FitCF: A Framework for Automatic Feature Importance-guided Counterfactual Example Generation

Anonymous ACL submission

Abstract

Counterfactual examples are widely used in natural language processing (NLP) as valuable data to improve models, and in explainable artificial intelligence (XAI) to understand model behavior. The automated generation of counterfactual examples remains a challenging task even for large language models (LLMs), despite their impressive performance on many tasks. In this paper, we first introduce ZeroCF, a faithful approach for leveraging important words derived from feature attribution methods to generate counterfactual examples in a zero-shot setting. Second, we present a new framework, FitCF[1], which further verifies aforementioned counterfactuals by label flip verification and then inserts them as demonstrations for few-shot prompting, outperforming two state-of-the-art baselines. Through ablation studies, we identify the importance of each of FitCF's core components in improving the quality of counterfactuals, as assessed through flip rate, perplexity, and similarity measures. Furthermore, we show the effectiveness of *LIME* and *Integrated Gradients* as backbone attribution methods for FitCF and find that the number of demonstrations has the largest effect on performance. Finally, we reveal a strong correlation between the faithfulness of feature attribution scores and the quality of generated counterfactuals.

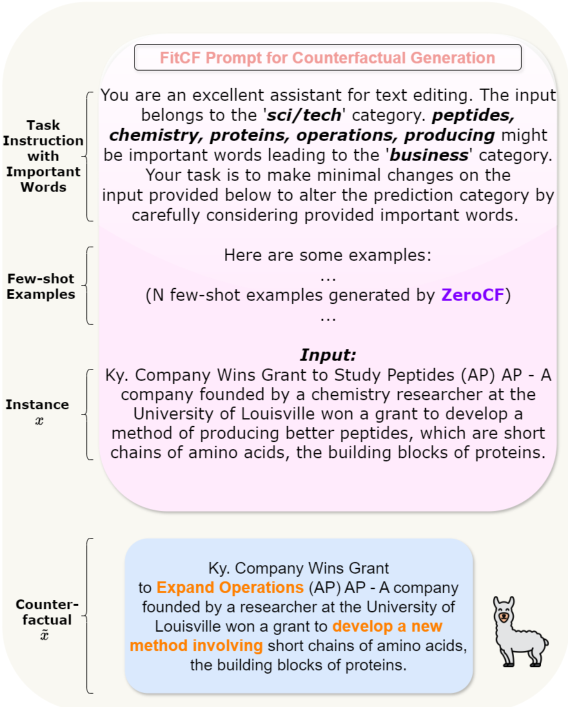


Figure 1: Given an instance x from the AG News dataset classified as "***sci/tech***", our ZeroCF approach generates few-shot examples, whose important words are determined by *LIME* for a BERT model. FitCF then generates a counterfactual $\tilde{x}$ on this basis. The edits to original instance x are highlighted in orange, yielding $\tilde{x}$ which is classified as "***business***".

1 Introduction

The advent of increasingly complex and opaque LLMs has triggered a critical need for explainability and interpretability of such models. Counterfactuals, which are minimally edited inputs that yield different predictions compared to reference inputs (Miller, 2019; Ross et al., 2021; Madsen et al., 2022) are widely used in XAI and NLP. Applications include creating new data points for improving models in terms of performance (Kaushik et al., 2020; Sachdeva et al., 2024) and robustness (Gardner et al., 2020; Ross et al., 2021) and understanding the black-box nature of models (Wu et al., 2021; Wang et al., 2024). Crowd-sourcing counterfactuals can be costly, inefficient, and impractical (Chen et al., 2023), particularly in specialized domain such as medicine. LLM-based counterfactual generation offers a more efficient and scalable alternative. Despite advancements in counterfactual generation techniques and the demonstrated versatility of LLMs across tasks (Wu et al., 2021; Bhan et al., 2023; Li et al., 2024), the efficacy of LLMs

[1]Code: https://anonymous.4open.science/r/FitCF

in producing high-quality counterfactuals in a zero-shot setting, as well as the effective construction of valid counterfactuals as demonstrations to enable few-shot prompting, remains an open question (Bhattacharjee et al., 2024b). Additionally, the combination of widely used interpretability methods, with the goal to exploit their combined benefits, has been insufficiently explored within XAI research (Treviso et al., 2023; Baeumel et al., 2023; Bhan et al., 2023).

To this end, we first present ZEROCF, a method to combine feature importance with counterfactual generation by leveraging important words identified through feature attribution scores for a fine-tuned BERT (Devlin et al., 2019) on the target dataset, evaluated on four representative feature importance methods (§4.4). The generation of counterfactuals with ZEROCF is performed by prompting LLMs with extracted important words in a zero-shot setting without any auxiliary counterfactual data (§3.1). We then propose the FITCF framework (Figure 1), which uses ZEROCF-generated counterfactuals following a label flip verification step as demonstrations for few-shot prompting without relying on human-crafted examples (§3.2).

Secondly, we evaluate ZEROCF and FITCF on two NLP tasks - news topic classification and sentiment analysis - using two baselines, POLYJUICE (Wu et al., 2021) and FIZLE (Bhattacharjee et al., 2024b). The automatic evaluation employs three automated metrics: Label flip rate, fluency, and edit distance. Both ZEROCF and FITCF significantly outperform POLYJUICE, with ZEROCF surpassing FIZLE in most cases and FITCF consistently exceeding both baselines and ZEROCF.

Thirdly, we perform ablation studies on three key components of FITCF: (1) Important words; (2) the number of demonstrations; (3) label flip verification. The results reveal that all three components contribute positively to improving the quality of counterfactuals, as measured by label flip rate, fluency, and edit distance, with the number of demonstrations being the most influential. In addition, FITCF exhibits greater robustness and achieves higher overall quality when combined with *LIME* and *SHAP* compared to its combination with *Gradient* and *Integrated Gradients*.

Lastly, we conduct a correlation analysis between the quality of generated counterfactuals and the faithfulness of feature attribution scores as used in ZEROCF and FITCF. The analysis reveals that *LIME* and *SHAP* can produce more faithful feature attribution scores compared to *Gradient* and *Integrated Gradients*. Furthermore, we observe a strong correlation between the faithfulness of these feature attribution scores and the quality of counterfactuals generated by FITCF.

2 Related Work

Counterfactual Generation MICE generates contrastive edits that change the prediction to a given contrast prediction (Ross et al., 2021). POLYJUICE uses a fine-tuned GPT-2 (Radford et al., 2019) to specify the type of edit needed to generate counterfactual examples (Wu et al., 2021). DISCO (Chen et al., 2023) uses the GPT-3 fill-in-the-blank mode (Brown et al., 2020), which is not available in most open-source LLMs (Chen et al., 2023). Bhattacharjee et al. (2024a) identify the latent features in the input text and the input features associated with the latent features to generate counterfactual examples, which is criticized due to the additional level of complexity with no significant performance gain (Delaunay et al., 2024). FIZLE (Bhattacharjee et al., 2024b) shares the most similarity with FITCF and uses LLMs as pseudo-oracles to generate counterfactuals with the assistance of LLM-generated important words in a zero-shot setting.

Combination of Interpretability Methods Recent works have explored the possibility to combine different XAI methods. Wang et al. (2021) propose a feature importance-aware attack, which disrupts important features that consistently influence the model's decisions. Gressel et al. (2023) identify perturbations in the feature space to produce evasion attacks. Treviso et al. (2023) present the framework, CREST, to generate counterfactual examples by combining rationalization with span-level masked language modeling. Krishna et al. (2023) employ various post-hoc explanations for rationalization, extending beyond counterfactuals, in contrast to CREST. Bhan et al. (2023) propose a method to determine impactful input tokens with respect to generated counterfactual examples. In contrast, FITCF uses feature importance to guide counterfactual example generation.

3 Methodology

3.1 ZEROCF

Bhattacharjee et al. (2024b) introduced FIZLE, which generates counterfactuals in a zero-shot setting by prompting the LLM with important words

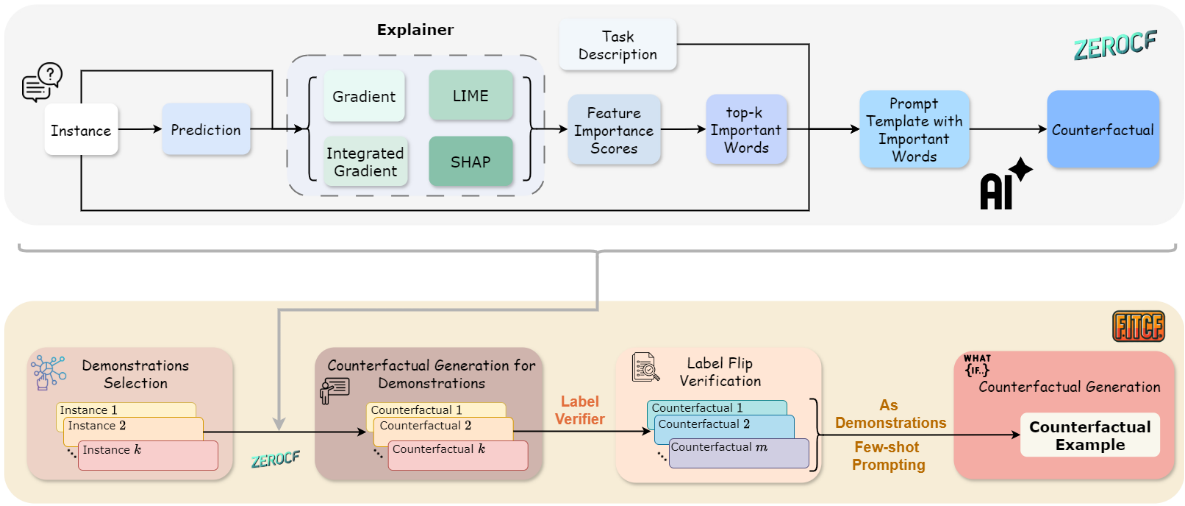


Figure 2: The upper part of the figure illustrates how counterfactuals are generated by ZEROCF using important words extracted by the explainer (BERT) through various feature important methods (*Gradient, Integrated Gradients, LIME, SHAP*). Lower part of the figure shows the pipeline of FITCF involving demonstrations selection, automatic construction of counterfactual examples by ZEROCF, label flip verification, and counterfactual generation.

identified by the LLM itself. However, these extracted words may be unfaithful or hallucinated (Li et al., 2023)[2]. To address this limitation, we propose ZEROCF (Figure 2; examples are provided in Table 7), which relies on the most attributed words based on feature attribution scores determined by various explanation methods for the predictions of a BERT model fine-tuned on the target dataset. Feature importance involves determining how significant an input feature is for a given output (Madsen et al., 2022), which we find to enhance the counterfactual generation process (§6.1).

Prediction Given an input x from the dataset $\mathcal{D}$, we leverage a BERT model $\mathcal{M}_{\mathcal{D}}$ fine-tuned on $\mathcal{D}$[3] to obtain the prediction y_{pred} for the given input x:

$$y_{pred} = \mathcal{M}_{\mathcal{D}}(x) \tag{1}$$

Feature Attribution Scores Then we deploy an explainer $\mathcal{E}$ with access to the model $\mathcal{M}_{\mathcal{D}}$, which employs various feature importance methods f (§4.4) to acquire feature attributions scores s based on the prediction y_{pred} and the given input x:

$$s = \mathcal{E}(x, y_{pred}, f, \mathcal{M}_{\mathcal{D}}) \tag{2}$$

Counterfactual Generation Finally, we identify the top-attributed words[4] w based on feature attribution scores s and deploy an LLM $\mathcal{L}$ in a zero-shot setting to generate the counterfactual $\tilde{x}$ with the prompt p (§A.1), which consists of task instruction i, words w, the prediction y_{pred}, and the input x:

$$\tilde{x} = \mathcal{L}(p) \tag{3}$$

3.2 FITCF

While ZEROCF mitigates the issue of hallucinated important words extracted by the LLM, the counterfactuals generated by ZEROCF may fail to flip the prediction, e.g., due to the limited capability of zero-shot prompting (Brown et al., 2020). To address it, we propose FITCF (Figure 1, Figure 2), inspired by Auto-CoT (Zhang et al., 2023), which combines two interpretability methods, feature importance and counterfactual examples, leveraging their complementary advantages and automatically constructs demonstrations by ZEROCF incorporating label-flip verification. Verified demonstrations subsequently enable few-shot prompting in FITCF.

top-k Examples Sampling In order to diversify demonstration selection (An et al., 2023; Zhang et al., 2023) and construct demonstrations automatically, we first convert all instances from the

[2] Applying `Llama3-8B` with FIZLE on AG News, we find that for 64.5% of the instances, a subset of generated important words is hallucinated, i.e., absent from the original input.

[3] Detailed information, e.g., accuracy, about the deployed BERT models is provided in Appendix B.

[4] The top attributed words are further post-processed by replacing the "[CLS]" and "[SEP]" special tokens if any, with the subsequent attributed words and by merging tokenized subwords if one of them is a top attributed word.

dataset $\mathcal{D}$ into sentence embeddings using SBERT[5], and then apply k-means clustering on these sentence embeddings to form c clusters[6]. Afterwards, we select a total of k instances which are closest to the centroid of each cluster[7]. In such a way, we diversify the demonstrations, potentially mitigating any misleading effects caused by ZEROCF, which may produce flawed counterfactuals. Finally, ZEROCF is employed to generate counterfactuals for the k selected instances using simple heuristics.

Label Flip Verification Subsequently, in order to validate the generated counterfactuals and to prevent incorrect counterfactuals from misleading the LLM (Turpin et al., 2023), we employ the same BERT model $\mathcal{M}_\mathcal{D}$ (§3.1) to make predictions on k generated counterfactuals $\mathcal{C} = \{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_k\}$ and the original input $\mathcal{X} = \{x_1, x_2, ..., x_k\}$ individually and assess whether the labels are inconsistent:

$$\forall i \in \{1, 2, \cdots, k\} : \hat{y}_{x_i} = \mathcal{M}_\mathcal{D}(x_i) \qquad (4)$$

$$\forall i \in \{1, 2, \cdots, k\} : \hat{y}_{\tilde{x}_i} = \mathcal{M}_\mathcal{D}(\tilde{x}_i) \qquad (5)$$

The generated counterfactuals $\tilde{x}_i$, where the predicted labels remain consistent $\hat{y}_{\tilde{x}_i} = \hat{y}_{x_i}$, are excluded from the demonstrations for further process to ensure the validity of the generated counterfactuals. In the end, we obtain m counterfactuals, where $m \leq k$. To maintain a consistent number of demonstrations (ℓ) for each input, if $m < \ell$, additional examples are iteratively selected based on their proximity to the cluster centroid, until the required number of demonstrations is achieved.

Counterfactual Generation For a given input x, ℓ input-counterfactual pairs generated by ZEROCF are used as demonstrations, along with important words w extracted based on the feature attribution scores s generated by BERT (§3.1), to prompt the LLM to generate the counterfactual for the input x in a few-shot setting (Figure 2, §A.2).

3.3 Considerations for Choice of Models

In ZEROCF, feature attributions are generated for a BERT model's predictions, based on which important words are then extracted(§3.1). Moreover, in FITCF, the same BERT model serves as a label flip verifier (§3.2). We emphasize that any model capable of performing classification tasks effectively can be used as a label flip verifier or for generating feature attribution scores[8].

4 Experimental Setup

4.1 Baselines

We employ the following two approaches as baselines for FITCF.

Polyjuice POLYJUICE allows users to control perturbation types and deploys a GPT-2[9] to generate counterfactuals by framing the task as a conditional text generation problem (Wu et al., 2021).

FIZLE FIZLE employs an LLM to identify important words and prompts the LLM with these words in a zero-shot setting to generate counterfactuals (Bhattacharjee et al., 2024b).

4.2 Dataset

Following Nguyen et al. (2024); Bhattacharjee et al. (2024b), we demonstrate the validity of ZEROCF and FITCF by applying them to two NLP tasks: News topic classification and sentiment analysis[10].

AG News AG News (Zhang et al., 2015) contains news articles created by combining the titles and description fields of articles from four categories: *World*, *Sports*, *Business*, and *Sci/Tech*.

SST2 SST2 (Socher et al., 2013) is part of the larger Stanford Sentiment Treebank and focuses specifically on binary sentiment classification of natural language movie reviews. Each sentence is labeled as either *negative* or *positive*.

4.3 Models for Counterfactual generation

We select three open source state-of-the-art instruction fine-tuned LLMs with increasing parameter sizes[11]: `Llama3-8B` (AI@Meta, 2024), and `Qwen2.5-{32B,72B}` (Team, 2024).

[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[6] Clustering visualizations are given in Appendix C.

[7] Selected examples and their corresponding counterfactuals for a given instance are provided in Appendix D.

[8] For encoder-only architectures like the BERT model employed in our study, tools like FERRET (Attanasio et al., 2023) can be used to derive feature attribution scores (§4.4). For encoder-decoder or decoder-only architectures, tools like INSEQ (Sarti et al., 2023) can generate such scores.

[9] Although POLYJUICE utilizes a relatively small model, GPT-2, for generating counterfactuals, we fairly consider it a suitable baseline for FITCF, since the deployed GPT-2 is **fine-tuned** on a counterfactual example dataset.

[10] Details on label distributions and example instances from the datasets used can be found in Appendix E.

[11] More details about deployed models and inference time are provided in Appendix F.

4.4 Feature Importance

FERRET (Attanasio et al., 2023) is a framework that provides post-hoc explanations for LLMs and can evaluate these explanations based on faithfulness and plausibility. We use FERRET to generate feature attribution scores, selecting the following feature importance methods f: *Gradient* (Simonyan et al., 2014), *LIME* (Ribeiro et al., 2016), *Integrated Gradients* (Sundararajan et al., 2017), and *SHAP* (Lundberg and Lee, 2017).

5 Evaluation

5.1 Automatic Evaluation

The generated counterfactuals are evaluated using the following three automated metrics.

Soft Label Flip Rate The Soft Label Flip Rate (SLFR) measures the frequency at which newly perturbed examples alter the original label to a different label (Ge et al., 2021; Nguyen et al., 2024; Bhattacharjee et al., 2024a). For a dataset with N instances, we calculate SLFR as follows:

$$SLFR = \frac{1}{N}\sum_{n=1}^{N}\mathbb{1}(y_k^{'} \neq y_k)$$

where $\mathbb{1}$ is the indicator function, y_k is the original label and $y_k^{'}$ is the predicted label after the perturbation. Note that we use the same LLM for both counterfactual generation and classification[12].

Perplexity Perplexity (PPL) is defined as the exponential of the average negative log-likelihood of a sequence. PPL can measure the naturalness of the text distribution and how fluently the model can output the next word given the previous words (Fan et al., 2018). Given a sequence $X = (x_0, x_1, \cdots, x_t)$, PPL of X is calculated as:

$$PPL(X) = \exp\left\{\frac{1}{t}\sum_{i}^{t}\log p_\theta(x_i|x_{<i})\right\}$$

Following Wang et al. (2023); Nguyen et al. (2024); Bhattacharjee et al. (2024b), we deploy `GPT-2` to calculate PPL in our experiments due to its proven effectiveness in capturing such text distributions.

Textual Similarity (TS) The counterfactual $\tilde{x}$ should be as similar as the original input x (Madaan et al., 2021), where lower distances indicate greater similarity. We use normalized word-level Levenshtein distances d to capture all edits, which is widely used by the research community (Ross et al., 2021; Treviso et al., 2023):

$$TS = \frac{1}{N}\sum_{i=1}^{N}\frac{d(x_i, \tilde{x}_i)}{|x_i|} \quad (6)$$

[12]The accuracy and error rate of the deployed LLMs, along with the prompt instruction used are provided in Appendix G.

5.2 Ablation Study

As illustrated in Figure 2, FITCF comprises three core components: *Important words*; *demonstrations*; and *label flip verification*. Accordingly, we conduct a comprehensive ablation study to evaluate the importance of each component individually. The experiments are conducted using `Qwen2.5-72B`, as `Qwen2.5-72B` particularly struggles to generate high-quality counterfactual examples compared to `Llama3-8B` and `Qwen2.5-32B` (Table 1, Table 3).

5.2.1 Effect of Important Words

To assess the contribution of important words identified by `BERT` using different feature importance methods to counterfactual generation, we conduct the experiment using FITCF omitting any pre-identified important words.

5.2.2 Effect of Number of Demonstrations

In FITCF, as c clusters are obtained through clustering, and due to the difficulty and complexity of counterfactual example generation, we set the number of demonstrations to twice the number of clusters ($2c$) for each dataset (§3.2; Figure 4), which results in 10 demonstrations for AG News and 8 for SST2, respectively. To examine the effect of the number of demonstrations and assess the necessity of doubling the number of demonstrations to $2c$, we further evaluate the quality of counterfactual examples generated by FITCF, with the number of demonstrations set to the number of clusters (c).

5.2.3 Effect of Label Flip Verification

To ensure the validity of the selected demonstrations and prevent incorrect examples from misleading the LLM (Rubin et al., 2022; Turpin et al., 2023), FITCF incorporates a label flip verifier (§3.2). This verifier is implemented using a fine-tuned `BERT` model (Table 6) trained on the target dataset. To assess the impact of label flip verification, we conduct an ablation study by excluding label flip verification for comparative analysis.

Model	Approach	Method	AG News ($PPL = 95.72$) SLFR ↑	PPL ↓	TS ↓	SST2 ($PPL = 309.53$) SLFR ↑	PPL ↓	TS ↓
GPT2	POLYJUICE	-	18.60%	121.76	0.50	29.00%	258.32	0.71
Llama3-8B	FIZLE	-	93.50%	123.67	0.61	95.50%	202.22	0.52
	ZEROCF	Gradient	93.50%	102.56	0.38	97.50%	239.15	0.46
	ZEROCF	IG	95.50%	109.09	0.27	**99.50%**	222.51	**0.42**
	ZEROCF	LIME	97.50%	107.72	0.39	97.00%	264.91	0.42
	ZEROCF	SHAP	**98.00%**	**99.08**	**0.27**	94.00%	**204.76**	0.46
	FITCF	Gradient	94.50%	86.90	0.21	99.80%	159.57	**0.47**
	FITCF	IG	**96.00%**	87.67	0.23	100.00%	161.88	0.48
	FITCF	LIME	95.50%	**75.15**	**0.19**	**100.00%**	**151.22**	0.48
	FITCF	SHAP	94.00%	260.57	0.21	100.00%	157.36	0.49
Qwen2.5-32B	FIZLE	-	49.00%	53.07	1.14	86.80%	167.51	0.66
	ZEROCF	Gradient	**68.00%**	62.63	2.10	70.50%	205.06	**0.48**
	ZEROCF	IG	51.00%	**60.45**	**0.76**	91.00%	222.57	0.64
	ZEROCF	LIME	56.00%	63.75	0.84	90.50%	576.59	0.62
	ZEROCF	SHAP	55.50%	61.68	0.79	**93.00%**	**191.00**	0.60
	FITCF	Gradient	56.00%	62.97	0.73	89.00%	214.25	0.51
	FITCF	IG	57.50%	**57.01**	**0.68**	**90.50%**	221.64	**0.49**
	FITCF	LIME	56.00%	57.45	0.79	89.50%	174.34	0.52
	FITCF	SHAP	**62.00%**	57.64	0.78	89.50%	**157.09**	0.52
Qwen2.5-72B	FIZLE	-	21.50%	84.09	0.22	92.00%	257.91	0.43
	ZEROCF	Gradient	16.67%	74.19	**0.21**	88.50%	263.47	0.34
	ZEROCF	IG	24.50%	92.47	0.22	**92.00%**	**281.10**	0.46
	ZEROCF	LIME	23.00%	**72.73**	0.71	85.00%	289.20	0.30
	ZEROCF	SHAP	**25.00%**	73.92	0.74	86.50%	319.60	**0.22**
	FITCF	Gradient	**77.00%**	62.13	0.99	96.00%	595.71	**0.38**
	FITCF	IG	42.00%	63.54	**0.33**	95.00%	**207.55**	0.39
	FITCF	LIME	45.00%	**61.54**	0.35	**96.50%**	240.94	0.41
	FITCF	SHAP	38.96%	67.28	0.34	96.50%	590.94	0.39

Table 1: Automatic evaluation results of counterfactuals generated by FIZLE, ZEROCF, and FITCF with `Llama3-8B`, `Qwen2.5-32B`, and `Qwen2.5-72B` using Soft Label Flip Rate (SLFR), Perplexity (PPL), and Textual Similarity (TS) on AG News and SST2. Bold faced values indicate for each approach, which feature importance method is the best performing according to the respective metric.

5.3 Correlation Analysis

As we deploy various feature importance methods to generate counterfactuals synergistically (Figure 2), which can then be applied as demonstrations in FITCF, we investigate the correlation between the quality of the feature attribution scores and the quality of generated counterfactuals. The feature attribution scores are evaluated based on faithfulness using FERRET (Attanasio et al., 2023). For faithfulness evaluation, we employ three metrics: *comprehensiveness*, *sufficiency* (DeYoung et al., 2020) and *Kendall's τ correlation with Leave-One-Out token removal* (Jain and Wallace, 2019).

6 Results

6.1 Automatic Evaluation

Table 1 demonstrates that our proposed approaches, ZEROCF and FITCF, consistently outperform POLYJUICE easily, which exhibits relatively low SLFR. For AG News dataset using `Qwen2.5-32B`, the edit distance is comparatively higher than that of POLYJUICE, and the other baseline, FIZLE, also shows a larger edit distance compared to POLYJUICE. For SST2 dataset, `Qwen2.5-72B` tends to generate counterfactuals that are less natural and fluent when leveraging ZEROCF and FITCF. Interestingly, `Llama3-8B`, the smallest model among all evaluated LLMs, achieves the best overall performance. In contrast, `Qwen2.5-72B` generally underperforms compared to both `Llama3-8B` and `Qwen2.5-32B`, as `Qwen2.5-72B` has a stronger capability to discern the underlying context, making it less prone to flipping labels (App. D, Table 10).

Additionally, we observe that ZEROCF does not outperform FIZLE in some cases, e.g., with `Qwen2.5-72B` on SST2 dataset. However, in most cases, ZEROCF offers noticeable advantages in enhancing the quality of counterfactuals compared to FIZLE. Furthermore, we find that *Integrated Gradients* and *SHAP* contribute more positively to the

Dataset	Method	SLFR	PPL	TS
AG News	Gradient	41.50% (↓35.50%)	67.85 (↓5.72)	0.36 (↑0.63)
	IG	37.50% (↓4.50%)	67.85 (↓4.31)	0.37 (↑0.62)
	LIME	40.68% (↓4.32%)	66.08 (↓2.54)	0.35 (↑0.02)
	SHAP	37.00% (↓1.96%)	84.14 (↓16.86)	0.51 (↓0.17)
SST2	Gradient	93.50% (↓2.50%)	214.27 (↑381.44)	0.42 (↓0.04)
	IG	95.00% (- 0.00%)	214.27 (↓6.72)	0.42 (↓0.02)
	LIME	95.50% (↓1.00%)	278.78 (↓37.84)	0.41 (-0.00)
	SHAP	96.00% (↓0.50%)	290.57 (↑-300.37)	0.43 (↓0.04)

Table 2: Automatic evaluation results of counterfactuals generated by FitCF using `Qwen2.5-72B`, with demonstrations generated by ZeroCF without specifying *important words*.

Dataset	Method	SLFR	PPL	TS
AG News	Gradient	13.50% (↓63.50%)	66.74 (↓4.61)	0.27 (↑0.72)
	IG	15.50% (↓22.00 %)	64.28 (↓0.74)	0.27 (↑0.06)
	LIME	18.00% (↓27.00%)	68.28 (↓6.74)	0.27 (↓0.08)
	SHAP	14.00% (↓24.96%)	64.06 (↑3.22)	0.28 (↑0.06)
SST2	Gradient	89.00% (↓7.00%)	235.08 (↑360.63)	0.36 (↑0.02)
	IG	93.50% (↓1.50 %)	266.09 (↓58.54)	0.39 (-0.00)
	LIME	91.50% (↓5.00%)	250.70 (↓9.76)	0.39 (↑0.02)
	SHAP	92.00% (↓4.50%)	583.42 (↑7.52)	0.38 (↑0.01)

Table 3: Automatic evaluation results of counterfactuals generated by FitCF with `Qwen2.5-72B` using c demonstrations.

quality of counterfactuals, on average[13], compared to other feature importance methods.

Importantly, FitCF emerges as the most effective method for generating high-quality counterfactuals, consistently outperforming both baselines and ZeroCF across all evaluated settings, underscoring its robustness and effectiveness. This demonstrates the advantage of combining feature importance with the counterfactual generation process. Under the FitCF framework, *Integrated Gradients* and *LIME* illustrate superior performance in generating counterfactuals compared to the other two approaches.

6.2 Ablation Study

The results of the ablation studies are presented in Table 2, 3, 4, where for PPL and TS, an upward (*downward*) arrow signifies that a decrease (*increase*) in the value corresponds to an improvement (*deterioration*) in both metrics.

6.2.1 Effect of Important Words

Table 2 shows that for AG News, SLFR decreases across all methods, with the most significant decline observed when using *Gradient*. Concurrently, PPL improves and edit distances generally increases, suggesting that the generated counterfactuals diverge more from the original text, except when using *SHAP*. In contrast, for SST2, SLFR remains consistently high, with slight decreases. PPL exhibited mixed results, with both notable increases and decreases depending on the method, reflecting variability in fluency. Meanwhile, edit distance either decreases or remains unchanged. Overall, FitCF with *SHAP* demonstrates the highest robustness when important words are not specified, whereas *Gradient* is particularly sensitive to the inclusion of important words.

[13] We do not consider the number of times a feature importance method achieves the maximum value in tables, but rather the average ranking of a method across all datasets.

Dataset	Method	SLFR	PPL	TS
AG News	Gradient	34.00% (↓43.00%)	63.27 (↓1.14)	0.33 (↑0.66)
	IG	40.50% (↓1.50%)	64.65 (↓1.11)	0.35 (↓0.02)
	LIME	42.50% (↓2.50%)	65.23 (↓3.69)	0.35 (- 0.00)
	SHAP	34.00% (↓4.96%)	65.30 (↑1.98)	0.34 (- 0.00)
SST2	Gradient	94.50% (↓1.50%)	222.52 (↑373.19)	0.36 (↑0.02)
	IG	94.50% (↓2.00%)	240.11 (↓32.56)	0.39 (- 0.00)
	LIME	96.00% (↓0.50%)	245.79 (↓4.85)	0.40 (↑0.01)
	SHAP	94.50% (↓2.00%)	281.65 (↑309.29)	0.38 (↑0.01)

Table 4: Automatic evaluation results of counterfactuals generated by FitCF using `Qwen2.5-72B`, without *label flip verification*.

6.2.2 Effect of Number of Demonstrations

As shown in Table 3, we find that the number of demonstrations plays an critical role in the performance of FitCF. For AG News, SLFR declines precipitously when the number of clusters (c) is used as the number of demonstrations (§5.2.2), while the edit distance shows a slight improvement. In comparison, for SST2, the degree of SLFR diminishment is less conspicuous.

Furthermore, Table 3 reveals that in general, FitCF with *Integrated Gradients* and *SHAP* exhibits greater robustness compared to *Gradient* and *LIME*. In particular, FitCF with *Gradient* demonstrates the highest sensitivity, with a strong decline in quality as the number of demonstrations decreases.

6.2.3 Effect of Label Flip Verification

Table 4 divulges trends similar to those observed in Table 2 (§6.2.1). Omitting label flip verification leads to decreases in SLFR across both datasets, highlighting the importance of this step. However, skipping label flip verification occasionally results in lower PPL for certain methods, suggesting improved fluency in some cases.

Meanwhile, the decrease in SLFR is more pronounced for AG News, particularly with the *Gradient* method, which shows the largest SLFR drop alongside increases in PPL. Conversely, *Integrated Gradients* and *LIME* present minimal impact on SLFR, indicating a relative reliance on label flip verification to maintain consistent performance.

Model	Dataset / Method	AG News comp.	AG News suff.	AG News τ (loo)	SST2 comp.	SST2 suff.	SST2 τ (loo)
Llama3	Gradient	0.20	0.13	0.06	0.21	0.25	-0.03
	IG	0.38	0.03	0.07	-0.52	0.05	0.22
	LIME	0.61	-0.02	0.16	0.68	0.02	0.29
	SHAP	0.62	-0.02	0.16	0.60	0.03	0.25
Qwen-32B	Gradient	0.12	0.12	0.07	0.20	0.23	-0.03
	IG	0.32	0.03	0.05	0.50	0.04	0.21
	LIME	0.53	-0.01	0.12	0.67	0.01	0.29
	SHAP	0.53	-0.01	0.08	0.59	0.02	0.25
Qwen-72B	Gradient	0.12	0.12	0.07	0.20	0.23	-0.03
	IG	0.32	0.03	0.05	0.50	0.04	0.21
	LIME	0.53	-0.01	0.12	0.67	0.01	0.29
	SHAP	0.53	-0.01	0.07	0.59	0.02	0.25

Table 5: Faithfulness evaluation results based on *Comprehensiveness* (comp.), *Sufficiency* (suff.) and *Kendall's τ correlation with Leave-One-Out token removal* (τ (loo)) for counterfactuals generated by FITCF using Llama3-8B, Qwen2.5-32B, and Qwen2.5-72B on AG News and SST2 datasets.

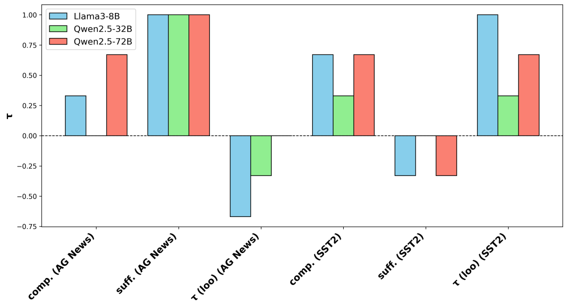


Figure 3: A Kendall's tau (τ) that quantifies the degree of correspondence between the ranking of generated counterfactuals' *quality* and the ranking of *feature attribution evaluation results* is reported.

6.3 Discussion

Important words identified through feature attribution scores for BERT are more effective and less prone to hallucination for counterfactual generation compared to those self-generated by LLMs. Through ablation studies on the three core components of FITCF, we conclude that the number of demonstrations generated by ZEROCF has the most significant impact on the performance of FITCF. While specifying important words and applying label flip verification also contribute to FITCF's effectiveness, their influence is less marked compared to the number of demonstrations. While SLFR decreases across three tables, the edit distance gets improved overall, except for SST, where no important words are specified. This indicates that without a certain component, the counterfactuals generated by FITCF are generally less edited, resulting in less successful label flips. Moreover, FITCF with *Gradient* proves to be the least robust, showing substantial drops in SLFR, when any of the three components is removed. In contrast, FITCF with *LIME* and *SHAP* demonstrate greater robustness and consistently produce high-quality counterfactuals.

6.4 Correlation Analysis

From Table 5, we discover that *LIME* and *SHAP* consistently outperform *Gradient* and *Integrated Gradients* in terms of comprehensiveness and τ (loo) across all models and datasets, which aligns with our findings in §6.3. In addition, the comprehensiveness and sufficiency scores exhibit less variation across three models for AG News, though they are generally lower than those for SST2. In contrast, τ (loo) scores for SST2 are slightly higher compared to AG News. Furthermore, for AG News, a strong correlation ($\tau = 1$) is observed in Figure 3 between the quality of generated counterfactuals and sufficiency, while for SST2, both comprehensive and τ (loo) demonstrate notable correlations with counterfactual quality. We conclude that the faithfulness of feature attribution scores is generally strongly correlated with the quality of counterfactuals generated with the auxiliary assistance of extracted important words using FITCF.

7 Conclusion

We first introduced ZEROCF, an approach that leverages important words derived from feature attribution methods for counterfactual example generation in a zero-shot setting. Building on this, we proposed FITCF, a framework that automatically constructs high-quality demonstrations using ZEROCF, eliminating the need for human-annotated ground truth for counterfactual generation. FITCF validates counterfactuals via label flip verification for their suitability as demonstrations in a few-shot setting. Empirically, FITCF outperforms two baselines POLYJUICE and FIZLE, and our own ZEROCF. Through ablation studies, we identified the three core components of FITCF - number of demonstrations, important words, and label flip verification - as critical to enhancing counterfactual quality. Moreover, we evaluated the faithfulness of feature attribution scores and found that *LIME* and *Integrated Gradients* are the most effective feature importance methods for FITCF, consistently producing the most faithful feature attribution scores. Finally, our analysis revealed a strong correlation between the faithfulness of feature attribution scores and the quality of the generated counterfactuals.

Limitations

We conducted experiments exclusively using datasets in English. In other languages, the current approach may not offer the same advantages.

The deployed BERT models perform well on fine-tuned tasks (Table 6). However, the LLMs used are not as effective as classifiers compared to BERT models (Table 10) (Shin et al., 2020). The quality of the generated counterfactual examples may be affected by the fact that, given an instance, LLMs perceive the label as flipped, even though the actual label is not flipped.

In ZeroCF and FitCF, feature attribution scores are determined by an explanation method for the predictions of a BERT model fine-tuned on the target dataset and the same BERT model is used to verify label flips. The potential contribution of other language models to performing both tasks in ZeroCF and FitCF, however, remains unexplored.

Future work includes investigating the correlation between additional dimensions of feature attribution scores, such as *plausibility*, *coherence* and *insightfulness*, and the quality of counterfactuals through user studies (Domnich et al., 2024). We also plan to explore the potential of language models with architectures beyond encoder-only models as a foundation for feature attributions to be used in ZeroCF and FitCF.

References

AI@Meta. 2024. Llama 3 model card.

Shengnan An, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Jian-Guang Lou, and Dongmei Zhang. 2023. How do in-context examples affect compositional generalization? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11027–11052, Toronto, Canada. Association for Computational Linguistics.

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*. Association for Computational Linguistics.

Tanja Baeumel, Soniya Vijayakumar, Josef van Genabith, Guenter Neumann, and Simon Ostermann. 2023. Investigating the encoding of words in BERT's neurons using feature textualization. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 261–270, Singapore. Association for Computational Linguistics.

Milan Bhan, Jean-noel Vittaut, Nicolas Chesneau, and Marie-jeanne Lesot. 2023. Enhancing textual counterfactual explanation intelligibility through counterfactual feature importance. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 221–231, Toronto, Canada. Association for Computational Linguistics.

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

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2024b. Zero-shot llm-guided counterfactual generation for text. *Preprint*, arXiv:2405.04793.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2023. DISCO: Distilling counterfactuals with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528, Toronto, Canada. Association for Computational Linguistics.

Julien Delaunay, Luis Galárraga, and Christine Largouët. 2024. Does it make sense to explain a black box with another black box? *Preprint*, arXiv:2404.14943.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Marharyta Domnich, Julius Valja, Rasmus Moorits Veski, Giacomo Magnifico, Kadi Tulver, Eduard Barbu, and Raul Vicente. 2024. Towards unifying evaluation of counterfactual explanations: Leveraging large language models for human-centric assessments. *Preprint*, arXiv:2410.21131.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Yingqiang Ge, Shuchang Liu, Zelong Li, Shuyuan Xu, Shijie Geng, Yunqi Li, Juntao Tan, Fei Sun, and Yongfeng Zhang. 2021. Counterfactual evaluation for explainable ai. *Preprint*, arXiv:2109.01962.

Gilad Gressel, Niranjan Hegde, Archana Sreekumar, Rishikumar Radhakrishnan, Kalyani Harikumar, Anjali S., and Krishnashree Achuthan. 2023. Feature importance guided attack: A model agnostic adversarial attack. *Preprint*, arXiv:2106.14815.

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*.

Satyapriya Krishna, Jiaqi Ma, Dylan Slack, Asma Ghandeharioun, Sameer Singh, and Himabindu Lakkaraju. 2023. Post hoc explanations of language models can improve language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 65468–65483. Curran Associates, Inc.

Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. 2023. A survey of large language models attribution. *Preprint*, arXiv:2311.03731.

Yongqi Li, Mayi Xu, Xin Miao, Shen Zhou, and Tieyun Qian. 2024. Prompting large language models for counterfactual generation: An empirical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13201–13221, Torino, Italia. ELRA and ICCL.

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.

Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. 2021. Generate your counterfactuals: Towards controlled counterfactual generation for text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13516–13524.

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

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Van Bach Nguyen, Paul Youssef, Christin Seifert, and Jörg Schlötterer. 2024. 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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

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.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Rachneet Sachdeva, Martin Tutek, and Iryna Gurevych. 2024. CATfOOD: Counterfactual augmented training for improving out-of-domain performance and calibration. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1876–1898, St. Julian's, Malta. Association for Computational Linguistics.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, Oskar van der Wal, Malvina Nissim, and Arianna Bisazza. 2023. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 421–435, Toronto, Canada. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

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.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Qianli Wang, Tatiana Anikina, Nils Feldhus, Josef Genabith, Leonhard Hennig, and Sebastian Möller. 2024. LLMCheckup: Conversational examination of large language models via interpretability tools and self-explanations. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 89–104, Mexico City, Mexico. Association for Computational Linguistics.

Yequan Wang, Jiawen Deng, Aixin Sun, and Xuying Meng. 2023. Perplexity from plm is unreliable for evaluating text quality. *Preprint*, arXiv:2210.05892.

Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. 2021. Feature importance-aware transferable adversarial attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7639–7648.

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.

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.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

A Prompt Instruction

A.1 Prompt for ZEROCF

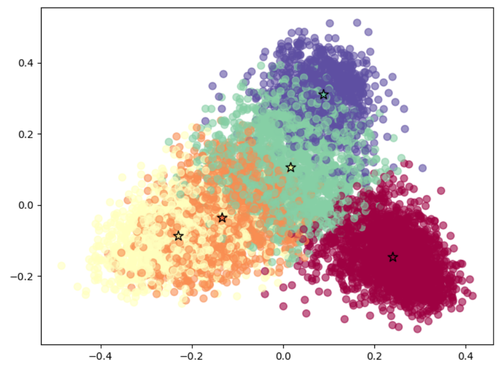
```
You are an excellent assistant for text
editing. You are given an input from the
{dataset} dataset, classified into one of
{len(labels)} categories:
{', '.join(labels)}. The input belongs to
the '{prediction}' category.
{important_words} might be important words
leading to the '{prediction}' category.

Your task is to make minimal changes on the
below provided input to alter the
prediction category by carefully
considering provided important words.
Please output only the edited input.

Input: {input_text}
```

A.2 Prompt for FITCF

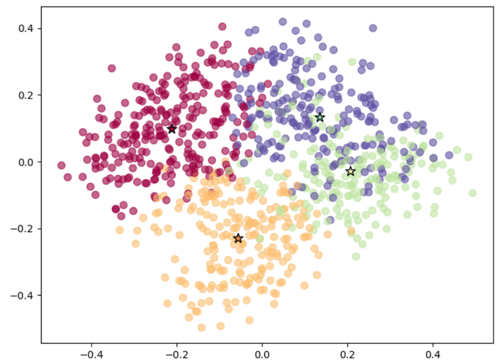
```
You are an excellent assistant for text
editing. You are given an input from the
{dataset} dataset, classified into one of
{len(labels)} categories:
{', '.join(labels)}. The input belongs to
the '{prediction}' category.
{important_words} might be important words
leading to the '{prediction}' category.

Your task is to make minimal changes on the
input provided below to alter the
prediction category to '{counterpart}' by
carefully considering provided important
words and examples. Please output the
edited input only!

Below are some examples consisting of
original and edited input.

[original input] {original_input_1}
[edit input] {edit_input_1}
...
[original input] {input_text}
[edit input]
```

B Detailed Information of Deployed BERT

Table 6 displays BERT models used for AG News and SST2 datasets with their validation accuracies. As both BERT models demonstrate strong performance in accuracy, we can use them as classifiers (§3.1) and label flip verifiers (§3.2).



(a) AG News



(b) SST2

Figure 4: Visualization of clustering in AG News and SST2, where stars denote cluster centroids.

C Visualization of Clustering

Figure 4 visualizes the clustering of sentence embeddings from AG News, and SST2 datasets, with their dimensions reduced to two using PCA. The illustrations suggest that generic patterns already exist, with instances from various clusters contributing to these patterns.

D Demonstration Selection by FITCF

Table 7 shows the most similar demonstrations selected from each cluster, as shown in Figure 4 for the question *"Rivals Try to Turn Tables on Charles Schwab By MICHAEL LIEDTKE SAN FRANCISCO (AP) – With its low prices and iconoclastic attitude, discount stock broker Charles Schwab Corp. (SCH) represented an annoying stone in Wall Street's wing-tipped shoes for decades..."* from AG News.

The decrease in SLFR performance while using a strong LLM can be attributed to the ad-

Dataset	Model	Accuracy	Link
AG News	textattack/bert-base-uncased-ag-news	93.03%	https://huggingface.co/textattack/bert-base-uncased-ag-news
SST2	gchhablani/bert-base-cased-finetuned-sst2	92.32%	https://huggingface.co/gchhablani/bert-base-cased-finetuned-sst2

Table 6: BERT models used for AG News and SST2 datasets, with accuracy validated on their respective testsets.

Text	Counterfactual
Bovina ends two-year wait. Seventh-seeded Russian Elena Bovina won her first title in two years by beating France's Nathalie Dechy 6-2 2-6 7-5 in the final of the **Pilot Pen tournament**.	Bovina ends two-year wait. Seventh-seeded Russian Elena Bovina won her first title in two years by beating France's Nathalie Dechy 6-2 2-6 7-5 in the final of the International Event .
Wall St.'s Nest Egg - the Housing Sector NEW YORK (Reuters) - If there were any doubts that we're still living in the era of the stay-at-home economy, the rows of empty seats at the Athens Olympics should help erase them.	The Olympics - the Housing Sector NEW YORK (Reuters) - If there were any doubts that we're still living in the era of the stay-at-home economy, the rows of empty seats at the Athens Olympics should help erase them.
French Take Gold, Bronze in Single **Kayak** ATHENS, Greece - Winning on whitewater runs in the family for Frenchman Benoit Peschier, though an Olympic gold is something new. Peschier paddled his one-man kayak aggressively but penalty free in both his semifinal and final runs on the man-made **Olympic** ...	French Take Gold, Bronze in Single Kayaking Competition ATHENS, Greece - Winning on whitewater runs in the family for Frenchman Benoit Peschier, though an Olympic gold is something new. Peschier paddled his one-man kayak aggressively but without penalty in both his semifinal and final runs on the man-made Olympic course .
Japanese Utility Plans IPO in October (AP) AP - Electric Power Development Co., a former state-run utility, **said Friday it** is planning an initial public offering on the Tokyo Stock Exchange in October, a deal that could be the country's biggest new stock listing in six years.	Electric Power Development Co., a former state-run utility, is planning an initial public offering on the Tokyo Stock Exchange in October, a deal that could be the country's biggest new stock listing in six years.
Afghan women make brief **Olympic** debut Afghan women **made a short-lived** debut in the Olympic Games on Wednesday as 18-year-old **judo wild-card** Friba Razayee was defeated after 45 seconds of her first match in the under-70kg middleweight.	Afghan women make brief debut in international relations as 18-year-old Friba Razayee was defeated after 45 seconds of her first match in the under-70kg middleweight.

Table 7: The most similar demonstrations selected from each cluster for the question *"Rivals Try to Turn Tables on Charles Schwab By MICHAEL LIEDTKE SAN FRANCISCO (AP) – With its low prices and iconoclastic attitude, discount stock broker Charles Schwab Corp. (SCH) represented an annoying stone in Wall Street's wing-tipped shoes for decades..."* from AG News. Corresponding counterfactuals are generated by Qwen2.5-72B using ZEROCF. Differences are marked in **bold** and edits are highlighted in red .

vanced contextual understanding of such models, e.g., Qwen2.5-72B. These models are more adept at discerning the underlying context of inputs and therefore less likely to incorrectly flip labels. For instance, as shown in Table 7, the second example remains clearly related to **business**, as the main topic—Housing Sector—is still evident, even though "*Wall St.'s Nest Egg*" is replaced with "*The Olympic*".

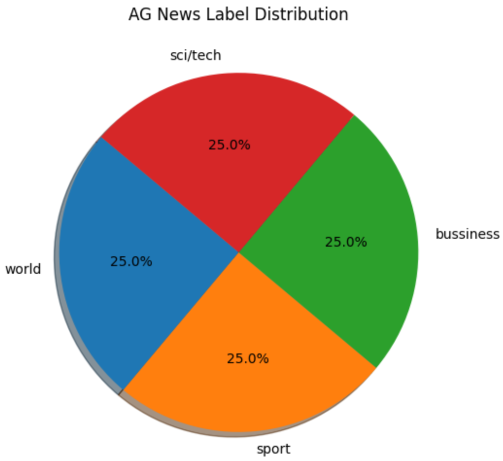
E Dataset

E.1 Label Distribution

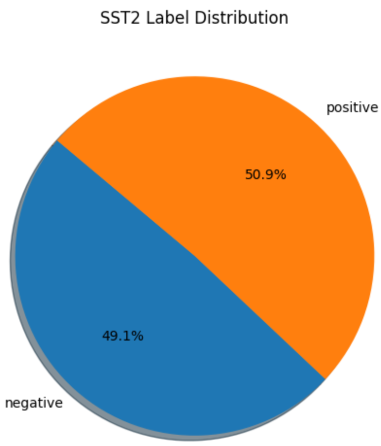
Figure 5 shows the label distributions of AG News and SST2 validation sets.

E.2 Dataset Example

Figure 6 demonstrates example instances and gold labels from AG News and SST2 datasets.



(a) AG News



(b) SST2

Figure 5: Label distribution of AG News and SST2.

Figure 6: Example instances from AG News and SST2.

F Experiment

F.1 Models

Table 8 demonstrates LLMs that are used for ZEROCF and FITCF. To reduce memory consumption, we use a GPTQ-quantized version (Frantar et al., 2023). All LLMs are directly downloaded from Huggingface and run on a single NVIDIA RTXA6000, A100 or H100 GPU.

F.2 Inference Time

Table 9 shows inference time for ZEROCF and FITCF using `Llama3-8B`, `Qwen2.5-32B` and `Qwen2.5-32B` on AG News and SST2.

G Calculation of Label Flip Rate

We use the same LLM to serve as both the flip label verifier and the counterfactual generator (§5.1). To validate deployed LLMs' classification performance, we evaluate them on the AG News and SST2 datasets. Subsequently, we detail the prompt instructions used for flip label verification.

G.1 Classification Performance of LLMs

Table 10 displays the accuracy score and error rate on AG News and SST2 datasets using `Llama3-8B`, `Qwen2.5-32B`, and `Qwen2.5-72B`. Our findings indicate that `Qwen2.5-32B` demonstrates the best classification performance with the lowest error rate, whereas `Llama3-8B` has the poorest classification performance. Notably, `Qwen2.5-72B` is the only LLM that generates predictions outside the predefined labels on SST2.

G.2 Prompt Instruction

```
You are an excellent assistant for text
classification. You are provided with an
```

Name	Citation	Size	Link
Llama3	AI@Meta (2024)	8B	https://huggingface.co/meta-llama/Meta-Llama-3-8B
Qwen2.5	Team (2024)	32B	https://huggingface.co/Qwen/Qwen2.5-32B-Instruct-GPTQ-Int4
Qwen2.5	Team (2024)	72B	https://huggingface.co/Qwen/Qwen2.5-72B-Instruct-GPTQ-Int4

Table 8: Three open sourced LLMs used in ZeroCF and FitCF.

	AG News		SST2	
	ZeroCF	FitCF	ZeroCF	FitCF
Llama3-8B	8h	13h	2h	5h
Qwen2.5-32B	9h	17h	7h	12h
Qwen2.5-72B	38h	47h	8h	16h

Table 9: Inference time for ZeroCF and FitCF using Llama3-8B, Qwen2.5-32B and Qwen2.5-32B on AG News and SST2.

```
original and an edited instance from the
{dataset_name} dataset. Each instance
belongs to one of {len(labels)} categories:
{', '.join(labels)}. Determine if the
predicted classifications of the original
and edited instances are different.
[original instance] '{instance}'
[edited instance] '{counterfactual}'
Respond with 'yes' if they are different.
Response with 'no' if they are the same.
Answer 'yes' or 'no' only!
```

Dataset	**Model**	**Accuracy**	**Error Rate**
AG News	Llama3-8B	<u>72.39%</u>	<u>0.70%</u>
	Qwen2.5-32B	**80.73%**	**0.28%**
	Qwen2.5-72B	79.12%	0.47%
SST2	Llama3-8B	<u>89.75%</u>	0.00%
	Qwen2.5-32B	**94.61%**	**0.00%**
	Qwen2.5-72B	94.27%	<u>0.11%</u>

Table 10: Accuracy score and error rate on AG News and SST2 datasets across three runs on the validation set using `Llama3-8B`, `Qwen2.5-32B`, and `Qwen2.5-72B` in a *zero-shot* setting. The error rate is calculated by counting the number of instances where the predicted label falls outside the pre-defined label set.