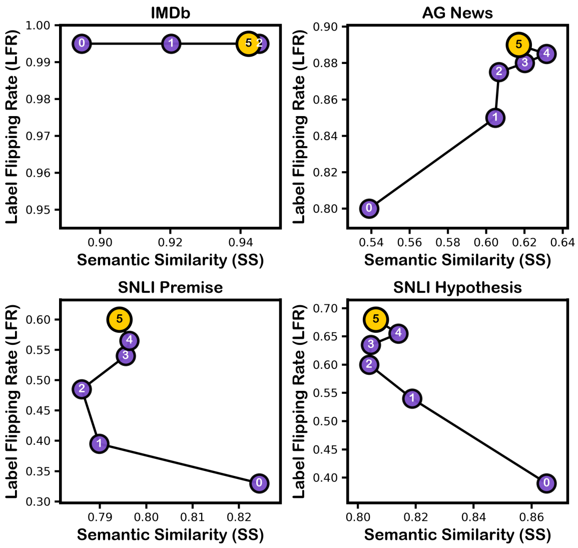

Figure 10: Trade-off between semantic similarity and label flipping rate. Markers show iteration $\mathcal{K}$. Results are from *i*Flip-NL using `LLaMA3.3-70B`.

with attribution signals in sentiment and topic classification, whereas this alignment is less stable in NLI due to the complexity of entailment reasoning.

I Highlighted Words from Attribution Methods

Table 22 illustrate how different attribution methods identify most important words within the same IMDb review.

J Additional Examples

This section presents illustrative case studies of counterfactual generation, showcasing both successful refinements and common failure patterns of the proposed framework.

J.1 Example 1: Undesired Flip

Observation: Table 23 shows that the framework can produce label flips, but not always toward the desired direction. The label was flipped from **NEUTRAL** to **ENTAILMENT** instead of the expected **CONTRADICTION**.

J.2 Example 2: Failed Flip

Observation: Table 24 shows a common failure case: the generator tends to produce longer and more elaborate counterfactuals, but the label remains unchanged despite multiple refinement rounds.

J.3 Example 3: No Early Stopping (Flip-Flop)

Observation: Table 25 shows another failure pattern: even though the generator kept producing essentially the same counterfactual, the classifier's prediction oscillated between **NEUTRAL** and **CONTRADICTION**, suggesting instability near the decision boundary.

J.4 Example 4: No Early Stopping (Flip-Flop)

Observation: Table 26 illustrates a flip-flop failure pattern: the classifier initially flipped correctly to **CONTRADICTION**, but subsequent refinements pushed the prediction back to **NEUTRAL**, showing instability in the refinement loop.

K Prompts

This section presents the full set of prompts employed in our counterfactual generation pipeline.

Method	IMDb Review
AttenLRP	This movie is on the level with Welcome Home Roxy Carmichael for biggest pieces of garbage that have ever hit the silver screen. If these guys weren't Adam Sandler's gay friends, this script would have ended up where it should have: as some big time movie exec's toilet paper. I hate this movie, it makes me want to injure people. I will admit that I have high standards, but honestly I'd rather watch Step Up 2. The ultra sad part was when I logged onto IMDb and read that you pieces of trash actually gave this movie a 6.9 rating. This is a testament to all of the retards in our society that will go watch terrible movies that are just hour and a half long dick, fart, and weed jokes with little to no originality. After seeing this rating, I would like to suggest Tyler Perry's House of Pain to all of you guys who enjoyed this film; you'll see some high quality humor there on about the same level of this abhorrent abomination.
SHAP	This movie is on the level with Welcome Home Roxy Carmichael for biggest pieces of garbage that have ever hit the silver screen. If these guys weren't Adam Sandler's gay friends, this script would have ended up where it should have: as some big time movie exec's toilet paper. I hate this movie, it makes me want to injure people. I will admit that I have high standards, but honestly I'd rather watch Step Up 2. The ultra sad part was when I logged onto IMDb and read that you pieces of trash actually gave this movie a 6.9 rating. This is a testament to all of the retards in our society that will go watch terrible movies that are just hour and a half long dick, fart, and weed jokes with little to no originality. After seeing this rating, I would like to suggest Tyler Perry's House of Pain to all of you guys who enjoyed this film; you'll see some high quality humor there on about the same level of this abomination.
LIME	This movie is on the level with Welcome Home Roxy Carmichael for biggest pieces of garbage that have ever hit the silver screen. If these guys weren't Adam Sandler's gay friends, this script would have ended up where it should have: as some big time movie exec's toilet paper. I hate this movie, it makes me want to injure people. I will admit that I have high standards, but honestly I'd rather watch Step Up 2. The ultra sad part was when I logged onto IMDb and read that you pieces of trash actually gave this movie a 6.9 rating. This is a testament to all of the retards in our society that will go watch terrible movies that are just hour and a half long dick, fart, and weed jokes with little to no originality. After seeing this rating, I would like to suggest Tyler Perry's House of Pain to all of you guys who enjoyed this film; you'll see some high quality humor there on about the same level of this abhorrent abomination.
Grad×Input	This movie is on the level with Welcome Home Roxy Carmichael for biggest pieces of garbage that have ever hit the silver screen. If these guys weren't Adam Sandler's gay friends, this script would have ended up where it should have: as some big time movie exec's toilet paper. I hate this movie, it makes me want to injure people. I will admit that I have high standards, but honestly I'd rather watch Step Up 2. The ultra sad part was when I logged onto IMDb and read that you pieces of trash actually gave this movie a 6.9 rating. This is a testament to all of the retards in our society that will go watch terrible movies that are just hour and a half long dick, fart, and weed jokes with little to no originality. After seeing this rating, I would like to suggest Tyler Perry's House of Pain to all of you guys who enjoyed this film; you'll see some high quality humor there on about the same level of this abhorrent abomination.

Table 22: IMDb review with Top-k important words highlighted under different attribution methods.

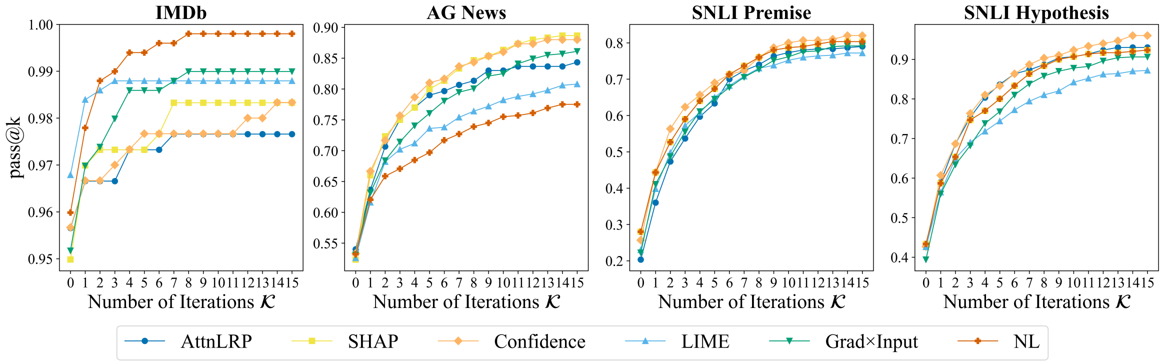

Figure 11: Iterative refinement effectiveness on OLMo2-7B. pass@k curves for each dataset (IMDb, AG News, SNLI), evaluated up to $\mathcal{K} = 15$ refinement steps.

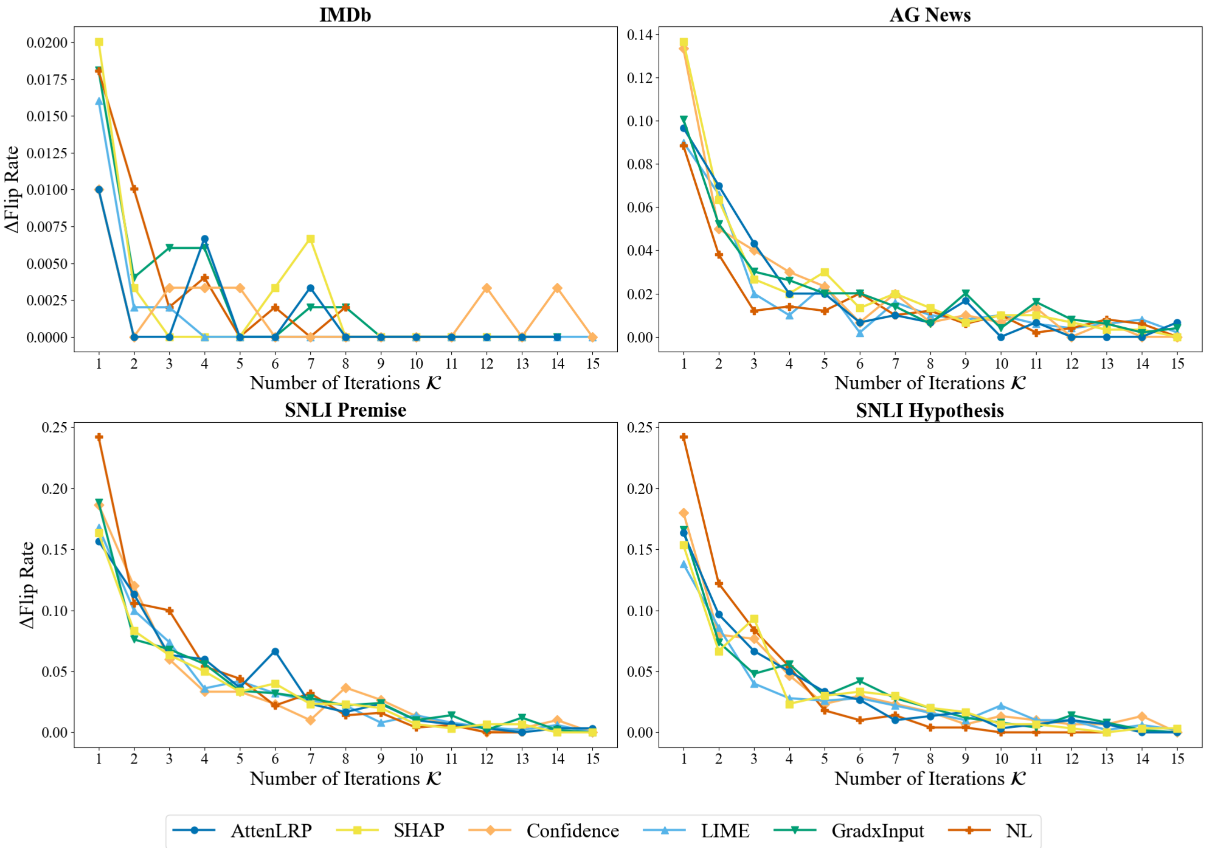

Figure 12: Change in flip rate (ΔFlip Rate) across iterative refinement rounds with OLMo2-7B for different feedback signals. The results show how much additional gain each refinement step contributes beyond the previous round.

The Base Counterfactual Prompt in Section K.1 is used to initially generate a counterfactual example. The subsequent refinement stage then improves this initial counterfactual using one of our three feedback mechanisms: the Confidence-Based Refinement Prompt in Section K.2, the Attribution-Based Refinement Prompt in Section K.3, and the Natural-Language Feedback Refinement Prompt in Section K.4.

K.1 Base Counterfactual Prompt

```
A classifier has determined that the label
of the following text is [ORIGINAL_LABEL].
Please flip it to [TARGET_LABEL] with
minimal changes. Wrap your answer in
<cf>...</cf>.
Original input:
[INPUT_TEXT]
Hint: [TASK-SPECIFIC HINT]
```

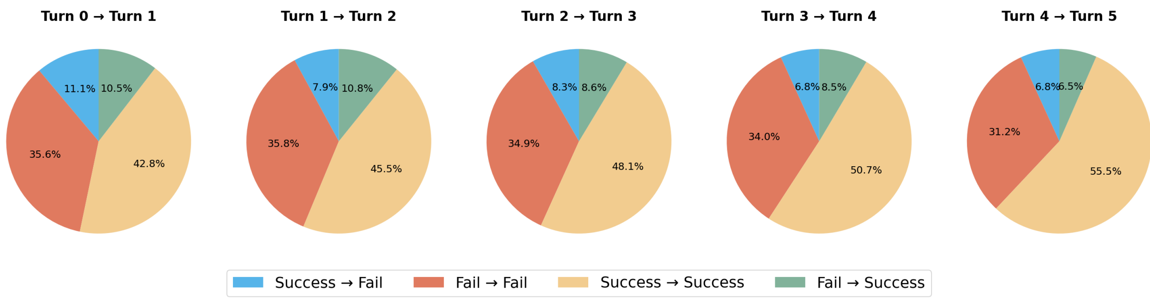

Figure 13: Iterative refinement with `OLMo2-7B` **without early stopping**, averaged across all datasets and all feedback signals. A substantial share of *Success* → *Fail* transitions indicates reduced stability compared to early-stopping.

Round	**Counterfactual Example**	**Classifier Prediction**
Original	Premise: A small girl with short brown hair is pushing a pink scooter over grass with brown leaves with trees in the background. Hypothesis: The girl is outside.	NEUTRAL (Expected: Contradiction)
Round 1	Premise: A small girl with short brown hair is inside playing with a pink scooter. Hypothesis: The girl is outside.	NEUTRAL (Expected: Contradiction)
Round 2	Counterfactual: *invalid output*	NEUTRAL
Round 3	Premise: A small girl with short brown hair is playing inside with a pink scooter. Hypothesis: The girl is outside.	NEUTRAL (Expected: Contradiction)
Round 4	Premise: A small girl with short brown hair is indoors riding a pink scooter. Hypothesis: The girl is outside.	**ENTAILMENT** (Expected: Contradiction)

Table 23: Iterative counterfactual refinement on an SNLI Premise. Top-k important words identified by AttnLRP are highlighted. *i*`Flip` flips the label, but not always to the desired class.

K.2 Confidence-based Refinement Prompt

```
The classifier now predicts your text as
[PRED_LABEL] ([CONF]% confidence), which is
the desired label.
Your task is to minimize the edits compared
to the original while preserving the current
label: [PRED_LABEL].
Make the revision as close as possible to
the original, but do not revert to the
original label: [ORIGINAL_LABEL].
Wrap ONLY the final text in <cf>...</cf>.
Original input (classifier label
[ORIGINAL_LABEL]):
[INPUT_TEXT]
Current counterfactual:
[CF_TEXT]
Hint: [TASK-SPECIFIC HINT]
```

K.3 Attribution-based Refinement Prompt

```
The classifier now predicts your text as
[PRED_LABEL], which is the desired label.
Key words influencing this prediction:
[TOP_WORDS].
Your task is to minimize the edits compared
to the original while preserving the current
label: [PRED_LABEL].
Make the revision as close as possible to
the original, but do not revert to the
original label: [ORIGINAL_LABEL].
Wrap ONLY the final text in <cf>...</cf>.
Original input (classifier label
[ORIGINAL_LABEL]):
[INPUT_TEXT]
Current counterfactual:
[CF_TEXT]
Hint: [TASK-SPECIFIC HINT]
```

K.4 Natural-Language Feedback Refinement Prompt

Natural-language refinement first asks the model to generate NL feedback analyzing how to minimally achieve the target label. This feedback is then used to guide a second prompt that produces the counterfactual.

Round	Counterfactual Example	Classifier Prediction
Original	Premise: A ballerina standing on the toes of one foot with one arm up in the air and one leg lifted behind her. Hypothesis: The ballerina hates this.	CONTRADICTION (Expected: Neutral)
Round 1	Premise: A ballerina stands gracefully on the toes of one foot, her arm reaching towards the sky, while the other leg is lifted behind her, a look of intense concentration on her face as she perfectly captures the beauty and intricacy of the moment in dance, amidst a bustling cityscape outside the studio window, completely unaware of the chaos beyond her focus. Hypothesis: The ballerina hates this.	CONTRADICTION (Expected: Neutral)
Round 2	Premise: A ballerina stands gracefully on the toes of one foot, her arm reaching towards the sky, while the other leg is lifted behind her, a look of intense concentration on her face as she perfectly captures the beauty and intricacy of the moment in dance, happening during her first solo performance in front of a packed theater, completely unaware of the standing ovation from the enthralled audience outside the proscenium arch. Hypothesis: The ballerina hates this.	CONTRADICTION (Expected: Neutral)
Round 3	Premise: A ballerina stands gracefully on the toes of one foot, her arm reaching towards the sky, while the other leg is lifted behind her, capturing the serene beauty of a peaceful morning scene in the park, completely unaware of the beautiful sunrise illuminating the clouds overhead. Hypothesis: The ballerina hates this.	CONTRADICTION (Expected: Neutral)
Round 4	Premise: A ballerina stands gracefully, her arm elegantly raised, foot poised on the tip, as she prepares for a grand performance on a bustling city stage, her eyes filled with determination and focus. The bright lights above illuminate her, casting long shadows on the polished wooden floor. The audience's anticipation is palpable in the silent, charged air of the theatre. Hypothesis: The ballerina hates this.	CONTRADICTION (Expected: Neutral)
Round 5	Premise: A ballerina stands gracefully on a beach, one arm stretched upward towards the sky as if greeting the sun, the other hand resting lightly on her hip, amidst a sea of beachgoers. The air is salty and warm, and the sand is soft beneath her feet. Around her, people are lounging, laughing, and enjoying the freedom of the day. The city's skyline is visible in the distance, with skyscrapers gleaming against the azure sky. Hypothesis: The ballerina hates this.	CONTRADICTION (Expected: Neutral)

Table 24: Iterative counterfactual refinement on an SNLI Premise. Top-k important words identified by AttnLRP are highlighted. Despite multiple refinements, the prediction remains CONTRADICTION instead of flipping to the expected NEUTRAL.

K.4.1 Feedback Generation Prompt (NL Feedback)

```
Analyze the current counterfactual and
suggest improvements to achieve the target
label [TARGET_LABEL].
You should make the smallest possible edits
to the text while still achieving the
target.
Wrap your reasoning inside
<think>...</think>.

Original (classifier label
[ORIGINAL_LABEL]):
[INPUT_TEXT]

Current counterfactual:
[CF_TEXT]
```

K.4.2 Counterfactual Editing Prompt (Using NL Feedback)

```
Based on the feedback below, revise the text
so that it flips to  [TARGET_LABEL].

Feedback:
[FEEDBACK_TEXT]

Wrap ONLY the final counterfactual in
<cf>...</cf>.

Original text (classifier label:
[ORIGINAL_LABEL]):
[INPUT_TEXT]

Current counterfactual:
[CF_TEXT]
```

Round	Counterfactual Example	Classifier Prediction
Original	Premise: A man in a business suit silently meditating in front of the camera. Hypothesis: A business man is giving a heated interview for the local news.	NEUTRAL (Expected: Contradiction)
Round 1	Premise: A man in a business suit silently meditating in front of the camera. Hypothesis: A business man is giving a heated interview for the local news.	NEUTRAL (Expected: Contradiction)
Round 2	Premise: A man in a business suit silently meditating in front of the camera. Hypothesis: A business man is giving a heated interview for the local news.	**CONTRADICTION** (Expected: Contradiction)
Round 3	Premise: A man in a business suit silently meditating in front of the camera. Hypothesis: A business man is giving a heated interview for the local news.	NEUTRAL (Expected: Contradiction)
Round 4	Premise: A man in a business suit silently meditating in front of the camera. Hypothesis: A business man is giving a heated interview for the local news.	**CONTRADICTION** (Expected: Contradiction)
Round 5	Premise: A man in a business suit silently meditating in front of the camera. Hypothesis: A business man is giving a heated interview for the local news.	NEUTRAL (Expected: Contradiction)

Table 25: Iterative counterfactual refinement on an SNLI premise. Top-k important words identified by AttnLRP are highlighted. Despite producing nearly identical counterfactuals, the classifier oscillates between CONTRADICTION and NEUTRAL, illustrating instability near the decision boundary.

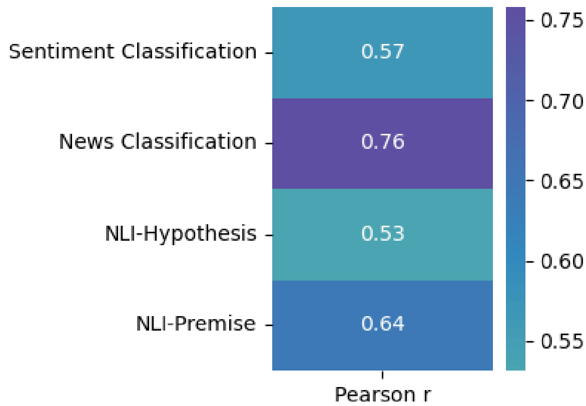

Figure 14: Pearson correlation between augmentation performance and counterfactual quality across datasets.

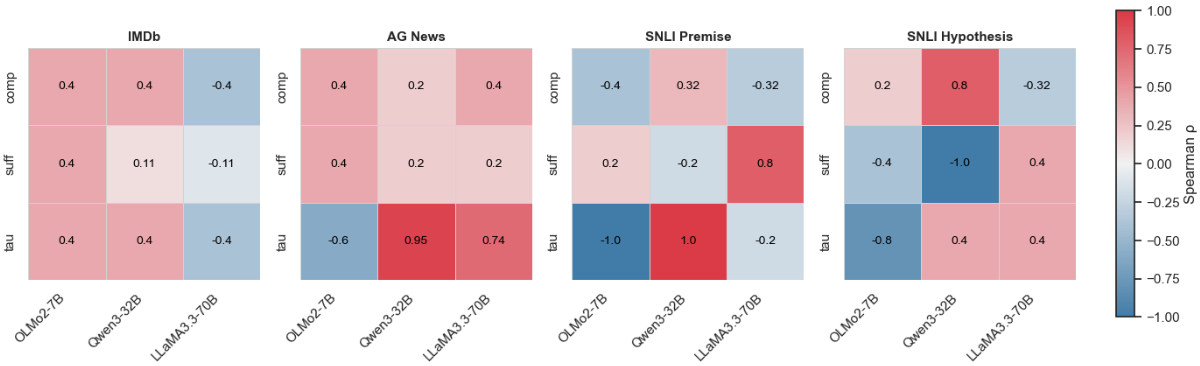

Figure 15: Spearman correlation between faithfulness and counterfactual quality across datasets.

Round	Counterfactual Example	Prediction (Expected)
Original	Premise: Man in robe making designs on dirt ground. Hypothesis: A man making designs in the dirt outside.	NEUTRAL
Round 1	Hypothesis: A man making designs in the dirt outdoors.	NEUTRAL (Expected: Contradiction)
Round 2	Hypothesis: A man is carefully cleaning a spotless carpet indoors.	**CONTRADICTION** (Expected: Contradiction)
Round 3	Hypothesis: A man is carefully cleaning a spotless carpet indoors while hlthere's a man in a robe making designs on a dirt ground outside.	NEUTRAL (Expected: Contradiction)
Round 4	Hypothesis: A man is meticulously cleaning a spotless carpet indoors while simultaneously, there's a man in a robe making designs on a dirt ground outside, in total contradiction to the original scene where there's no mention of indoor cleaning and only a man making designs on dirt.	NEUTRAL (Expected: Contradiction)
Round 5	Hypothesis: A man indoors is vacuuming a spotless carpet while another man outside is making intricate designs on a dirt ground, representing a total contradiction as it juxtaposes cleaning indoors with design-making outdoors, and a spotless carpet with a dirt ground.	NEUTRAL (Expected: Contradiction)

Table 26: Iterative counterfactual refinement on an SNLI Hypothesis. Top-k important words identified by AttnLRP are highlighted. The classifier briefly flipped to CONTRADICTION in Round 2 but reverted back to NEUTRAL.

L User Study Guidelines

User Study Guidelines

Dear participants,

Thank you for participating in our user study. This study focuses on evaluating model-generated explanations, specifically counterfactual explanations (definitions and examples follow). We generate counterfactuals using three approaches in total, and you will evaluate 10 counterfactuals generated by each approach. You will subjectively rate each explanation based on the following three dimensions: **completeness**, **understandability**, and **cohesiveness**.

Counterfactual Example (CFE): the smallest change to an input that would cause a model to output a different prediction. It is important to note that "change in prediction" refers to the model's predicted label, not the ground truth label.

Simple Example (Sentiment Classification):

- Original: "The movie was boring." → Model predicts **negative**.
- CFE: "The movie was great." → Model predicts **positive**.

In our questionnaire, the task is more complex: the model modifies news sentences from AG News, which has four categories: Business, Sports, World, Sci/Tech. You will manually rate CFEs using the following metrics. Higher scores indicate higher quality.

1. Completeness (1–6): The explanation is sufficient in explaining the outcome. In other words, does the CFE provide enough information to clearly support the new predicted label?

- **1** – The CFE is not enough to justify the new label; explanation is insufficient
- **3** – Partially justifies the new label; some key information is missing
- **6** – Fully justifies the new label; enough information to clearly support the prediction shift

We give two special cases to help you rate:

- **Case I**:
  Original: "The company reported lower profits this quarter." → Model predicts Business.
  CFE: "The company reported lower profits this sports quarter." → Model predicts Sports.
  We may rate this CFE a low completeness score (e.g., 1–2). Because the topic of CFE is still Business, though it did mention sports quarter.
- **Case II**:
  Original: "The president met with foreign leaders to discuss regional security." → Model predicts World.
  CFE: "The president met with foreign leaders to discuss new technology." → Model predicts Sci/Tech.
  We may rate this CFE a low completeness score (e.g., 1–2). The sentence still mainly describes political or diplomatic meetings; the small mention of "technology" is not enough to justify Sci/Tech.

2. Overall satisfaction (1–6): This scenario effectively explains how to reach a different outcome.

- **1** – Does not explain how to reach a different outcome
- **3** – Explains it partially but lacks clarity
- **6** – Clearly and fully explains how to reach a different outcome

We give one special case to help you rate:

- **Case I**:
  Original: "Leaders met to discuss international security and trade sanctions." → Model predicts World.
  CFE: "Leaders met to discuss international sports security and trade sanctions." → Model predicts World.
  We may rate this CFE a low Overall Satisfaction score (e.g., 1–2), because the change from "security" to "sports security" is too minor to meaningfully alter the sentence's topic.

3. Feasibility (1–6): The actions suggested by the explanation are practical, realistic to implement and actionable.

- **1** – unrealistic or impossible
- **3** – partly feasible but unclear or difficult
- **6** – fully realistic and actionable

We give one special case to help you rate:

- **Case I**:
  Original: "Apple Inc. won an international innovation award." → Model predicts Business.
  CFE: "Apple Inc. won the World Cup" → Model predicts Sports.
  We may rate this CFE a low score (e.g., 1–2), because the change is unrealistic. A company cannot win the World Cup.

M LLM-as-a-Judge Evaluation

We adopt an LLM-as-a-judge protocol to assess the explanatory quality of counterfactual from three subjective perspectives: *Completeness*, *Overall Satisfaction*, and *Feasibility*.

M.1 Prompt

The following prompt instructs the judge model to rate each CFE using integer scores from 1 to 6.

```
You are an evaluation model
(LLM-as-a-judge). You will be given:
- Original text
- Counterfactual example (CFE): A minimally
edited version of the input text that
results in a change in the model's
prediction.
- Original prediction and counterfactual
prediction
- Whether the prediction flips (isFlip)

Your task is to score the CFE scenario's
explanatory quality on three metrics. Use
integer scores from 1 to 6 (1 = lowest, 6 =
highest).
Metrics:
1) Completeness (1-6): The explanation is
sufficient in explaining the outcome.
2) Overall satisfaction (1-6): This scenario
effectively explains how to reach a
different outcome.
3) Feasibility (1-6): The actions suggested
by the explanation are practical, realistic
to implement and actionable.

Output MUST be exactly in the following
format (three tags, integers only):
<completeness>score</completeness><satisfaction>score</satisfaction><feasibility>score</feasibility>

Do not output anything else.
```

M.2 Evaluation Across Judge Models

Table 27 reports the mean ratings for all counterfactuals across three judge models.

Model	Method	Complet.	Overall Sat.	Feasib.	Kripp. α
Gemma3-27B	CGG	3.4	3.4	4.5	0.649
	FIZLE	3.2	3.2	3.3	
	*i*Flip-NL	**5.2**	**5.3**	**5.1**	
GPT-OSS-120B	CGG	2.7	2.6	2.9	0.666
	FIZLE	2.7	2.7	3.3	
	*i*Flip-NL	**3.4**	**3.4**	**3.5**	
DeepSeek-R1	CGG	3.5	3.3	4.4	0.571
	FIZLE	3.4	3.3	4.8	
	*i*Flip-NL	**5.0**	**4.7**	**4.6**	

Table 27: Evaluation results on *Completeness*, *Overall Satisfaction*, and *Feasibility* across three judge models. Best results are bolded. Krippendorff's α denotes Krippendorff's alpha measuring agreement between the judge model ratings and human ratings.

N Error Analysis

Subsection N.1 and N.2 discusses Incomplete and Unrealistic CFE Candidates, and Subsection N.4 shows how iterative refinement can enhance the completeness and feasibility of CFE candidates.

N.1 Incomplete CFE Candidates

As illustrated in Table 28, a common failure case occurs when only partial entity mentions are edited, resulting in incomplete modifications that fail to sufficiently alter the model's predictions.

Dataset	AG News
Method	CGG
Original	Senate approves tax relief bill for manufacturers The Senate today passed a far-reaching, $136 billion corporate tax package that cuts taxes for businesses ranging from film companies to bow-and-arrow makers while closing tax loopholes and bringing US exporters in line with **Prediction:** business
CFE	Senate approves research funding bill for innovators The Senate today passed a far-reaching, $136 billion scientific research package that cuts funding for projects ranging from space companies to robotics makers while closing research gaps and bringing US scientists in line with **Prediction:** business
Issue	Although many business-related entities are replaced with scientific terms, enough business cues remain and the overall prediction does not change.

Table 28: Error analysis for incomplete CFE candidates

N.2 Unnatural CFE Candidates

As shown in Table 29, another common failure case is when modifications produce counterfactuals that are unnatural. The generated CFE candidates contain contexts that are inconsistent with real-world facts and scenarios.

Dataset	AG News
Method	FIZLE
Original	IBM Claims Computing Crown (The Motley Fool) The Motley Fool - IBM (NYSE: IBM - News) has new bragging rights. Press reports indicate that the technology giant has created the world's fastest supercomputer two years after a Japanese computer claimed that title. **Prediction:** sci/tech
CFE	IBM Claims Tennis Crown (The Sports Fool) The Sports Fool - IBM (NYSE: IBM - News) has new bragging rights. Press reports indicate that the tennis giant has created the world's fastest serve two years after a Japanese player claimed that title. **Prediction:** sci/tech
Issue	The CFE unrealistically portrays IBM, a technology company, as a tennis giant, creating a scenario that conflicts with real world.

Table 29: Error analysis for unnatural CFE candidates.

N.3 Per-annotator Human and LLM Ratings

Table 31 presents two concrete examples that report detailed per-annotator ratings from a human user study (H1–H3) and an LLM-as-a-judge evaluation (L1–L3; `Gemma3-27B`, `GPT-OSS-120B`, and `DeepSeek-R1`) on three dimensions: *Completeness*, *Overall Satisfaction*, and *Feasibility*.

N.4 Enhancing Completeness and Feasibility with Iterative Refinement

As shown in Table 30, iterative refinement helps address both incompleteness and unnatural counterfactual scenarios in CFE candidates.

Dataset	AG News
Method	*i*`Flip`
Original	Gold Fields wins appeal to fight takeover Gold Fields Ltd. won an appeal on Friday in its battle to stave off a hostile $7.1 billion takeover by Harmony Gold Mining Co. that would create the world #39;s largest gold mining company. **Prediction:** business
CFE (a) (*Iter.* $\mathcal{K}=1$)	Gold Fields wins global appeal to address international concerns Gold Fields Ltd. won a global appeal on Friday in its effort to address international concerns by Harmony Gold Mining Co. that would impact global stability, sparking concerns at the United Nations and among nations worldwide. **Prediction:** business
CFE (b) (*Iter.* $\mathcal{K}=2$)	Gold Fields wins global appeal to address international concerns, sparking widespread discussion on the world stage about the global implications of mining operations and their potential impact on international stability, prompting concerns at the United Nations and among nations worldwide. **Prediction:** world
Observation	Although **CFE (a)** adds world-related context, it is still mainly about the legal appeal, so the model keeps the *business* label and the claim about affecting global stability is not very realistic. After one round of iterative refinement, **CFE (b)** becomes more complete and realistic by highlighting the global debate around mining operations, and consequently the model prediction changes to *world*.

Table 30: Error analysis for Iterative Refinement.

Dataset	AG News
Original	*"Senate approves tax relief bill for manufacturers The Senate today passed a far-reaching, $136 billion corporate tax package that cuts taxes for businesses ranging from film companies to bow-and-arrow makers while closing tax loopholes and bringing US exporters in line with "*
CFE by CGG	*"Senate approves research funding bill for innovators The Senate today passed a far-reaching, $136 billion scientific research package that cuts funding for projects ranging from space companies to robotics makers while closing research gaps and bringing US scientists in line with"* **Orig. Pred** business — **Completeness** H1–H3: 3/2/4 L1–L3: 3/1/2 **CFE Pred** business — **Overall Satisfaction** H1–H3: 2/2/3 L1–L3: 3/1/2 **isFlip** Yes — **Feasibility** H1–H3: 3/2/2 L1–L3: 4/1/4
CFE by FIZLE	*"Sports fans are approval of tax relief bill for sports manufacturers The Senate today passed a far-reaching, $136 billion corporate tax package that cuts taxes for businesses ranging from sports equipment companies to athletic wear makers while closing tax loopholes and bringing US sports exporters in line with international standards, which is expected to boost the sports industry and create new jobs for athletes and sports enthusiasts."* **Orig. Pred** business — **Completeness** H1–H3: 6/3/1 L1–L3: 6/5/5 **CFE Pred** sport — **Overall Satisfaction** H1–H3: 3/3/5 L1–L3: 6/5/5 **isFlip** No — **Feasibility** H1–H3: 3/2/3 L1–L3: 6/5/5
CFE by *i*Flip	*"Senate approves research funding bill for sci/tech innovators. The Senate today passed a far-reaching, $136 billion research package that funds cutting-edge projects in areas like quantum computing, biotechnology, and advanced materials for companies ranging from artificial intelligence startups to renewable energy labs, while supporting collaborations with institutions like NASA, Google, and MIT, and enhancing the global competitiveness of US tech exporters."* **Orig. Pred** business — **Completeness** H1–H3: 5/5/6 L1–L3: 6/5/6 **CFE Pred** sci/tech — **Overall Satisfaction** H1–H3: 6/6/6 L1–L3: 6/5/6 **isFlip** Yes — **Feasibility** H1–H3: 6/5/6 L1–L3: 6/4/6

Table 31: Detailed ratings from three human annotators (H1–H3) and three LLM judges (L1–L3), where L1=Gemma3-27B, L2=GPT-OSS-120B, and L3=DeepSeek-R1.