LLMs for Generating and Evaluating Counterfactuals: A Comprehensive Study

Van Bach Nguyen[†*] **Paul Youssef**[†*] **Christin Seifert**[†] **Jörg Schlötterer**[†‡]
[†]University of Marburg, [‡]University of Mannheim
{vanbach.nguyen, paul.youssef, christin.seifert, joerg.schloetterer}@uni-marburg.de

Abstract

As NLP models become more complex, understanding their decisions becomes more crucial. Counterfactuals (CFs), where minimal changes to inputs flip a model's prediction, offer a way to explain these models. While Large Language Models (LLMs) have shown remarkable performance in NLP tasks, their efficacy in generating high-quality CFs remains uncertain. This work fills this gap by investigating how well LLMs generate CFs for three tasks. We conduct a comprehensive comparison of several common LLMs, and evaluate their CFs, assessing both intrinsic metrics, and the impact of these CFs on data augmentation. Moreover, we analyze differences between human and LLM-generated CFs, providing insights for future research directions. Our results show that LLMs generate fluent CFs, but struggle to keep the induced changes minimal. Generating CFs for Sentiment Analysis (SA) is less challenging than NLI and Hate Speech (HS) where LLMs show weaknesses in generating CFs that flip the original label. This also reflects on the data augmentation performance, where we observe a large gap between augmenting with human and LLM CFs. Furthermore, we evaluate LLMs' ability to assess CFs in a mislabelled data setting, and show that they have a strong bias towards agreeing with the provided labels. GPT4 is more robust against this bias, but it shows strong preference to its own generations. Our analysis suggests that safety training is causing GPT4 to prefer its generations, since these generations do not contain harmful content. Our findings reveal several limitations and point to potential future work directions.

1 Introduction

The growing popularity of artificial intelligence (AI) and increasingly complex "black-box" models have triggered a critical need for interpretability. As Miller (2019) highlights, explanations often seek to understand why an event P occurred instead of an alternative Q. Ideally, explanations should demonstrate how minimal changes to an instance could have led to different outcomes. In the context of textual data, this translates to introducing minimal modifications to the text through word additions, replacements, or deletions, to flip the label assigned by a given classifier. Counterfactual generation in NLP aims to foster an understanding of models, thereby facilitating their improvement (Kaushik et al., 2020), debugging (Ross et al., 2021), or rectification (Balashankar et al., 2023).

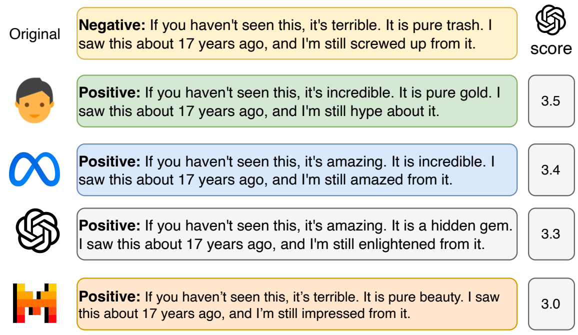

Figure 1: Counterfactual for Sentiment Analysis from several LLMs with their evaluation scores from GPT4.

In the field of NLP, LLMs have consistently demonstrated remarkable performance across diverse tasks. However, despite significant advancements in counterfactual generation methods, the efficacy of LLMs in producing high-quality counterfactuals (CFs) remains an open question. Our study bridges this gap by rigorously assessing the inherent capability of LLMs to generate CFs and identifying the most effective ones. We conduct a comprehensive comparison of several common LLMs, spanning different sizes and accessibility levels, evaluating their performance specifically on the counterfactual generation task. Our assessment encompasses standard metrics for CFs quality, as well as an in-depth evaluation of language fluency tailored to the context of counterfactual generation.

[*]Equal contribution

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 14809–14824
November 12-16, 2024 ©2024 Association for Computational Linguistics

Furthermore, we extend our analysis to data augmentation. We consider generating CFs for 3 tasks in this study: Sentiment Analysis (SA), Natural Language Inference (NLI), and Hate Speech (HS).

Our analysis demonstrates that LLMs are able to generate fluent text. However, they have difficulties in inducing minimal changes. Generating CFs for SA is less challenging than NLI and HS, where LLMs exhibit weaknesses in generating CFs that flip the labels. For data augmentation, SA CFs from LLMs can be an alternative to human CFs, as they are able to achieve similar performance, while on NLI and HS further improvements are needed. Furthermore, we show a positive correlation between keeping minimal changes and data augmentation performance. This suggests a new direction to generate improved data for augmentation, potentially leading to more efficient augmentation approaches.

We further assess the ability of LLMs to act as evaluators of CFs. We show a sample of CFs from different LLMs with the corresponding scores in Figure 1. By conducting controlled experiments, we show that LLMs have a strong bias to agree with the given labels, even if these labels are incorrect. GPT4 demonstrates strong preference to its own generations. Our analysis suggests that one reason for this preference is safety training, i.e., GPT4 prefers its own generations, because these generations do not contain any harmful content. Finally, to facilitate further research, we contribute a new dataset of CFs generated by various LLMs.[1]

2 Evaluation Methodology

We conduct a multi-faceted evaluation, considering several use cases where CFs could be beneficial.

2.1 Intrinsic Evaluation

Given a fixed classifier f and a dataset with N samples $(x_1, x_2, \ldots, x_N)$, $x = (z_1, z_2, \ldots, z_n)$ represents a sequence of n tokens with a ground truth label y. A valid counterfactual x' should: (1) achieve the desired target label y' with (2) minimal changes, and (3) align with likely feature distributions (Molnar, 2022). To evaluate these three desiderata, we consider the intrinsic properties of *Flip Rate*, *Textual Similarity*, and *Perplexity* as also suggested in a benchmark for counterfactual evaluation (Nguyen et al., 2024):

[1]https://github.com/aix-group/llms-for-cfs/

Flip Rate (FR): measures how effectively a method can change labels of instances with respect to a pretrained classifier. FR is defined as the percentage of generated instances where the labels are flipped over the total number of instances N (Bhattacharjee et al., 2024):

$$FR = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[f(x_i) = y']$$

Textual Similarity (TS): quantifies the closeness between an original instance and the counterfactual. Lower distances indicate greater similarity. We use the Levenshtein distance for d to quantify the token-distance between the original instance x and the counterfactual x'. This choice is motivated by the Levenshtein distance's ability to capture all type of edits (insertions, deletions, or substitutions) and also its widespread use in related work (Ross et al., 2021; Treviso et al., 2023):

$$TS = \frac{1}{N} \sum_{i=1}^{N} \frac{d(x_i, x_i')}{|x_i|}$$

Perplexity (PPL): To ensure that the generated text is plausible, realistic, and follows a natural text distribution, we leverage perplexity from GPT-2 because of its effectiveness in capturing such distributions. (Radford et al., 2019)[2]

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

2.2 Data Augmentation

After detecting failures in task-specific models, CFs can be used to augment the training data, and help close potential flaws in the reasoning of these models (Kaushik et al., 2020). Additionally, data augmentation with CFs increases generalization and OOD performance (Sen et al., 2021; Ding et al., 2024). In this evaluation, we examine how augmenting original training data with human and LLMs-generated CFs reflects on the performance of task-specific models.

2.3 LLMs for CFs Evaluation

Evaluation with LLMs has been shown to be a valid alternative to human evaluation on various

[2]While GPT-2 is used for simplicity in this study, any other LLM can be substituted as long as it demonstrates strong text generation capabilities

tasks like open-ended story generation and adversarial attacks (Chiang and Lee, 2023), open-ended questions (Zheng et al., 2023), translation (Kocmi and Federmann, 2023) and natural language generation (Liusie et al., 2024). In this work, we examine how well LLMs can evaluate CFs. Detecting mistakes in CFs with LLMs opens the door for iteratively refining CFs (Madaan et al., 2023).

For assessing LLMs in CFs evaluation, we leverage them to evaluate two sets of CFs. An *honest* set of CFs from humans, and a *corrupted* set, where we corrupt the ground truth labels. We compare the scores between the two sets and draw conclusions about the realiablity of LLMs for evaluating CFs.

3 Experimental Setup

3.1 Data

We compare CFs generated by LLMs against CFs generated by crowd workers (Kaushik et al., 2020) and experts (Gardner et al., 2020) (hereinafter referred to as "Human Crowd" and "Human Experts" respectively).

Sentiment Analysis (SA). We experiment with the IMDb dataset (Maas et al., 2011). For better comparability, we use the data splits from Kaushik et al. (2020).

Natural Language Inference (NLI). We experiment with SNLI (Bowman et al., 2015). Here too, we use the data splits from Kaushik et al. (2020).

Hate Speech (HS). We use the dataset from (Vidgen et al., 2021), which includes human CFs.

3.2 Generating Countefactuals

In order to make our study LLMs-focused and computationally feasible, we decided to generate counterfactual in a way that fulfills the following criteria:

- Generated CFs can be used for data augmentation (an evaluation aspect)
- Generating CFs does not require human intervention (e.g., specifying edits or labeling)
- Generating CFs does not require additional training in order to make the study computationally feasible
- The resulting CFs should depend only on the evaluated LLM in order to exclude any other confounding factors

To create the prompt for the LLMs to generate CFs, we combine two techniques: (1) Selecting the closest factual instance to the current instance (Liu et al., 2022). Since the provided example has a crucial effect on performance (Liu et al., 2022), we select the closest factual/counterfactual pair that has been generated by humans. We use Sentence-BERT (Reimers and Gurevych, 2019) to obtain the latent space representation, and then calculate the distance using cosine similarity from that latent space. (2) Chain-of-Thought (CoT) prompting (Wei et al., 2022), showing the necessary steps to generate a counterfactual instance based on a factual one, since it has been shown to help LLMs reason better and provide higher-quality answers. An overview of the process for generating CFs is depicted in Figure 2.

Specifically, we use the validation set in each dataset as a reference to select the closest example when generating CFs for the train and test sets. After obtaining the pair of closest instances, we apply CoT prompting by defining three steps to generate the counterfactual:

- Step 1: Identify all of the important words that contribute to flipping the label.
- Step 2: Find replacements for the words identified in Step 1 that lead to the target label.
- Step 3: Replace the words from Step 2 in the original text to obtain the counterfactual instance.

This prompt aligns with other work (Ross et al., 2021; Treviso et al., 2023; Li et al., 2024), which involve identifying significant words that impact the label and altering them to flip the label, thereby generating counterfactual instances. The prompt examples can be found in the Appendix E.

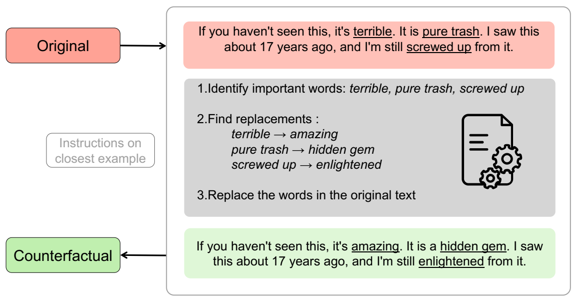

Figure 2: An overview of CFs generation process. Step-by-step instructions are shown on closest example.

3.3 LLMs

We compare open-source LLMs with closed-source LLMs. We choose LLAMA-2 (Touvron et al., 2023) and Mistral (Jiang et al., 2023) as representatives for open-source models, and GPT-3.5 and GPT-4[3] as representatives for closed-source LLMs. Table 1 summarizes the properties of each LLM.

Model	Size	HF	Instruct	OS
LLAMA2	7B/70B	✓	✓	✓
Mistral	7B/56B	✗	✓	✓
GPT3.5	-	✓	✓	✗
GPT4	-	✓	✓	✗

Table 1: Characteristics of Large Language Models (LLMs, including Size, Human Feedback (HF), Instruction, and Open-Source (OS).

4 Results and Discussion

4.1 Intrinsic Evaluation

We show the results for the intrinsic evaluation in Table 2. For flip rate, we use SOTA BERT-based models from (Morris et al., 2020) (SA and NLI) and (Vidgen et al., 2021) (HS).

The obtained perplexity values reflect the high fluency of LLMs, some of which are even more fluent than humans.[4]The perplexity of HS is significantly higher than that of other datasets due to the informal nature of tweets, where users often use slang, uncommon words, or elongated words for emphasis. Distance values show that LLMs do not necessarily adhere to conducting minimal changes. One exception here is GPT3.5, whose average distance values resemble that of human-generated CFs. The large distance values for LLM-generated CFs could be explained by their tendency to overgenerate (Guerreiro et al., 2023).

In terms of flip rate, we notice that some LLM-generated CFs can have a higher flip rate than human-generated CFs on SA, whereas the opposite can be observed on NLI. Meanwhile, LLM-generated CFs can reach moderate FR in HS. NLI CFs could be more difficult to generate than SA and HS CFs, which explains the gap in flip rate between LLMs and humans on the one hand, and GPT4 and other LLMs on the other hand (this is especially apparent on the *NLI - hypothesis*). This suggests that GPT4 should be the preferred choice to generate CFs for explaining a model's behavior. Furthermore, across all datasets, LLMs struggle to flip the label while keeping the changes minimal, i.e., they often need to make many modifications to flip the label. We examine the LLM-generated CFs in more detail in Section 4.4.

This part of the evaluation shows us that LLMs are able to generate fluent CFs, but struggle to induce minimal changes. It also demonstrates that it is challenging to generate NLI and HS CFs that flip the label, whereas generating SA CFs is less difficult.

4.2 Data Augmentation

We train on both original training data and CFs from different LLMs to see if augmenting the training data leads to an improved performance. For comparison, we conduct data augmentation with human CFs as well. The results for SA, NLI and HS are shown in Table 3, 4 and 5 respectively.

SA. On the crowd CFs and expert CFs test sets for SA including LLM-generated CFs lead to improved performance. LLAMA2 7B provide the most useful CFs for data augmentation, but other LLMs perform similarly. However, augmenting with human CFs works the best. On the original test set, augmenting with CFs does not improve performance. This shows that the gains in performance from data augmentation are visible only if the test set contains challenging examples.

NLI. On the *crowd premise* test set of NLI, which consists of CFs that were generated by changing the premise only, we notice that most of the LLM-generated CFs help improve the model's performance by a good margin (> 7 pp). The gap to augmenting with human CFs, however, is still large ($\sim$ 9 pp). On the *crowd hypothesis* test set, all LLMs lead to a lower performance. Here too, there is a large gap to human CFs ($\sim$ 16 pp). On the *original* test set, augmenting with LLM-generated CFs hurts performance, while augmenting with human-generated CFs bring good improvements ($\sim$ 5 pp). This shows how high-quality human CFs improve the model's capabilities, and points to a problem with LLM-generated CFs for NLI.

HS. Training with LLM-generated CFs does not bring substantial improvements on the CFs and the original test sets. Conversely, training with human CFs leads to significant improvements on both test

[3]We use API from https://openai.com/

[4]Note that the shown perplexity values are based on GPT-2

	SA			NLI - premise			NLI - hypothesis			Hate Speech		
	PPL ↓	TS ↓	FR ↑	PPL ↓	TS ↓	FR ↑	PPL ↓	TS ↓	FR ↑	PPL ↓	TS ↓	FR ↑
Human Experts	51.07	0.16	81.15	-	-	-	-	-	-	-	-	-
Human Crowd	48.03	0.14	85.66	74.89	0.17	59.13	65.67	0.19	79.75	229.05	0.31	87.39
GPT3.5	49.53	**0.16**	79.51	71.62	**0.15**	35.50	51.30	**0.19**	41.50	235.52	**0.16**	54.05
GPT4	49.05	0.29	94.03	73.39	0.28	**57.12**	58.35	0.21	**65.88**	**209.49**	0.49	**76.54**
LLAMA2 7B	46.99	0.64	78.26	70.34	0.36	41.02	59.60	0.28	38.64	-	-	-
LLAMA2 70B	**33.88**	1.37	93.48	**63.17**	0.21	41.07	58.54	0.23	46.62	-	-	-
Mistral 7B	48.55	1.06	95.13	78.34	0.36	37.71	**39.06**	0.46	44.11	365.15	0.69	67.41
Mistral 56B	35.63	0.57	**95.45**	65.37	0.23	27.46	57.65	0.21	31.55	401.63	0.56	70.30

Table 2: Metrics for intrinsic evaluation. **PPL** is perpelexity using GPT-2. **TS** is Levenshtein distance. **FR** is flip rate with respect to a SOTA classifier.[5]

sets. On this task too, LLM-generated CFs fall short of human CFs, indicating that there remains significant room for improvement.

Connection with intrinsic metrics. We examine the relation between data augmentation performance on the one hand and perplexity and Levenshtein distance on the other hand. The correlation values in Table 6 suggest that CFs with lower distance (to the factual instances) bring more improvements. Indeed, classifiers could be insensitive to small changes (Glockner et al., 2018), and having such examples in the training can make classifiers more robust. The negative correlation between accuracy and perplexity suggests that more fluent CFs are less effective in improving the classifier's performance. This indicates that classifiers primarily focus on the content rather than grammatical structure or coherence, especially in NLI tasks where the (factual) instances are mere image captions that are not necessarily fluent or grammatical texts

In summary, most LLMs produce CFs that come close to human CFs in terms of data augmentation performance on SA. On NLI and HS, the results are less positive: LLM-generated CFs bring no improvements in most cases, and the gap to human CFs is still large. CFs with less changes to the factual instances are more beneficial for data augmentation.

4.3 LLMs for CFs Evaluation

We examine how reliable are LLMs for CFs evaluation by asking them to evaluate two sets of human CFs: an *honest* set and a *corrupted* set. The "honest set" refers to a collection of human CFs, for which the ground truth labels are provided, whereas

[5]LLAMA-2 is unable to generate counterfactuals for HS due to its safety mechanism.

	Test Data		
	Crowd CFs	Expert CFs	Orig.
Original only	91.68 ± 1.07	86.31 ± 1.62	**90.20** ± 0.67
Human Crowd	**95.94** ± 0.37	**92.01** ± 1.09	89.63 ± 0.85
GPT3.5	94.55 ± 0.96	89.88 ± 1.47	89.30 ± 0.51
GPT4	93.52 ± 0.89	89.10 ± 0.76	**89.88** ± 0.57
LLAMA2 7B	**95.29** ± 0.72	**90.37** ± 1.57	88.89 ± 1.35
LLAMA2 70B	94.18 ± 0.27	88.89 ± 1.02	89.39 ± 0.44
Mistral 7B	93.93 ± 0.62	88.61 ± 1.68	89.22 ± 0.72
Mistral 56B	93.40 ± 1.02	88.20 ± 0.79	89.84 ± 0.79

Table 3: Data augmentation results for SA. Classification model is trained on original and LLMs or human-generated CFs with Accuracy as a metric.

the "corrupted set" consists of instances, for which wrong labels differing from the gold labels are provided. In the context of NLI, the third label, distinct from both the target and factual labels, is selected for inclusion in the corrupted set. For SA, the reverse label is chosen while the factual label remains undisclosed. Initially, we prompt GPT3.5 and GPT4 to assess whether the provided CFs accurately represent the target labels by assigning a score from 1 to 4 (cf. Appendix E). Here, a score of 1 or 2 indicates disagreement (complete or partial) with the target label, while a score of 3 or 4 indicates agreement (partial or complete) with the target label. Ideally, the evaluation LLMs should give high scores to the honest set, and low scores to the corrupted set. We show the distributions for disagreements and agreements in Table 7.

On SA, both LLMs perform well, but GPT4 exhibits higher sensitivity to the corrupted examples. On NLI, we notice that GPT3.5 gives high flip label scores to humans CFs with both correct and incorrect labels. GPT4 performs much better, but still exhibits high tendency to agree with wrong labels ($\sim 40\%$). The results can be explained by the ten-

	Test Data		
	crowd Premise	crowd Hypothesis	Orig.
Original only	43.60 ± 3.87	59.75 ± 3.06	71.85 ± 1.33
Human Crowd	**63.42** ± 2.74	**70.53** ± 1.02	**76.65** ± 2.04
GPT3.5	54.42 ± 1.86	49.68 ± 2.64	53.00 ± 2.61
GPT4	53.10 ± 1.85	**54.50** ± 1.28	**63.50** ± 1.31
LLAMA2 7B	52.85 ± 1.29	49.45 ± 2.03	58.15 ± 2.23
LLAMA2 70B	**54.58** ± 3.69	49.02 ± 2.96	58.05 ± 0.78
Mistral 7B	51.05 ± 2.89	46.52 ± 2.51	58.50 ± 2.50
Mistral 56B	51.35 ± 1.79	45.45 ± 1.07	48.65 ± 1.88

Table 4: Data augmentation results for NLI. Classification model is trained on original and LLMs or human-generated CFs with Accuracy metric.

	Test Data	
	CFs	Orig.
Original only	83.27 ± 2.66	70.28 ± 0.60
Human	**94.27** ± 0.20	**94.30** ± 0.14
GPT3.5	81.00 ± 2.87	**70.29** ± 0.96
GPT4	**86.00** ± 3.20	69.33 ± 0.49
Mistral 7B	84.32 ± 2.52	69.90 ± 0.75
Mistral 56B	82.86 ± 1.78	68.58 ± 0.97

Table 5: Data augmentation results for Hate Speech. Classification model is trained on original and LLMs or human-generated CFs with accuracy as a metric.

dency of LLMs to agree with the provided answers, especially on reasoning tasks (Zheng et al., 2023). To verify this, we prompt both LLMs to classify the same set of NLI CFs by choosing one of the three labels (entailment, neutral, contradiction) using a similar prompt. The classification results in Table 8 show an improved performance compared to asking the same LLMs if they agree with incorrect labels (cf. Table 7). We also compare the flip label score distributions of GPT3.5 and GPT4 on the corrupted set in Table 11, and observe that even though GPT3.5 gives high scores to corrupted inputs it is less certain (most frequent score is 3), whereas GPT4 tends to be more certain and assigns mostly 1 or 4 (> 93%).

Evaluation with GPT4. We conduct a wide-scale CFs evaluation with GPT4. Besides verifying the target label **FL**, we also ask GPT4 to judge if there are any unnecessary alterations **UA**, and if the CF is realistic **RS**. For these aspects, we use a scoring scheme ranging from 1 to 4, where higher scores indicate better performance. The results for the GPT4 evaluation can be found in Table 9.

The evaluation scores from GPT4 show that

Compared values	**SA**	**NLI**	**HS**
Accuracy & −PPL	-0.26	-0.56	-0.10
Accuracy & −TS	0.49	0.52	0.60

Table 6: Spearman correlations between intrinsic metrics and data augmentation performance.

GPT4 prefers LLM-CFs, and especially its own generations, which are given the highest scores on most datasets. On SA, Mistral 56B scores the highest with LLAMA 70B and GPT4 having slightly lower scores. On NLI, human CFs take the second position after GPT4. On HS, GPT4 performs the best, while human CFs are given the second lowest score on average. GPT4 might have a bias to prefer its own generations (Panickssery et al., 2024). We further investigate this bias in Section 4.4.3. To further verify the evaluation scores from GPT4, we calculate the correlations between GPT4 scores and the scores from the intrinsic evaluation.

The correlations in Table 10 indicate strong correlation for label flipping on SA and NLI, but weak correlation on HS. This suggests that GPT4 highly agrees with the classifier. Minimal changes show weak correlation with Levenshtein distance on SA and HS, with moderate correlation on NLI, implying that GPT4 is not necessarily sensitive to small changes. GPT4 shows weak to moderate positive correlation on realisticness with perplexity on HS and SA, and moderate negative correlation on NLI. This discrepancy might be due to the nature of the different texts, i.e., while SA contains long movie reviews, NLI contains short image captions and HS contains tweets.

LLMs show a high tendency to agree with the provided labels even if these are incorrect, especially on tasks that require reasoning such as NLI. The correlation between GPT4 evaluation scores and automated metrics for label flipping, textual distance, and fluency varies across tasks .

4.4 Qualitative Analysis

4.4.1 CFs for NLI

We look into a selected set of examples based on the evaluation from GPT4. For each LLM, we pick 2 NLI examples with the highest/lowest scores. We end up with 28 examples. We identify three categories of errors based on this sample :

- **Copy-Paste:** When asked to generate a CF, and change the label from *contradiction to*

LLM/Set	Task	1&2	3&4	Avg.
GPT3.5				
Honest	SA	3.61	96.39	3.43
Corrupted	SA	77.42	22.58	1.61
Honest	premise	0.63	99.37	3.57
Corrupted	premise	5.56	94.44	3.13
Honest	hypothesis	1.38	98.62	3.56
Corrupted	hypothesis	3.53	96.47	3.28
GPT4				
Honest	SA	7.53	92.47	3.66
Corrupted	SA	98.93	1.08	1.04
Honest	premise	12.31	87.69	3.58
Corrupted	premise	59.51	40.50	2.19
Honest	hypothesis	4.50	95.50	3.81
Corrupted	hypothesis	57.87	42.12	2.29

Table 7: Flip label scores distribution for GPT3.5 and GPT4 on honest and corrupted sets.

Set	LLM	Part	Acc.
Honest	GPT3.5	premise	54.90
Honest	GPT3.5	hypothesis	63.08
Honest	GPT4	premise	59.25
Honest	GPT4	hypothesis	75.75

Table 8: Classification performance on human CFs. Note the improved performance compared to asking LLMs if they agree with a given label (cf. Table 7).

entailment, LLMs will use the unchanged part (premise or hypothesis) as output. This a clever but lazy way to flip the label to *entailment*, since two identical sentences would naturally have the label *entailment*. These CFs were given perfect scores by GPT4. Table 14 in the Appendix shows the percentage of copy-paste CFs for all LLMs (at most 4.27% for GPT3.5).

- **Negation:** When asked to to change the label from *entailment to contradiction*, LLMs would negate the premise/hypothesis. The negation does not make sense in the observed CFs.

- **Inconsistency:** These examples contain contradictory or illogical sentences, but GPT4 sometimes incorrectly assigned high scores.

We show examples for each error category in Table 12.

4.4.2 Evaluation Scores

We also look into the evaluation scores from GPT4 on the same set of examples. We show correct and incorrect evaluations in Table 15. GPT4 assigns high scores to contradictory examples, which partially fulfill the target label, and low scores to examples which contain valid minimal changes. GPT4 could be insensitive to such small changes.

4.4.3 Bias in GPT4 Scores

Given GPT4's preference towards its own generations (cf. Table 9), we conduct a qualitative analysis to examine if we agree with the scores given by GPT4 on a set of SA CFs. More specifically, we examine a set of expert CFs that were given lower scores than their corresponding GPT4 CFs on all three metrics. On 12 out of 14 instances we do not agree with the scores given by GPT4. We notice that GPT4 unnecessarily changes some parts of the movie reviews, and introduces changes that do not make sense in the wider context of the reviews. We also noticed that GPT4 changes/omits parts containing potentially harmful content (e.g., torture, sexual content, etc.). Hence, we believe GPT4 prefers its own generations, because these generations do not contain any harmful content (despite safety not being an evaluation criteria).

5 Related Work

Large Language Models. LLMs have demonstrated impressive capabilities across a diverse natural language processing tasks, such as question answering, wherein the model needs to retrieve relevant information from its training data and generate a concise response, or text summarization, which distills lengthy texts into concise summaries while retaining crucial information (Maynez et al., 2023). However, the task of CFs generation has not been comprehensively evaluated for LLMs. A large number of LLMs exist, exhibiting variations in model size, architecture, training dataset, the incorporation of human feedback loops and accessibility (open-source or proprietary) (Zhao et al., 2023). Consequently, there is a necessity to conduct comparative evaluations across different models on a standardized task. Since the architectures of the LLMs under consideration are predominantly similar, and the training datasets are either known public sources or undisclosed, the primary focus

	SA				NLI - premise				NLI - hypothesis				Hate Speech			
	FL	UA	RS	Avg.	FL	UA	RS	Avg.	FL	UA	RS	Avg.	FL	UA	RS	Avg.
Expert Humans	3.54	2.69	2.49	2.91	-	-	-	-	-	-	-	-	-	-	-	-
Crowd Humans	3.66	2.95	2.58	3.06	<u>3.58</u>	**3.88**	**3.86**	<u>3.77</u>	<u>3.81</u>	<u>3.96</u>	3.81	<u>3.86</u>	3.04	3.54	3.19	3.26
GPT3.5	3.58	2.91	2.65	3.05	2.51	3.82	3.69	3.34	3.19	3.93	3.74	3.62	1.78	<u>3.58</u>	3.02	2.79
GPT4	3.79	**3.15**	2.91	3.28	**3.68**	<u>3.83</u>	<u>3.84</u>	**3.78**	**3.96**	**3.98**	**3.92**	**3.95**	**3.65**	**3.73**	**3.63**	**3.67**
LLAMA2 7B	3.60	2.74	2.63	2.99	2.96	3.38	3.67	3.34	3.23	3.74	3.66	3.54	-	-	-	-
LLAMA2 70B	<u>3.87</u>	3.05	**2.96**	<u>3.29</u>	3.07	3.68	3.77	3.51	3.60	3.89	3.75	3.75	-	-	-	-
Mistral 7B	3.85	2.84	2.69	3.13	2.97	3.63	3.74	3.45	3.50	3.70	3.65	3.62	<u>3.32</u>	<u>3.58</u>	<u>3.40</u>	<u>3.43</u>
Mistral 56B	**3.88**	<u>3.07</u>	<u>2.94</u>	**3.30**	2.71	3.81	3.75	3.42	2.95	3.94	<u>3.84</u>	3.58	3.31	3.44	3.25	3.33

Table 9: Scores for evaluation with GPT4. **FL** refers to flipping label score, **UA** to unncessary alteration, **RS** is the realisticness score, and **Avg.** is the average of the three scores. Best score for each task is in **bold**. Second best score is <u>underlined</u>.

Compared Values	SA	NLI	HS
FL & FR	0.86	0.92	0.30
UA & -TS	0.18	0.60	0.10
RS & -PPL	0.43	-0.26	0.20

Table 10: Spearman correlations between intrinsic metrics and GPT-4 evaluation scores. **PPL** and **TS** scores are negated so that higher is better.

LLM/Score	1	2	3	4
GPT3.5	0.70	3.85	69.61	25.84
GPT4	55.50	3.19	2.94	38.37

Table 11: Flip label score distributions on the corrupted set of NLI. Distribution is an average of the distributions on the premise and hypothesis sets.

of this study is to compare LLMs that are different in model size, the implementation of human feedback, and accessibility. To enhance the performance of LLMs across various tasks, in-context learning (ICL) techniques have been employed to optimize the prompts provided to these models. Numerous prompt engineering approaches during the inference phase have been proposed, either by selecting the demonstration instances, or formatting the prompt in form of instruction or reasoning steps (Dong et al., 2022). In this study, leverage chain-of-thought prompting (CoT) (Wei et al., 2022) and selecting closest instance retrieval strategies(Liu et al., 2022) to optimize the generation process.

CFs generation methods. There exists several methods for generating CFs, but most of them are desigend for a specific LLM. The CFs generated by MICE (Ross et al., 2021) are intended for debugging models, and not for data augmentation. Polyjuice (Wu et al., 2021) requires specifying the type of edits that should be conducted, and the resulting CFs should be manually labeled. (Robeer et al., 2021) DISCO (Chen et al., 2023) uses GPT-3's fill-in-the-blanks mode, which is unavailable in most open source LLMs and would require adapting them. CREST (Treviso et al., 2023) depends on a rationalizer module and the editor module is a masked LM that needs to be further trained. Instead, we decided to prompt LLMs to generate CFs by providing instructions and an example. We provide more details in Section 3.2.

LLMs for CFs generation (Li et al., 2024) investigated the strengths and weaknesses of LLMs as CFs generators. Additionally, they disclosed the factors that impact LLMs during CFs generation, including both intrinsic properties of LLMs and prompt design considerations. However, this study lacks intrinsic evaluation of CFs and omits comparison with human-generated CFs. Sachdeva et al. (2024) leverage LLMs to generate CFs for extractive question answering, showing that data augmentation with CFs improve OOD performance, and that this improvement correlates with the diversity of the generated CFs. Prior work by Bhattacharjee et al. (2024) investigated the capability of GPT models in generating CFs for explanations by optimizing their prompts. However, their analysis was limited to the GPT family and did not consider downstream tasks or comparison with human-generated CFs. In this work, we conduct a more comprehensive evaluation of LLMs on multiple aspects, including intrinsic metrics of CFs explanation quality and performance on downstream tasks. Furthermore, we compare the LLM-generated CFs against those produced by humans, and propose a novel approach to evaluate CFs using LLMs.

Original Premise	Original Hypothesis	Counterfactual	Original Label	Target Label	FL	UA	RS	LLM	Changed Part
Copy-Paste									
A photographer and assistant take a photo of a man in a suit and a woman in a dress in front of large white columns and a yellow wall.	The photographer is taking a photo of a dog.	The photographer is taking a photo of a dog.	contradiction	entailment	4.0	4.0	4.0	GPT4	premise
At least eight people are sitting at a table in a building that overlooks the city.	People sitting on the floor	At least eight people are sitting at a table on the floor in a building that overlooks the city.	contradiction	entailment	4.0	4.0	4.0	Mistral 7B	hypothesis
Inconsistency									
At least eight people are sitting at a table in a building that overlooks the city.	People sitting on the floor	At least eight people are sitting at a table in a building that overlooks the city. People can be sitting on the floor.	contradiction	entailment	4.0	4.0	4.0	Mistral 7B	premise
An old man in a blue shirt and white pants glowers by a glass door.	A man resting outside	A glass man by a door.	neutral	contradiction	1.0	1.0	1.0	Human Crowd	hypothesis
Negation									
Two men in costumes with fake carrot noses, top hats, sunglasses and white fur coats that contain white electrical lights.	People in costumes	Two women in costumes with real carrot noses, no hats, no sunglasses, no coats, no lights.	entailment	contradiction	1.0	2.0	4.0	LLAMA2 70B	premise

Table 12: Categorization of a sample of incorrect NLI CFs with evaluation scores from GPT4.

6 Conclusion

In this work, we investigated the use of various LLMs for CFs generation. Our results show that LLMs generate fluent CFs, but struggle to keep the induced changes minimal. Generating CFs for SA is less challenging than NLI and HS, where LLMs show weaknesses in generating CFs that change the original label. CFs from LLMs can replace human CFs for the purpose of data augmentation on SA and achieve similar performance, while on NLI and HS further improvements are needed. Further, our results suggest that CFs with minimal changes are essential for data augmentation. We also showed that when asked to asses CFs, LLMs exhibit a strong bias towards agreeing with the provided label even if this label is incorrect. GPT4 appears to be more robust than GPT3.5 against this bias. Furthermore, we showed that GPT4 scores its own generations higher and that safety training might be one reason for this preference, i.e., GPT4 prefers its own generations, because they do not contain any harmful content. Future work should focus on (i) leveraging LLMs for higher quality NLI and HS CFs, which correctly change the label and keep changes minimal, (ii) assessing the evaluation abilities of LLMs in mislabeled data settings, and (iii) investigating the effects of safety training on LLMs as evaluators.

7 Limitations

We used the default parameters for generating counterfactuals. Experimenting with different parameters might have a non-negligble effect on the results. We included various LLMs in our experiments to be inclusive and be able to compare open-source and closed LLMs. However, these LLMs might have been exposed, during their training, to the data we use from (Kaushik et al., 2020). In this regard, the training data of most open-source and all closed-source LLMs remains unknown. In our qualitative analysis (see Section 4.4), we noticed that GPT4 generated a CF that is identical to a human CF from (Kaushik et al., 2020).

References

Ananth Balashankar, Xuezhi Wang, Yao Qin, Ben Packer, Nithum Thain, Ed Chi, Jilin Chen, and Alex Beutel. 2023. Improving classifier robustness through active generative counterfactual data augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 127–139, Singapore. Association for Computational Linguistics.

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2024. Towards llm-guided causal explainability for black-box text classifiers.

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.

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.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using llms: Data perspectives, learning paradigms and challenges. *arXiv preprint arXiv:2403.02990*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

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.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in Large Multilingual Translation Models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

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*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Yongqi Li, Mayi Xu, Xin Miao, Shen Zhou, and Tieyun Qian. 2024. Prompting large language models for counterfactual generation: An empirical study.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta. Association for Computational Linguistics.

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 *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Joshua Maynez, Priyanka Agrawal, and Sebastian Gehrmann. 2023. Benchmarking large language model capabilities for conditional generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9194–9213, Toronto, Canada. Association for Computational Linguistics.

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

Christoph Molnar. 2022. *Interpretable Machine Learning*, 2 edition.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

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

Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Marcel Robeer, Floris Bex, and Ad Feelders. 2021. Generating realistic natural language counterfactuals. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3611–3625, Punta Cana, Dominican Republic. 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.

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.

Indira Sen, Mattia Samory, Fabian Flöck, Claudia Wagner, and Isabelle Augenstein. 2021. How does counterfactually augmented data impact models for social computing constructs? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 325–344, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, and Moya Chen et. al. 2023. Llama 2: Open foundation and fine-tuned chat 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.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. 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 1667–1682, Online. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

A Successful Generations

Table 13 shows how often LLMs successfully generated CFs, i.e., how often they adhered to the predefined template in the prompt.

Test split	**Success Rate**
SA	
GPT3.5	100.00
GPT4	99.59
LLAMA2 7B	98.98
LLAMA2 70B	81.76
MISTRAL 7B	84.22
MISTRAL 56B	94.67
NLI	
changed premise	
GPT3.5	100.00
GPT4	100.00
LLAMA2 7B	96.00
LLAMA2 70B	98.00
MISTRAL 7B	96.12
MISTRAL 56B	99.25
changed hypothesis	
GPT3.5	100.00
GPT4	100.00
LLAMA2 7B	99.62
LLAMA2 70B	94.38
MISTRAL 7B	94.38
MISTRAL 56B	98.25
HS	
GPT3.5	63.21
GPT4	76.28
MISTRAL 7B	80.44
MISTRAL 56B	81.44

Table 13: Success rate in generating CFs. We consider generations that do not adhere to the pre-defined template in the prompt as failed generations.

LLM	**changed part**	**percentage**
Crowd	premise	0.00
Crowd	hypothesis	0.00
GPT3.5	premise	1.63
GPT3.5	hypothesis	4.27
GPT4	premise	4.14
GPT4	hypothesis	2.25
LLAMA2 7B	premise	3.00
LLAMA2 7B	hypothesis	2.26
LLAMA2 70B	premise	2.04
LLAMA2 70B	hypothesis	1.06
MISTRAL 7B	premise	0.91
MISTRAL 7B	hypothesis	1.59
MISTRAL 56B	premise	0.63
MISTRAL 56B	hypothesis	1.40

Table 14: Percentage of CFs for each LLM where the CFs were a copy of the premise/hypothesis.

B Further Analysis

Table 14 shows the percentage of copy/paste examples in NLI CFs. 0

C Hyperparameter Tuning

In order to evaluate how beneficial are the generated counterfactuals when used for data augmentation, we train several models with and without the generated counterfactuals. When training these, we conduct hyperparameter optimization, and choose the best model based on its performance on a validation set. We considers the following hyperparameters: batch size = [16, 32, 64], training epochs = [2,3,4,5], learning rate $U(1e-5, 5e-5)$, weight decay $U(0, 0.3)$. In our experiments, we use HuggingFace (Wolf et al., 2020) for training and evaluation of perplexity. For distance, we use spaCy [6].

D Examples

We show categorize examples in Table 12. Some correct and incorrect evaluation examples are depicted in Figure 15. An example of LLMs evaluation on an SA example is shown in Figure 3.

E Prompt

Prompts for generating SA CFs and NLI CFs are shown in Figure 4 and 5 respectively. Prompts for evaluating SA CFs and NLI CFs are shown in Figure 7 and 6 respectively.

[6] https://github.com/explosion/spaCy

Premise	Hypothesis	Counterfactual	Original Label	Target Label	FL	UA	RS	LLM	Changed Part
Correct									
An older man sifts water through his fingers as a young toddler watches from inside the water.	A man is near a toddler	An older man sifts water through his fingers as a young toddler is far from any toddler.	entailment	contradiction	1.0	1.0	1.0	GPT4	revised_premise
Incorrect									
At least eight people are sitting at a table in a building that overlooks the city.	People sitting on the floor	At least eight people are sitting at a table in a building that overlooks the city. People can be sitting on the floor.	contradiction	entailment	4.0	4.0	4.0	Mistral 7B	revised_premise

Table 15: Correct and incorrect evaluation examples from GPT4.

```
Text 1, Factual Label = Negative:  Waiting to go inside the theathre with tickets in my hand, I
expected an interesting sci-fi fantasy movie which could finally feed my appetite of movies regarding
robot-technology, instead I went disappointed by each aspect of it, once more proving that stunning
special effects can't help a boring plot, which by my opinion was the worse in this year.  Acting in
this movie also dissatisfied me, Will Smith didn't show anything new in this movie, yet I never saw
his acting to change since "Men In Black" which was his only success by my opinion.  He had to retire
since than, not spoiling his name with titles like "I,Robot" and "Men In Black 2".  4/10

Text 2, Target Label = Positive:  Waiting to go inside the theatre with tickets in my hand, I saw
an interesting sci-fi fantasy movie which could finally feed my appetite for movies regarding
robot-technology, I wasn't disappointed by any aspect of it, once more proving that stunning special
effects can help the plot, which by my opinion was the best this year.  Acting in this movie satisfied
me, Will Smith didn't show anything new in this movie, yet I never saw his acting to change since "Men
In Black" which was one of his biggest success in my opinion.  ".  8/10
Answer:
1.  4
2.  2
3.  2

Explanation for 3:  The overall sentiment toward the movie has been successfully inverted from
negative to positive.  However, the remark about Will Smith's acting not changing since "Men In Black"
remained unchanged, which conflicts with the positive sentiment of the rest of the review.  This
inconsistency may cause readers to perceive the review as somewhat unrealistic because the critique of
the actor's performance persisting in an otherwise positive review is unusual.
4.  3
5.  2
6.  2
```

Figure 3: GPT-4 evaluation for a crowd-generated text in a sentiment analysis task, with a good explanation, as it can identify the negative aspects in the revised review.

```
Request: Given a piece of text with the original sentiment in the form of "Sentiment: Text". Change
the text with minimal edits to get the target sentiment from the original sentiment. Do not make any
unneccesary changes.
(Original text)
Positive: This is an excellent film!Tom Hanks and Paul Newman performed great!I was really surprised
when Newman was beating on his son!That was a great scene and the shooting scenes were staged good.I
was very surprised about the end.Rent this film today as it is one of Tom Hanks' best!
Target: Negative

Step 1: Identify phrases, words leading to the Positive sentiment:
'an excellent',
'great!I',
'a great',
'good.I',
'end.Rent ',
'today',
'best!',

Step 2: Change these phrases, words, sentences to Negative sentiment with minimal changes:
'an excellent' to 'a terrible'
'great!I' to 'badly!I'
'a great' to 'an awful'
'good.I' to 'badly.I'
'end.Rent ' to 'end.Don't rent'
'today' to ''
'best!' to 'worst!'

Step 3: replace the phrases, words, sentences from step 1 in the original text by the phrases, words,
sentences in step 2:
Negative: This is a terrible film!Tom Hanks and Paul Newman performed badly!I was really surprised
when Newman was beating on his son!That was an awful scene and the shooting scenes were staged badly.I
was very surprised about the end.Don't rent this film as it is one of Tom Hanks' worst!

Request: Given a piece of text with the original sentiment in the form of "Sentiment: Text". Change
the text with minimal edits to get the target sentiment from the original sentiment. Do not make any
unnecessary changes.
(Original text)
Positive: Well, sorry for the mistake on the one line summary.......Run people, run..to your nearest
movie store, that is! This movie is an fabulous!! Imagine! Gary Busey in another low budget movie,
with an incredibly funny scenario...isn't that a dream? No (well yes), it is Plato's run...........I
give it **** out of *****.
Target: Negative
```

Figure 4: Prompt for CFs generation - SA

```
Given two sentences (premise and hypothesis) and their original relationship, determine whether they
entail, contradict, or are neutral to each other. Change the sentence2 with minimal edits to achieve
the target relation from the original one. Do not make any unnecessary changes. For example:
Original relation: entailment
Two original sentences: Brown-haired woman talking to man with backpack. A woman is talking to a man
Target relation: neutral
Target sentence: sentence2
Step 1: Identify phrases, words in the sentence2 leading to the entailment relation:
'man',

Step 2: Change these phrases, words to get neutral relation with minimal changes:
'man' to 'student.'

Step 3: replace the phrases, words from step 1 in the original text by the phrases, words, sentences
in step 2:
(Edited sentence2): A woman is talking to a student.
#####End Example####
Request: Given two sentences (premise and hypothesis) and their original relationship, determine
whether they entail, contradict, or are neutral to each other. Change the sentence2 with minimal edits
to achieve the neutral relation from the original one. Do not make any unnecessary changes. Do not
add anything else.
Original relation: entailment
Two original sentences: A blond woman speaking to a brunette woman with her arms crossed. A woman is
talking to another woman.
Target relation: neutral
Target sentence: sentence2
```

Figure 5: Prompt for CFs generation - NLI

```
Evaluating Counterfactuals
Natural Language Inference (NLI) is a fundamental task in natural language processing (NLP) that
involves determining the relationship between two pieces of text:  a premise and a hypothesis.  The
relation between the premise and the hypothesis is described using three different labels:
Entailment:  if the hypothesis is definitely true given the premise.
Example for Entailment:
Premise:  A soccer game with multiple males playing.
Hypothesis:  Some men are playing a sport.
Label:  Entailment
Neutral:  if the hypothesis might be true given the premise.
Example for Neutral:
Premise:  An older and younger man smiling.
Hypothesis:  Two men are smiling and laughing at the cats playing on the floor.
Label:  Neutral
Contradiction:  if the hypothesis is definitely false given the premise.
Example for Contradiction:
Premise:  A man inspects the uniform of a figure in some East Asian country.
Hypothesis:  The man is sleeping
Label:  Contradiction
Purpose of the Evaluation:
This evaluation aims to assess the quality of counterfactual texts that were generated by different
methods.  A counterfactual text is an alternative version of a text designed to change the label of
the original (factual) instance while maintaining high text quality.
Task Description:
You will receive two instances.  Each instance consists of two sentences:  a premise and a hypothesis.
Each instance can be classified with one of the three aforementioned labels (Entailment, Neutral,
Contradiction).
Factual instance (Instance 1):  An instance and its factual label.
Counterfactual instance (Instance 2):  A modified version of the factual instance designed to express
a different label, i.e., match the target label.
Read the two texts and answer the questions below:
Instance 1:
Premise:
Hypothesis:
Factual Label:
Instance 2:
Premise:
Hypothesis:
Target Label:
1.  To which extent do you agree that Instance 2 has the label ?
4-totally agree, 3-partially agree, 2-partially disagree, 1-totally disagree
2.  Are there any unnecessary changes (removals, additions, replacements of words) in the
counterfactual text (Instance 2) that do not contribute to changing the original factual label to
the target label?
4-no unnecessary changes, 3-few unnecessary changes, 2-many unnecessary changes, 1-significant number
of unnecessary changes
3.  How realistic is Instance 2?  A realistic instance would not include any imaginary actions/items.
4-very realistic, 3-partially realistic, 2-partially unrealistic, 1-very unrealistic
If you think it is (highly/partially) unrealistic, please provide a brief explanation.
Your evaluation for the provided counterfactual text:
1.  Please provide a number only
2.  Please provide a number only
3.  Please provide a number only
```

Figure 6: Prompt for CFs Evaluation - NLI

```
Evaluating Counterfactuals

Texts can be classified into different categories, e.g., positive vs.  negative sentiment.  In this
case the sentiment (positive/negative) is called the 'label'.  A counterfactual text is an alternative
version of a text designed to change the label of the original (factual) text while maintaining high
text quality.

Purpose of the Evaluation:
This evaluation aims to assess the quality of counterfactual texts that were generated by different
methods.

Task Description:
You will receive two texts.  Each text can either express a positive or a negative sentiment.
Factual Text (Text 1):  A movie review with its (ground truth) factual label.
Counterfactual Text (Text 2):  A modified version of the movie review designed to express the opposite
sentiment, i.e., match the target label.

A simple example:  Text 1, Factual Label = Negative:  This movie is very bad.
Text 2, Target Label = Positive:  This movie is great.

Read the two texts and answer the questions below:

Text 1, Factual Label = :

Text 2, Target Label = :

1.  To which extent do you agree that Text 2 has the label ?
4-totally agree, 3-partially agree, 2-partially disagree, 1-totally disagree

2.  Are there any unnecessary changes (removals, additions, replacements of words) in the
counterfactual text (Text 2) that do not contribute to changing the original factual label to the
target label?
4-no unnecessary changes, 3-few unnecessary changes, 2-many unnecessary changes, 1-significant number
of unnecessary changes

3.  How realistic is Text 2?  A realistic movie review would not read strange in any way on a movie
review website.
4- very realistic, 3-partially realistic, 2-partially unrealistic, 1-very unrealistic

If you think it is (highly/partially) unrealistic, please provide a brief explanation.

Additionally, assess the following aspects of the counterfactual text:
4.  Grammaticality:  how would you rate the grammatical accuracy of text 2?
4-Definitely correct, 3-Somewhat correct, 2-Somewhat incorrect, 1-Definitely incorrect
5.  Cohesiveness:  How well do the sentences in the text 2 fit together?
4-Highly cohesive, 3-Reasonably cohesive, 2-Somewhat disjointed, 1-Very poorly fit together
6.  Likability:  how likely are you to vote for text 2 on the movie review site?
4-Definitely would vote, 3-Likely to vote, 2-Unlikely to vote, 1-Definitely would not vote

Your evaluation for the provided counterfactual text:
1.  (Please provide a number only)
2.  (Please provide a number only)
3.  (Please provide a number only)
4.  (Please provide a number only)
5.  (Please provide a number only)
6.  (Please provide a number only)
```

Figure 7: Prompt for CFs Evaluation - SA