# CEval: A Benchmark for Evaluating Counterfactual Text Generation

**Van Bach Nguyen**
University of Marburg, Germany
vanbach.nguyen@uni-marburg.de

**Christin Seifert**
University of Marburg, Germany
christin.seifert@uni-marburg.de

**Jörg Schlötterer**
University of Marburg, Germany
University of Mannheim, Germany
joerg.schloetterer@uni-marburg.de

## Abstract

Counterfactual text generation aims to minimally change a text, such that it is classified differently. Assessing progress in method development for counterfactual text generation is hindered by a non-uniform usage of data sets and metrics in related work. We propose CEval, a benchmark for comparing counterfactual text generation methods. CEval unifies counterfactual and text quality metrics, includes common counterfactual datasets with human annotations, standard baselines (MICE, GDBA, CREST) and the open-source language model LLAMA-2. Our experiments found no perfect method for generating counterfactual text. Methods that excel at counterfactual metrics often produce lower-quality text while LLMs with simple prompts generate high-quality text but struggle with counterfactual criteria. By making CEval available as an open-source Python library, we encourage the community to contribute additional methods and maintain consistent evaluation in future work.[1]

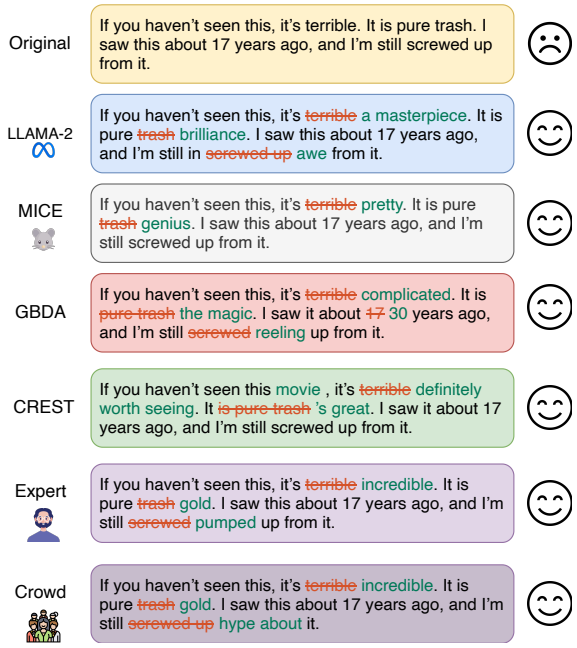


Figure 1: Examples of counterfactuals generated by different methods and human annotators that successfully flip the label from negative to positive for the same original instance.

## 1 Introduction

The rise of deep learning and complex "black-box" models has created a critical need for interpretability. As Miller (2019) notes, explanations often involve counterfactuals to understand why event *P* occurred instead of *Q*. Ideally, these explanations show how minimal changes in an instance could lead to different outcomes. For example, to explain why the review *"The film has funny moments and talented actors,* ***but it*** *feels long."* is negative rather than positive, a counterfactual like *"The film has funny moments and talented actors,* ***yet*** *feels* ***a bit*** *long."* can be used (see Fig. 1 for more counterfactual examples generated by different methods on the same original instance). This explanation highlights specific words to change and modifications needed for a positive sentiment . It also motivates counterfactual generation, which requires modifying an instance minimally to obtain a different model prediction. Besides explanations (Robeer et al., 2021), the NLP community uses counterfactuals for debugging models (Ross et al., 2021), data augmentation (Dixit et al., 2022; Chen et al., 2023; Bhattacharjee et al., 2024), and enhancing model robustness (Treviso et al., 2023; Wu et al., 2021). However, because it requires deciding where and how to change the text, with many possible modifications and a vast vocabulary. While many counterfactual generation methods for text data exist in the literature, they lack unified evaluation standards. Table 1 highlights inconsistencies in datasets, metrics, and baselines across different studies, making it difficult to compare different methods or select-

[1] https://github.com/aix-group/CEval-Counterfactual-Generation-Benchmark



*Proceedings of the 17th International Natural Language Generation Conference*, pages 55–69
September 23–27, 2024. 

| Method | Dataset | Metrics | Baseline |
|---|---|---|---|
| MICE (Ross et al., 2021) | IMDB, Race, Newgroups | Flip rate, Fluency, Minimality | MICE's variants |
| CF-GAN (Robeer et al., 2021) | HATESPEECH, SST-2, SNLI | Fidelity, Perceptibility, Naturalness | SEDC (Martens and Provost, 2014)<br>PWWS+ (Ren et al., 2019)<br>Polyjuice (Wu et al., 2021)<br>TextFooler (Jin et al., 2020) |
| CORE (Dixit et al., 2022) | IMDB, MNLI | Diversity, Closeness, Accuracy | Polyjuice (Wu et al., 2021)<br>GPT-3 (Brown et al., 2020)<br>Human-CAD |
| DISCO (Chen et al., 2023) | SNLI, WANLI | Flip Score, Diversity, Accuracy | Tailor (Ross et al., 2022)<br>Z-aug (Wu et al., 2022)<br>Human-CAD |

Table 1: Inconsistent use of datasets, metrics, and baselines across different methods.

ing the most suitable method for specific applications. To overcome these limitations, a comprehensive benchmark to thoroughly evaluate counterfactual generation methods is necessary. A benchmark that provides standardized datasets, metrics, and baselines, enabling fair and effective comparisons, and ultimately driving progress in counterfactual generation.

This work introduces CEval, the first comprehensive benchmark for evaluating methods that modify text to change classifier predictions, including contrastive explanations, counterfactual generation, and adversarial attacks. CEval offers a robust set of metrics, incorporating established metrics from the literature alongside a novel metric we propose that captures probability changes rather than hard flip rates. This set enables the assessment of both "counterfactual-ness" (e.g., label flipping ability) and textual quality (e.g., fluency, grammar, coherence). The benchmark includes curated datasets with human annotations and a strong baseline using a large language model with a simple prompt to ensure high evaluation standards. Using CEval, we systematically review and compare state-of-the-art methods, highlighting their strengths and weaknesses in generating counterfactual text. We analyze how automatically generated counterfactuals compare to human examples, revealing gaps and opportunities for improvement. We find that counterfactual generation methods often generate text that lacks in quality compared to simple prompt-based LLMs. In contrast, while the latter typically exhibit higher text quality, they may struggle to satisfy counterfactual metrics. These insights suggest exploring combinations of both paradigms into hybrid methods as promising direction for future research. By demonstrating that an open-source LLM can serve as an alternative to a closed-source LLM in text evaluation, we make the benchmark completely open-source, thereby promoting reproducibility and facilitating further research in this domain.

## 2 Related Work

Terms like "counterfactual" and "contrastive" generation are often used interchangeably in literature (Stepin et al., 2021) and our work adopts an inclusive definition. We define counterfactual generation as a process of generating a new instance $x'$, from the original instance $x$, that results in a different model prediction $y'$ with minimal changes. This definition includes counterfactual, contrastive generation, and adversarial attacks. Primarily, adversarial attacks focused on changing the label without considering text quality. Recent work like GBDA (Guo et al., 2021) focuses on producing adversarial text that is more natural by adding fluency and semantic similarity losses. Hence, we include GBDA in our benchmark. Technically, counterfactual generation methods for text fall into three categories:

**Masking and Filling Methods (MF):** These methods perform two steps: (1) identifying important words for masking by various techniques, such as selecting words with the highest gradient or training a separate rationalizer for the masking process and (2) replacing the masked words using a pre-trained language model with fill-in-the-blank capability. In step (1), MICE (Ross et al., 2021) and AutoCAD (Wen et al., 2022) use the gradient of the classifier. DoCoGen (Calderon et al., 2022) identifies all domain-specific terms by calculating a masking score for n-grams (where n$\leq$ 3) and

masks all n-grams with a masking score exceeding a threshold $\tau$. Meanwhile, CREST (Treviso et al., 2023) trains SPECTRA (Guerreiro and Martins, 2021) as a separate rationalizer to detect which phrases or words to mask. In step (2), each of these methods fine-tunes T5 to fill in the blanks created during masking. Additionally, Polyjuice (Wu et al., 2021) takes text with user-specified manual masking as input and fine-tunes a RoBERTa-based model to fill in the blanks using control codes.

**Conditional Distribution Methods (CD):** Methods like GBDA (Guo et al., 2021) and CF-GAN (Robeer et al., 2021) learn a conditional distribution for counterfactuals. The counterfactuals are obtained by sampling from this distribution based on a target label.

**Counterfactual Generation with Large Language Models:** Recently, there has been a trend towards using Large Language Models (LLMs) for counterfactual generation. Approaches like CORE (Dixit et al., 2022), DISCO (Chen et al., 2023) and FLARE (Bhattacharjee et al., 2024) optimize prompts fed into LLMs to generate the desired counterfactuals. This trend is driven by the versatile capabilities of LLMs in various tasks (Maynez et al., 2023).

Despite the diverse approaches proposed in generating counterfactuals across various studies, the common objective remains to generate high-quality counterfactuals. However, previous studies employed different metrics, baselines, and datasets, as illustrated in Table 1. Therefore, given the rapid growth of approaches in this field, establishing a unified evaluation standard becomes paramount. Existing benchmarks for counterfactual generation (Pawelczyk et al., 2021; Moreira et al., 2022) focus exclusively on tabular data with properties that are orthogonal to text (e.g., continuous value ranges). Hence, we introduce CEval to fill this gap and provide a standard evaluation framework specifically tailored to textual counterfactual generation. Our benchmark unifies metrics of both, counterfactual criteria and text quality assessment, including datasets with human annotations and a simple baseline from a large language model.

## 3 Benchmark Design

We focus on counterfactual generation for textual data, which involves editing given text with minimal modifications to produce new text that increases the probability of a predefined target label with respect to a black-box classifier. This process aims to generate a counterfactual, denoted as $x'$, that changes the classifier's predictions compared to the original text $x$.

Formally, given a fixed classifier $f$ and a dataset with $N$ samples $(x_1, x_2, \ldots, x_N)$, $x_i = (z_1, z_2, \ldots, z_n)$ represents a sequence of $n$ tokens. The original prediction is denoted as $f(x) = y$, while the counterfactual prediction is $y' \neq y$. The counterfactual generation process is represented by a method $e : (z_1, \ldots, z_n) \mapsto (z'_1, \ldots, z'_m)$, ensuring that $f(e(x)) = y'$. The resulting counterfactual example is $x' = (z'_1, \ldots, z'_m)$ with $m$ tokens.

A valid counterfactual instance should satisfy the following criteria (Molnar, 2022):

**Predictive Probability**: A counterfactual instance $x'$ should closely produce the predefined prediction $y'$. In other words, the counterfactual text should obtain the desired target label.

**Textual Similarity**: A counterfactual $x'$ should maintain as much similarity as possible to the original instance $x$ in terms of text distance. This ensures that the generated text remains coherent and contextually aligned with the original.

**Likelihood in Feature Space**: A counterfactual should exhibit feature values that resemble real-world text, indicating that $x'$ remains close to a common distribution for text. This criterion ensures that the generated text is plausible, realistic and consistent with typical language patterns.

**Diversity:** When an explanation is ineffective, humans can offer alternatives. Similarly, if a counterfactual is unrealistic or not actionable, it is beneficial to modify the original instance differently to provide diverse options (Mothilal et al., 2020). Therefore, an effective counterfactual method should present multiple ways to change a text instance to obtain the target label. Diversity is measures for a set of counterfactual instances.

### 3.1 Metrics

In CEval, we use two types of metrics: *counterfactual metrics*, which reflect the counterfactual criteria outlined above, and *textual quality metrics*, which assess the quality of the generated text, irrespective of its counterfactual properties.

#### 3.1.1 Counterfactual metrics

**Flip Rate (FR):** measures how effectively a method can change labels of instances with respect to a pretrained classifier. This metric represents the binary case of the *Predictive Probability* cri-

terion, determining whether the label changed or not and is commonly used in the literature (Treviso et al., 2023; Ross et al., 2021). 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):

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

where $\mathbb{1}$ is the indicator function.

**Probability Change ($\Delta$P):** While the flip rate offers a binary assessment of *Predictive Probability*, it does not capture the magnitude of change towards the desired prediction. Some instances may get really close to the target prediction but still fail to flip the label. For example, a review such as: *The movie looks great but has a confusing plot and slow pacing* is close to a positive label but remains negative. Consequently, its probability for the positive label should be larger than for a review like *This movie is terrible*, which is really negative. The Probability Change ($\Delta$P) metric captures such cases by quantifying the difference between the probability of the target label $y'$ for the original instance $x$ and the probability of the target label for the contrasting instance $x'$.

$$\Delta P = \frac{1}{N}\sum_{i=1}^{N}\big(P(y_i' \mid x_i', f) - P(y_i' \mid x_i, f)\big)$$

Here, $P(y \mid x, f)$ is the probability that classifier $f$ assigns to label $y$ on instance $x$.

**Token Distance (TD):** To measure *Textual Similarity*, we use the token-level Levenshtein distance $d(x, x')$ between the original instance $x$ and the counterfactual $x'$. This metric captures all types of text edits—insertions, deletions, and substitutions—making it ideal for evaluating minimal edits as counterfactual generation involves making these specific edits rather than completely rewriting the text. The Levenshtein distance is widely used in related work on counterfactual generation (e.g., Ross et al. (2021); Treviso et al. (2023)).

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

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

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

where $\log p_\theta(z_i \mid z_{<i})$ is the log-likelihood of token $z_i$ given the previous tokens $z_{<i}$.

**Diversity (Div):** We quantify diversity by measuring the token distance between pairs of generated counterfactuals. Given two counterfactuals, $x'^1$ and $x'^2$, for the same instance $x$, diversity is defined as the average pairwise distance between the sets of counterfactuals:

$$Div = \frac{1}{N}\sum_{i=1}^{N} d(x_i'^1, x_i'^2)$$

Here, $d(x_i'^1, x_i'^2)$ is the Levenshtein distance between the corresponding tokens of the two counterfactuals for the $i$-th instance.

#### 3.1.2 Text Quality Metrics

In addition to counterfactual evaluation metrics, we measure the quality of the generated text. *Text quality metrics* are designed to evaluate specific aspects of texts. Following (Chiang and Lee, 2023; Wang et al., 2023b), key text quality metrics for comprehensive insights into text quality are: 1) **Fluency** – natural and readable text flow; 2) **Cohesiveness** – logical and coherent structure and 3) **Grammar** – syntactical and grammatical accuracy.

Combined with counterfactual metrics, text quality metrics provide a comprehensive view on effectiveness and linguistic quality of generated counterfactuals. Evaluating these text quality metrics usually requires human annotations, which are costly and time-consuming. Recently, Chiang and Lee (2023); Huang et al. (2023); Wang et al. (2023b) showed that LLMs, specifically GPT-3/4 and ChatGPT, can serve as an alternative to human evaluation for assessing text quality using these metrics. In this work, we use *ChatGPT (gpt-3.5-turbo-0125)* with a temperature of 0.2 to evaluate the above textual quality metrics on a scale from 1 to 5 following (Chiang and Lee, 2023; Gilardi et al., 2023).

### 3.2 Datasets and Classifiers

We chose two benchmark datasets for different NLP tasks: sentiment analysis on IMDB (Maas

[2]While we use GPT-2 in this study, any other LLM with strong text generation capabilities is a viable drop-in replacement.

et al., 2011) and natural language inference (NLI) on SNLI (Bowman et al., 2015). For both datasets, human-generated counterfactuals from crowdsourcing (Kaushik et al., 2020) are available and for IMDB also from experts (Gardner et al., 2020). Additional datasets with pre-trained classifiers can be added to the benchmark.

IMDB contains diverse movie reviews from the IMDB website, along with corresponding sentiment labels (positive or negative) for each review. We selected the 488 instances with human-generated counterfactuals, balanced between 243 negative and 245 positive reviews (Maynez et al., 2023). Using a pre-trained BERT model[3] from TextAttack (Morris et al., 2020) with 89% accuracy, the counterfactual task is to minimally edit reviews to alter the classifier's prediction.

SNLI (Bowman et al., 2015) consists of sentence pairs labeled as entailment, contradiction, or neutral, requiring models to understand semantic relationships. Using a pre-trained BERT model[4] from TextAttack (Morris et al., 2020) with 90% accuracy, the counterfactual generation methods have to modify the premise or the hypothesis to change the classifier's label.

## 4 Counterfactual Methods Selection

In this section, we briefly describe the counterfactual generation methods we evaluate with our benchmark. We selected at least one representative for each of the categories *Masking and Filling (MF), Conditional Distribution (CD) and Large Language Models (LLMs)* (cf. Section 2) based on the following criteria:

- The authors provide reproducible source code.
- The method is problem agnostic and can be applied to multiple text classification tasks.
- The method has access to the underlying text classifier.

We used the criteria *reproducible code* and *problem agnostic* as hard filters and *access to the target classifier* as soft filter. A *problem agnostic* method is versatile enough to generate counterfactuals for various types of classification problems (whereas methods like Polyjuice (Wu et al., 2021) or Tailor (Ross et al., 2022) require control codes, which limits their flexibility). Methods without access to the target classifier are at disadvantage, as they have no information about the internals of the target classifier. Hence, wherever available, we opted for a method with access to the target classifier. The selection based on these criteria (cf. details in Appendix, Table 4) resulted in MICE, GDBA, CREST and LLAMA-2 as representative counterfactual generation methods. We briefly describe them in the following.

**MICE** (Ross et al., 2021) is a contrastive explanation generation method. It trains an editor to fill masked tokens in a text so that the final text changes the original label. The tokens to be masked are chosen based on the highest gradients contributing to the predictions, and binary search is used to find the minimum number of tokens to mask. This method requires access to the classifier to verify the label internally, representing a counterfactual generation method.

**GBDA** (Guo et al., 2021) is a gradient-based adversarial attack that uses a novel adversarial distribution for end-to-end optimization of adversarial loss and fluency constraints via gradient descent. Similar to MICE, this approach needs access to the classifier for internal label verification. This method represents the adversarial attack domain.

**CREST** (Treviso et al., 2023) follows a similar approach as MICE in first masking tokens that should be changed. Instead of using the highest gradient tokens to find the masks, the authors train a rationalizer using SPECTRA (Guerreiro and Martins, 2021). Then, they fill the blanks with T5 same as MICE. Given the popularity of the Mask and Filling type, we chose this method for a more comprehensive comparison.

**LLAMA-2** (Touvron et al., 2023): Large Language Models have shown good performance on many tasks with only simple prompts (Srivastava et al., 2023). Therefore, in this study, we use LLAMA-2 with simple one-shot learning as a baseline that is not specifically designed for counterfactual generation, but has strong language generation capabilities. The choice for LLAMA-2 as an open-source model is made in contrast to other studies that used closed-source LLMs.

The hyperparameters of each selected method can significantly impact the results, particularly for MICE (Ross et al., 2021) and CREST (Treviso et al., 2023). The percentage of masked tokens in both methods, representing the upper bound of changed tokens, directly influences the token distance and indirectly affects the flip rate: a lower

[3] https://huggingface.co/textattack/bert-base-uncased-imdb

[4] https://huggingface.co/textattack/bert-base-uncased-snli

| | | IMDB | | | | | | SNLI | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LLAMA-2 | MICE | GBDA | CREST | Expert | Crowd | LLAMA-2 | MICE | GBDA | CREST | Crowd |
| **CF Metrics** | Flip Rate ↑ | 0.7 | **1.0** | 0.97 | 0.71 | 0.81 | 0.85 | 0.39 | 0.85 | **0.94** | 0.39 | 0.75 |
| | ΔProbability ↑ | 0.69 | 0.91 | **0.96** | 0.70 | 0.80 | 0.84 | 0.33 | 0.65 | **0.86** | 0.10 | 0.64 |
| | Perplexity ↓ | **41.3** | 62.1 | 84.1 | 44.7 | 56.2 | 52.4 | **57.0** | 160 | 143 | 60.9 | 72.1 |
| | Distance ↓ | 73.9 | 38.5 | 46.1 | 70.5 | 29.3 | **25.0** | 6.15 | 5.64 | 4.85 | **3.53** | 4.06 |
| | Diversity ↑ | 61.6 | 48.4 | 47.6 | **86.6** | 38.7 | 38.7 | - | - | - | - | - |
| **Text Quality** | Grammar ↑ | 3.18 | 2.71 | 2.16 | 2.18 | 2.90 | 2.92 | 3.68 | 3.33 | 2.29 | 2.71 | 3.58 |
| | Cohesiveness ↑ | 3.12 | 2.81 | 2.38 | 2.27 | 2.99 | 2.95 | 3.61 | 3.31 | 2.03 | 2.74 | 3.60 |
| | Fluency ↑ | 3.13 | 2.79 | 2.37 | 2.33 | 2.99 | 2.92 | 3.59 | 3.33 | 2.17 | 2.70 | 3.56 |
| | *Average* ↑ | **3.14** | 2.77 | 2.30 | 2.27 | 2.96 | 2.93 | **3.63** | 3.33 | 2.16 | 2.72 | 3.58 |

Table 2: Results with counterfactual (CF) and text quality metrics on IMDB and SNLI. *Average* denotes average of text quality metrics, each scored on a scale 1-5 following (Chiang and Lee, 2023). We calculate diversity of the human groups by comparing expert with crowd counterfactuals and omit diversity on SNLI as it only has a single human counterfactual per instance (no expert annotations).

percentage allows fewer tokens to change, resulting in a smaller distance but potentially a lower flip rate. In our experiments, we maintain the hyperparameters as specified in the original papers of each method. In case of LLAMA-2, the temperature of LLMs affects word sampling: lower temperatures yield more deterministic results, while higher temperatures enhance creativity. For the comparison with other methods, we use a temperature of 1.0 and analyze the impact of varying temperatures at the end of the next section.

## 5 Results

We evaluate all counterfactual generation methods against human crowd-sourced and human expert generations. Note that MICE and GBDA have access to the prediction model during generation, while CREST employs a pre-trained T5 model for internal label verification and transfers its prediction to the target BERT model. In contrast, LLAMA-2 and both human evaluation groups (crowd and expert) generate counterfactual examples solely based on the provided text and prompt.

We start with an example to illustrate the methods' varying characteristics before discussing our observations from the quantitative results. Fig. 1 shows the shortest example in the IMDB dataset where all methods, including human edits, change the label of the original sentence on the generated counterfactual. For this simple instance, all methods and human groups agree on replacing negative words like terrible and trash with positive words, even though they differ in their choice of positive words. GDBA is the only exception, its replacements do not always convey a positive sentiment, which reduces text quality. Similarly, MICE and CREST fail to detect the negative phrase screwed up, which renders the text less cohesive and fluent than the text generated by LLAMA-2 and humans, who adapt this negative phrase as well. Besides correctly identifying important words, GDBA also replaces irrelevant words like 17 → 30, resulting in a larger edit distance. For a more complex example with higher variation of edits and generated text, see Table 9 in the Appendix.

**There is no single best method.** Table 2 shows that no single method consistently outperforms the others, even on a single dataset. Methods with access to the target classifier, such as MICE and GDBA, excel at flipping the label but generate "unnatural" text with lower quality and higher perplexity due to poor grammar and low cohesiveness. In contrast, humans and LLAMA-2 consistently produce higher quality text across most metrics on both datasets. The lower success rate of humans in flipping the label suggests limitations in the target classifier, as perfect flip rates would be expected for human-generated text, the "gold standard." Such potential issues are consistent with prior studies (Kaushik et al., 2020; Gardner et al., 2020). Additionally, LLMs used as evaluation proxies, such as ChatGPT and GPT-2 (which measures perplexity), prefer LLAMA-2's output over human-generated text on both the SNLI and IMDB datasets. This preference is observed across different evaluator temperatures, as shown in Table 3, suggesting an interesting direction for further research into bias of LLMs as evaluators.

**Diversity and distance are correlated.** On the IMDB dataset, CREST and LLAMA-2 exhibit the highest diversity but also the highest distance. In

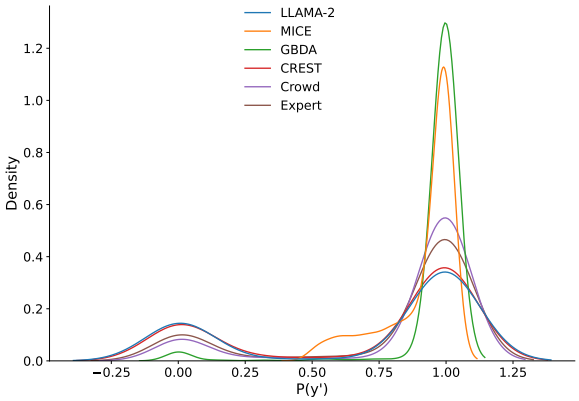


Figure 2: Distribution of target label probabilities of all methods on the IMDB dataset, including original text and human groups.

contrast, human-generated changes (crowd and expert) are minimal and the least diverse. The Pearson correlation between diversity and distance is 0.93, indicating a very strong correlation between these two metrics. This strong correlation is likely due to minimal changes limiting the amount of variation.

**Probability changes are mostly bimodal.** Interestingly, MICE has the highest flip rate (FR), but not the largest change in target label probability change ($\Delta$P) on the IMDB dataset. We observe a similar pattern when comparing LLAMA-2 and CREST on the SNLI dataset. CREST has an equal FR, despite LLAMA-2 inducing a larger $\Delta$P. A high FR combined with a low $\Delta$P suggests that the counterfactuals generated by the method are close to the decision boundary of the target classifier. Fig. 2 shows that only MICE generates a noticeable amount of instances that are close to the decision boundary ($P(y') = 0.5$). All others, including human groups, exhibit a bimodal pattern with narrow peaks at the two extremes. While the imperfect FR of human groups suggests limitations in the target classifier, the distribution pattern may indicate the source of those limitations: This pattern points to a poorly calibrated, overconfident target classifier, a common issue in today's deep learning architectures (Guo et al., 2017).

**Generated texts exhibit substantial differences.** Among automatically generated methods, MICE's counterfactuals are closest to the original texts[5] on the IMDB dataset, but still edit more tokens than humans (expert and crowd). The distance scores of CREST and LLAMA-2 are similar, as are those for MICE and GBDA, and for expert and crowd edits on the IMDB dataset. However, similar edit distances do not imply that these methods make the same edits. To investigate the similarity of edits by different methods, we calculated the average pairwise distance between all generated examples on the IMDB dataset, regardless of label flip success. The results are visualized in Fig. 3. Crowd

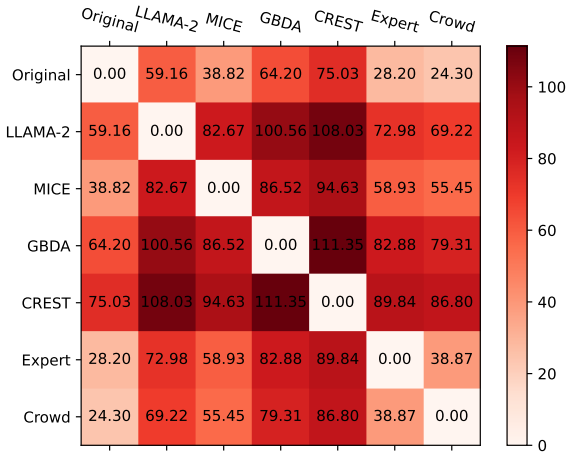


Figure 3: Avg. pairwise Levenshtein distance on IMDB.

and expert edits are highly similar, indicating substantial overlap in their modifications. MICE generated text is closest to human edits, which makes it the most promising candidate to serve as proxy for human-generated counterfactuals. GBDA and CREST have the largest distance to all other methods (including the original text) and to each other, i.e., their edits are largely distinct. This substantial difference in generated texts suggests that robustness analyses of the target classifier should always be conducted with multiple methods.

**Temperature affects counterfactual generation diversity** We compare LLAMA-2's temperature setting of 1.0 in Table 2 with additional values of 0.2 and 0.6 for *counterfactual generation* and observe that the diversity score of LLAMA-2 varies significantly with temperature changes: the lower the temperature, the lower the diversity. For a temperature of 0.2, diversity score is 28.3 and for temperature 0.6, diversity score is 44.4 (details in Appendix, Table 6). This finding aligns with the expectation that higher temperatures, which increase token sampling flexibility, enhance the diversity of generated text. In contrast, other metrics remain largely unchanged or show minor variations. For instance, average text quality is 3.15 at both temperatures of 0.6 and 0.2 on IMDB dataset.

[5] In Table 2 we report distance only for true counterfactuals.

| | Grammar | | | | Cohesiveness | | | | Fluency | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | GPT | | Mistral | | GPT | | Mistral | | GPT | | Mistral | |
| | *0.2* | *1.0* | *0.2* | *1.0* | *0.2* | *1.0* | *0.2* | *1.0* | *0.2* | *1.0* | *0.2* | *1.0* |
| Expert | 2.90 | 2.94 | 4.81 | 4.74 | 2.99 | 2.99 | 4.74 | 4.66 | 2.99 | 2.99 | 3.91 | 3.91 |
| Crowd | 2.92 | 2.89 | 4.88 | 4.79 | 2.95 | 2.98 | 4.78 | 4.68 | 2.92 | 2.94 | 3.83 | 3.81 |
| Crest | 2.18 | 2.15 | 4.05 | 3.96 | 2.27 | 2.30 | 3.95 | 3.91 | 2.33 | 2.37 | 3.36 | 3.34 |
| GBDA | 2.16 | 2.18 | 3.92 | 3.82 | 2.38 | 2.40 | 4.00 | 3.89 | 2.37 | 2.35 | 3.44 | 3.46 |
| Mice | 2.71 | 2.73 | 4.55 | 4.44 | 2.81 | 2.82 | 4.40 | 4.35 | 2.79 | 2.81 | 3.77 | 3.75 |
| LLAMA-2 | **3.18** | **3.19** | **4.90** | **4.86** | **3.12** | **3.11** | **4.83** | **4.74** | **3.13** | **3.12** | **4.00** | **3.96** |

Table 3: Comparison of text quality evaluation using Mistral and ChatGPT (GPT-3.5 Turbo) with different temperatures (0.2 and 1.0) on IMDB dataset.

## 6 Comparison of LLMs for Text Quality Evaluation

Evaluating text quality with ChatGPT has been shown to be effective (Huang et al., 2023; Gilardi et al., 2023). However, such evaluations come at high costs, limited control and customization constraints, and lack transparency. Therefore, we investigate an open-source LLM, Mistral-7B (Jiang et al., 2023) as an evaluation proxy.

**Mistral-7B is a valid alternative to ChatGPT** To validate Mistral's evaluation capability, we use Mistral to evaluate the counterfactuals generated by all methods and compare the assessment scores with those from ChatGPT. Specifically, we compare the average scores, the Pearson correlation on the scores of each instance, and the Spearman correlation of the ranking of each method on all text quality metrics on both datasets and two temperature settings of 0.2 and 1.0. Table 3 shows that Mistral-7B generally assigns higher scores than ChatGPT across all text quality metrics, though their scores are correlated. The Pearson correlation on the scores of each instance from the two models ranges from moderate to strong, with coefficients from 0.4 to 0.7, regardless of temperature settings (details in Appendix, Fig. 4). This implies that a text with high scores from Mistral is likely to receive high scores from ChatGPT as well. Furthermore, Spearman's rank correlation coefficients on the scores between the two models range from 0.89 to 1.0 , indicating a very strong correlation and partly even exactly identical rankings (details in Appendix Table 5).

To further validate Mistral-7B-instruct as a text quality evaluation proxy, we analyzed textual quality metrics on SNLI across two labels: contradiction and entailment. We hypothesized that entailment pairs exhibit higher cohesiveness and fluency than contradiction pairs, as entailment implies a logical relationship between the sentences. Our evaluation confirms that entailment pairs score significantly higher in text quality, particularly in cohesiveness and fluency, across all methods and human-generated texts. Detailed results are provided in Appendix, Table 7.

Given the moderate to strong correlation with ChatGPT scores, very strong correlation in rankings and the validation of textual quality on the SNLI dataset, Mistral-7B is a viable alternative for comparative counterfactual method evaluation.

**Text quality evaluation is robust to temperature variations** Since temperature influences the performance of LLMs during inference (Wang et al., 2023a), we evaluate its impact on their evaluation capabilities. Our study finds that text quality evaluation results are robust to temperature changes for both Mistral-7B and ChatGPT. We find a very strong correlation (Pearsons $\rho > 0.8$) between evaluation scores for different temperatures of the same model (Appendix Figures 4 and 5). Furthermore, the absolute scores remain similar across temperatures, as shown in Table 3.

## 7 Conclusion

We propose CEval to standardize the evaluation of counterfactual text generation, emphasizing the importance of both counterfactual metrics and text quality. Our benchmark facilitates standardized comparisons and analyzes the strengths and weaknesses of individual methods. Initial results show that counterfactual methods excel in counterfactual metrics but produce lower-quality text, while LLMs generate high-quality text but struggle to reliably flip labels. Combining these approaches could guide future research, such as using target classifier supervision to enhance LLM outputs. The diversity in method performance highlights the need for robustness analyses of target classifiers with mul-

tiple methods. Our findings also suggest that the target classifier may be poorly calibrated, warranting further investigation. Finally, we demonstrate that text quality evaluation using LLMs is robust to temperature changes. Additionally, we show that open-source LLMs, like Mistral, can serve as alternatives to closed-source models, such as ChatGPT, for evaluating text quality, thereby overcoming weaknesses of closed-source models, such as API deprecation or high costs. This leads to CEval being a fully open-source Python library, encouraging the community to contribute additional methods and to ensure that future work follows the same standards. For future work, we plan to integrate LLMs specifically designed for evaluation, such as Prometheus (Kim et al., 2023), as an option for assessing text quality. Furthermore, instead of only considering the difference between instances to measure diversity, the diversity metric can be expanded to incorporate the particular types of changes, such as negation and word replacements.

## Limitations

We employ default hyperparameters for each method and straightforward prompts with LLMs, which may not be optimal for the task at hand and could be further improved by hyperparameter optimization and prompt engineering.

This benchmark solely evaluates the quality of counterfactual text for explanation tasks. Further research is required to evaluate the performance of this text in other downstream tasks such as data augmentation with counterfactual examples or improving the robustness of the model using counterfactual examples. Additionally, we evaluate the metrics with a single BERT-based classifier. While this classifier achieves state-of-the-art classification accuracy, our results indicate that it might not be well calibrated. Estimating to which extent our findings can be generalized requires a combination of multiple diverse classifiers in the benchmark and the application in downstream tasks.

A potential exposure of ChatGPT or Mistral to the human counterfactual dataset is unlikely to impact our results, as we used these models only for evaluating text quality rather than counterfactual generation. The exposure of LLAMA-2 to human counterfactuals remains uncertain. If such exposure occurred, it could potentially influence our results for LLAMA-2, as it would help to generate better (human-like) counterfactuals. However, Fig. 3 shows a considerable distance between human-generated and LLAMA-generated counterfactuals, suggesting a low likelihood of such influence.

## Ethics Statement

We use the publicly available datasets IMDB and SNLI, and employ the benchmark to evaluate existing counterfactual generation methods. None of these methods declared any ethical concerns. While the benchmark is designed to evaluate counterfactual generation methods to advance research in explainable AI, it could be misused to select the best counterfactual methods for generating potentially harmful content. One such harmful application scenario could be the generation of counterfactuals to evade a fake news detector. However, if such evasion would actually be possible without a drastic change of the semantics, the major risk stems from the counterfactual generation methods rather than from their benchmark comparison.

We strongly believe that a benchmark evaluation should be as open, fair, transparent and reproducible as possible. Therefore, we make all our source code (including benchmark evaluation and method implementation) publicly available[1] and include the option to evaluate text quality metrics with the open-source LLM Mistral-7B (cf. Section 6).

## References


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. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Nitay Calderon, Eyal Ben-David, Amir Feder, and Roi Reichart. 2022. DoCoGen: Domain counterfactual generation for low resource domain adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

*Papers)*, pages 7727–7746. 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. 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. Association for Computational Linguistics.

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

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. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Nuno M. Guerreiro and André F. T. Martins. 2021. SPECTRA: Sparse structured text rationalization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6534–6550. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. 2021. Gradient-based Adversarial Attacks against Text Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5747–5757. Association for Computational Linguistics.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion proceedings of the ACM web conference 2023*, pages 294–297.

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.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8018–8025. Section: AAAI Technical Track: Natural Language Processing.

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

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

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. Association for Computational Linguistics.

Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. 2021. Generate Your Counterfactuals: Towards Controlled Counterfactual Generation for Text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13516–13524. Number: 15.

David Martens and Foster Provost. 2014. Explaining data-driven document classifications. *MIS Quarterly*, 38(1):73–100.

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. 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. Lulu.com.

Catarina Moreira, Yu-Liang Chou, Chihcheng Hsieh, Chun Ouyang, Joaquim Jorge, and João Madeiras Pereira. 2022. Benchmarking Counterfactual Algorithms for XAI: From White Box to Black Box. ArXiv:2203.02399 [cs].

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. Association for Computational Linguistics.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617.

Martin Pawelczyk, Sascha Bielawski, Johan Van den Heuvel, Tobias Richter, and Gjergji. Kasneci. 2021. CARLA: A Python Library to Benchmark Algorithmic Recourse and Counterfactual Explanation Algorithms. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

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

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097. 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. 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. Association for Computational Linguistics.

Alexis Ross, Tongshuang Wu, Hao Peng, Matthew Peters, and Matt Gardner. 2022. Tailor: Generating and perturbing text with semantic controls. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3194–3213. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, and et. al. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martín Pereira-Fariña. 2021. A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access*, 9:11974–12001. Conference Name: IEEE Access.

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. Association for Computational Linguistics.

Chi Wang, Xueqing Liu, and Ahmed Hassan Awadallah. 2023a. Cost-effective hyperparameter optimization for large language model generation inference. In *International Conference on Automated Machine Learning*, pages 21–1. PMLR.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023b. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11. Association for Computational Linguistics.

Jiaxin Wen, Yeshuang Zhu, Jinchao Zhang, Jie Zhou, and Minlie Huang. 2022. AutoCAD: Automatically generate counterfactuals for mitigating shortcut learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2302–2317. 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. Association for Computational Linguistics.

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2660–2676. Association for Computational Linguistics.

## A Generated Text Comparison Example

Table 9 presents examples where the majority of methods were unsuccessful in altering the original label. While LLAMA-2 and human evaluators both identify nonsensical words within the text, other methods overlook this aspect. In this intricate example, human crowdsource agreement with the human expert is not notably high, as their concurrence is limited to the term nonsensical. However, the human expert's observations exhibit more alignment with other methods, such as modifying denigrate akin to LLAMA-2, and replacing Sorry or nonsense as observed in MICE.

## B Method Selection Criteria

| Method | Type | Classifier Access | Reproducible code | Problem Agnosticity |
|---|---|---|---|---|
| **MICE** | MF | ✓ | ✓ | ✓ |
| CF-GAN | CD | ✓ | ✗ | ✓ |
| Polyjuice | MF | ✓ | ✓ | ✗ |
| **GBDA** | CD | ✓ | ✓ | ✓ |
| DISCO | LLM | ✗ | ✗ | ✓ |
| AutoCAD | MF | ✓ | ✗ | ✓ |
| CORE | MF | ✗ | ✗ | ✗ |
| DoCoGen | MF | ✓ | ✓ | ✗ |
| Tailor (Ross et al., 2022) | MF | ✓ | ✓ | ✗ |
| **CREST** | MF | ✓ | ✓ | ✓ |
| GYC(Madaan et al., 2021) | CD | ✓ | ✗ | ✓ |
| FLARE | LLM | ✗ | ✗ | ✓ |

Table 4: Comparison of Methods. Methods of different types that meet all inclusion criteria are highlighted in **bold** and are included in the benchmark.

## C Correlation of Mistral and ChatGPT

| **Temperature** | **0.2** | **1.0** |
|---|---|---|
| Grammar | 1.0 | 0.89 |
| Cohesiveness | 0.94 | 0.89 |
| Fluency | 1.0 | 0.94 |

Table 5: Spearman correlation of method rankings assigned by the LLM models Mistral and ChatGPT across different temperature settings, demonstrating very strong correlation.

## D Effect of Temperature

We evaluate the effect of temperature on the counterfactual generation process and text quality. Table 6 shows the results of LLAMA-2 with three different temperatures: 0.2, 0.6, and 1.0. Lower temperatures imply a higher likelihood of selecting the most frequent tokens and a lower likelihood of selecting less frequent tokens. Consequently, diversity is low at lower temperatures and high at higher temperatures. Perplexity is also correlated with temperature, while other metrics do not show a clear correlation. On the other hand, Figures 4 and 5 show the correlations between the same model at different temperatures, as well as the correlations between different models across various metrics. We observe a very strong correlation within the same model and a moderate correlation when using different models, suggesting that the evaluation is robust with respect to temperature.

| | | **IMDB** | | | **SNLI** | | |
|---|---|---|---|---|---|---|---|
| | | *0.2* | *0.6* | *1.0* | *0.2* | *0.6* | *1.0* |
| **CF Metrics** | Flip Rate ↑ | 0.68 | 0.65 | **0.70** | 0.38 | 0.40 | 0.39 |
| | ΔProbability ↑ | 0.67 | 0.66 | **0.69** | 0.32 | **0.33** | **0.33** |
| | Perplexity ↓ | 40.6 | **39.1** | 41.3 | **54.9** | 55.2 | 57.0 |
| | Distance ↓ | 50.7 | **48.9** | 58.0 | **4.36** | 4.48 | 4.78 |
| | Diversity ↑ | 28.3 | 44.4 | **61.6** | - | - | - |
| **Text Quality** | Grammar ↑ | **3.20** | 3.18 | 3.18 | 3.76 | **3.77** | 3.68 |
| | Cohesiveness ↑ | 3.14 | **3.15** | 3.12 | **3.71** | 3.69 | 3.61 |
| | Fluency ↑ | 3.12 | 3.11 | **3.13** | 3.66 | **3.71** | 3.59 |
| | *Average* ↑ | **3.15** | 3.15 | 3.14 | 3.71 | **3.72** | 3.63 |

Table 6: Comparison of LLAMA-2 counterfactual generation with different temperatures (0.2, 0.6, and 1.0). Temperature primarily affects diversity, with minimal impact on other metrics.

| | LLAMA-2 | | | MICE | | | GBDA | | | CREST | | | Crowd | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *E* | *N* | *C* | *E* | *N* | *C* | *E* | *N* | *C* | *E* | *N* | *C* | *E* | *N* | *C* |
| Grammar | 4.89 | **4.94** | 4.57 | **4.79** | 4.67 | 4.41 | **4.12** | 4.00 | 3.50 | **4.40** | 3.84 | 3.35 | **4.84** | 4.84 | 4.70 |
| Cohesiveness | **4.29** | 4.12 | 2.01 | **4.26** | 3.47 | 2.31 | **2.86** | 2.33 | 1.58 | **3.19** | 1.97 | 1.55 | **4.08** | 3.94 | 3.06 |
| Fluency | **4.99** | 4.86 | 4.38 | **4.90** | 4.67 | 4.38 | **4.61** | 4.07 | 3.56 | **4.43** | 3.73 | 3.13 | **4.95** | 4.83 | 4.30 |
| *Average* | **4.61** | 4.50 | 3.40 | **4.53** | 4.06 | 3.42 | **3.62** | 3.20 | 2.62 | **3.90** | 2.96 | 2.48 | **4.42** | 4.33 | 3.83 |

Table 7: Textual quality metrics to verify the LLMs evaluation. *E*: Entailment, *N*: Neutral, *C*: Contradiction

| | Grammar | | | | Cohesiveness | | | | Fluency | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | GPT | | Mistral | | GPT | | Mistral | | GPT | | Mistral | |
| | *0.2* | *1.0* | *0.2* | *1.0* | *0.2* | *1.0* | *0.2* | *1.0* | *0.2* | *1.0* | *0.2* | *1.0* |
| Crowd | 3.58 | 3.56 | 4.62 | **4.61** | 3.60 | 3.53 | 3.77 | **3.73** | 3.56 | 3.51 | **4.48** | **4.43** |
| Crest | 2.71 | 2.66 | 3.71 | 3.73 | 2.74 | 2.72 | 3.03 | 3.00 | 2.70 | 2.66 | 3.88 | 3.82 |
| GBDA | 2.29 | 2.31 | 3.27 | 3.22 | 2.03 | 2.08 | 2.10 | 2.20 | 2.17 | 2.16 | 3.37 | 3.31 |
| Mice | 3.33 | 3.32 | 4.44 | 4.39 | 3.31 | 3.31 | 3.50 | 3.46 | 3.33 | 3.34 | 4.38 | 4.29 |
| LLAMA-2 | **3.68** | **3.66** | **4.63** | 4.60 | **3.61** | **3.55** | 3.64 | 3.63 | **3.59** | **3.58** | 4.44 | 4.36 |

Table 8: Comparison of text quality evaluation using Mistral and ChatGPT (GPT-3.5 Turbo) with different temperatures (0.2 and 1.0) on SNLI dataset.

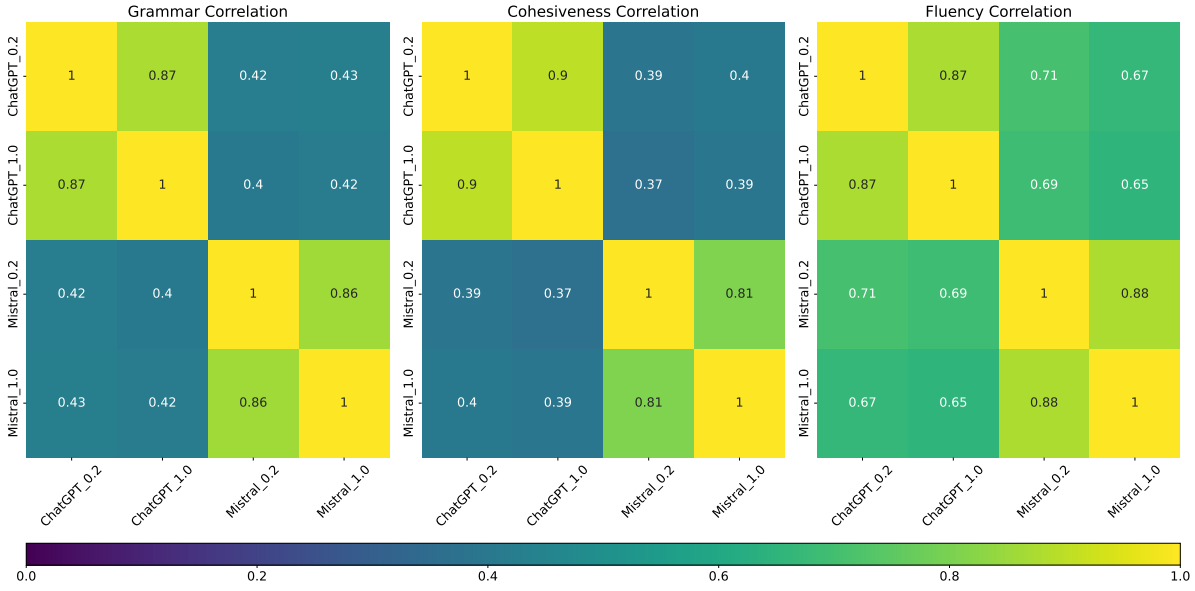


Figure 4: Pearson correlation between Mistral and ChatGPT in text quality evaluation with different temperatures (0.2 and 1.0) on the IMDB dataset. The same model with the different temperatures exhibits a strong correlation, meanwhile different models show a moderate correlation in evaluating text quality for counterfactual generation.

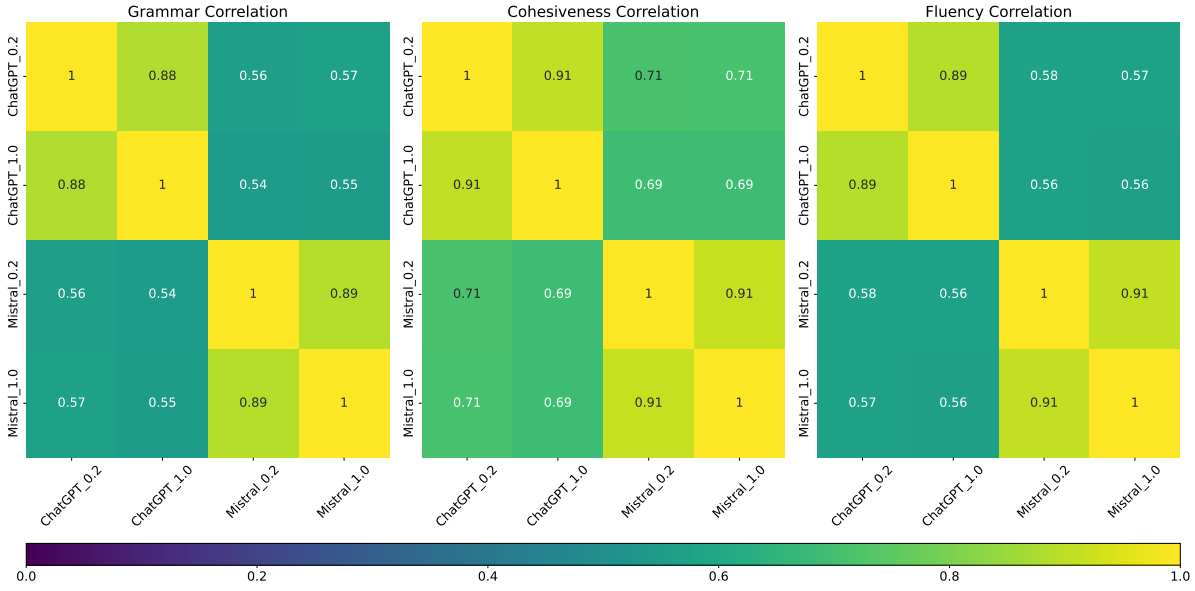


Figure 5: Pearson correlation between Mistral and ChatGPT in text quality evaluation with different temperatures (0.2 and 1.0) on the SNLI dataset. Text quality evaluation results of the same model with the different temperatures are strongly correlated; results from different models are moderately correlated.

| Method | Text | Predicted Label |
|---|---|---|
| Original | This movie frequently extrapolates quantum mechanics to justify nonsensical ideas, capped by such statements like "we all create our own reality". Sorry, folks, reality is what true for all of us, not just the credulous. The idea that "anything's possible" doesn't hold water on closer examination: if anything's possible, contrary things are thus possible and so nothing's possible. This leads to postmodernistic nonsense, which is nothing less than an attempt to denigrate established truths so that all ideas, well-founded and stupid, are equal. To quote sci-fi writer Philip K. Dick, who put it so well, "Reality is that which, when you stop believing in it, doesn't go away. | Negative |
| LLAMA-2 | This movie frequently extrapolates quantum mechanics to justify ~~nonsensical~~ inspiring ideas, capped by such statements like "we all create our own reality". Sorry, folks, reality is what true for all of us, not just the credulous. The idea that "anything's possible" doesn't hold water on closer examination: if anything's possible, contrary things are thus possible and so nothing's possible. This leads to postmodernistic nonsense, which is nothing less than an attempt to ~~denigrate~~ celebrate established truths so that all ideas, well-founded and stupid, are equal. To quote sci-fi writer Philip K. Dick, who put it so well, "Reality is that which, when you stop believing in it, doesn't go away. | Negative |
| MICE | This ~~movie frequently extrapolates~~ excellent film has nothing more to say than to condemn quantum mechanics to ~~justify~~ betray nonsensical ideas, ~~capped~~ accompanied by such statements like "we all create our own reality". ~~Sorry,~~ Hey, folks, reality is what true for all of us, not just the credulous. The idea that "anything's possible" doesn't hold water on closer examination: if anything's possible, contrary things are thus possible and so nothing's possible. This ~~leads~~ movie is intended to ~~postmodernistic nonsense, which~~ teach believers that embracing reality is nothing ~~less~~ than an ~~attempt~~ excuse to denigrate established truths so that all ideas, well-founded and ~~stupid~~ , doubtful , are equal. To quote sci-fi writer Philip K. Dick, who put it so well, "Reality is that which, when you stop believing in it, doesn't go away. | Positive |
| GBDA | this movie ~~frequently~~ still extrapolates ~~quantum~~ ~~mechanics~~ experimental depression to ~~justify~~ such nonsensical ideas, ~~capped~~ accompanied by ~~such~~ false statements like like " we all create our own reality " . ~~sorry,~~ ~~folks, reality~~ ". nonetheless, nonetheless, irony is ~~what~~ ~~true~~ what, for all of us, not just the credulous. the idea that " anything's possible " doesn't ~~hold~~ ~~water~~ ~~on~~ ~~closer~~ ~~examination:~~ go away for subjective assumptions : if anything's possible, ~~contrary~~ everyday things are ~~thus~~ ever possible and so ~~nothing's~~ everything's possible. this ~~leads~~ applies to ~~postmodernistic~~ postmodernist ~~nonsense,~~ authenticity, which is nothing less than an attempt to denigrate ~~established~~ ~~truths~~ cultural reality so that ~~all~~ those ideas, ~~well-founded~~ well - beautiful and ~~stupid,~~ beautiful, are ~~equal.~~ wonderful. to quote ~~sci-fi~~ ~~writer~~ sci - fi critic philip k. dick, who ~~put~~ points it so well, ~~"reality~~ " comedy is that which, when you stop believing in ~~it,~~ yourself, doesn't go away. | Positive |
| CREST | ~~This~~ ~~movie~~ ~~frequently~~ ~~extrapolates~~ ~~quantum~~ ~~mechanics~~ ~~to~~ ~~justify~~ ~~nonsensical~~ ~~ideas,~~ ~~capped~~ A quantum-sensical thriller, accompanied by such ~~statements~~ films like "we all create our own ~~reality"~~ world" . Sorry, folks, this ~~reality~~ is not what true for all of us, not just the the ~~credulous~~ credulity .The idea that "anything's possible" doesn't hold water on closer-to-end: ~~closer~~ ~~examination:~~ if anything's possible, contrary things are thus possible and so ~~nothing's~~ that's possible. ~~This~~ ~~leads~~ However, there is no less reason to definately ~~postmodernistic~~ ~~nonsense,~~ ~~which~~ ~~is~~ ~~nothing~~ ~~less~~ ~~than~~ ~~an~~ ~~attempt~~ ~~to~~ ~~denigrate~~ established ~~truths~~ characters so that ~~all~~ ~~ideas,~~ the characters, well-founded and ~~stupid,~~ well-meaning, are ~~equal.~~ not. To quote sci-fi writer Philip K. Dick, who put ~~it~~ ~~so~~ ~~well,~~ this film together, ~~"Reality,~~ "Really, ~~is~~ ~~that~~ ~~which,~~ when you stop believing in it, it doesn't go away. | Negative |
| Expert | This movie frequently extrapolates quantum mechanics to justify ~~nonsensical~~ futurist ideas, capped by such inspiring statements like "we all create our own reality". ~~Sorry,~~ Yes, folks, reality is this, what true for all of us, is what we just see, not just the credulous. The idea that "anything's possible" ~~doesn't~~ hold water even on closer examination: if anything's possible, contrary things are thus possible and so nothing's ~~possible.~~ possible but we're talking alternate universe. This leads to postmodernistic ~~nonsense,~~ theories, which ~~is~~ are nothing less than an attempt to ~~denigrate~~ elevate established truths so that all ideas, well-founded and stupid, are equal. To quote sci-fi writer Philip K. Dick, who put it so well, "Reality is that which, when you stop believing in it, doesn't go away. | Negative |
| Crowd | This movie frequently extrapolates quantum mechanics to justify ~~nonsensical~~ wise ideas, capped by such statements like "we all create our own reality". Sorry, folks, reality is what true for all of us, not just the credulous. The idea that "anything's possible" doesn't hold water on closer examination: if anything's possible, contrary things are thus possible and so nothing's possible. This leads to postmodernistic nonsense, which is nothing less than an attempt to denigrate established truths so that all ideas, well-founded and stupid, are equal. To quote sci-fi writer Philip K. Dick, who put it so well, "Reality is that which, when you stop believing in it, doesn't go away." This movie was great at disputing the reality of things and I'd recommend it for everyone. | Negative |

Table 9: Example for which most methods failed to flip the label