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.

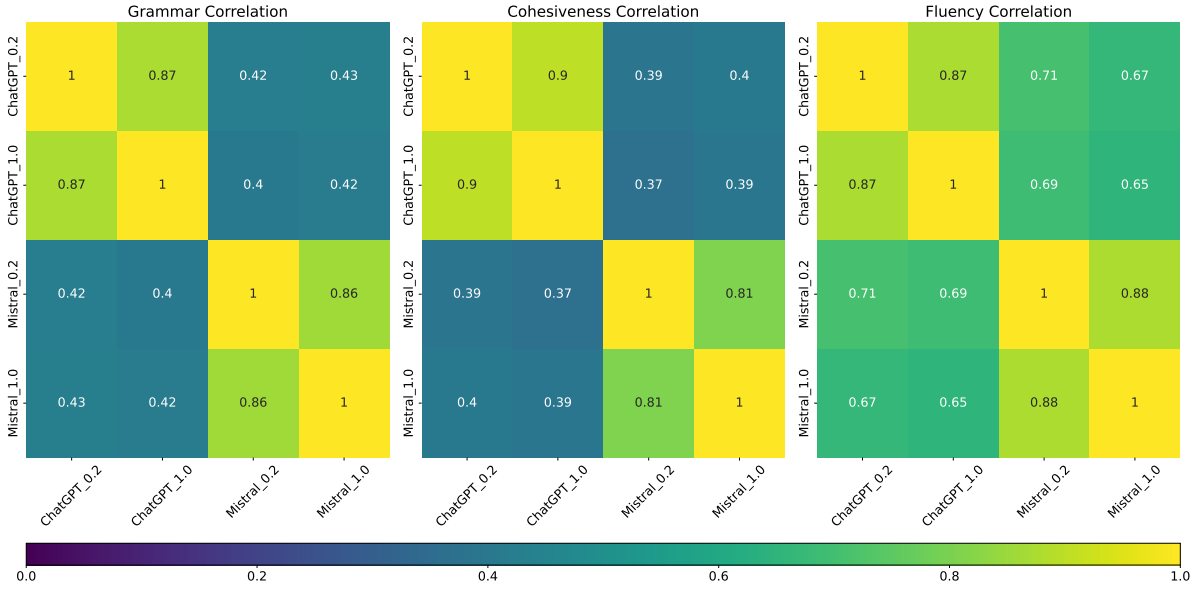

Figure 4: Pearson correlation between Mistral and ChatGPT in text quality evaluation with different temperatures (0.2 and 1.0) on the IMDB dataset. The same model with the different temperatures exhibits a strong correlation, meanwhile different models show a moderate correlation in evaluating text quality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

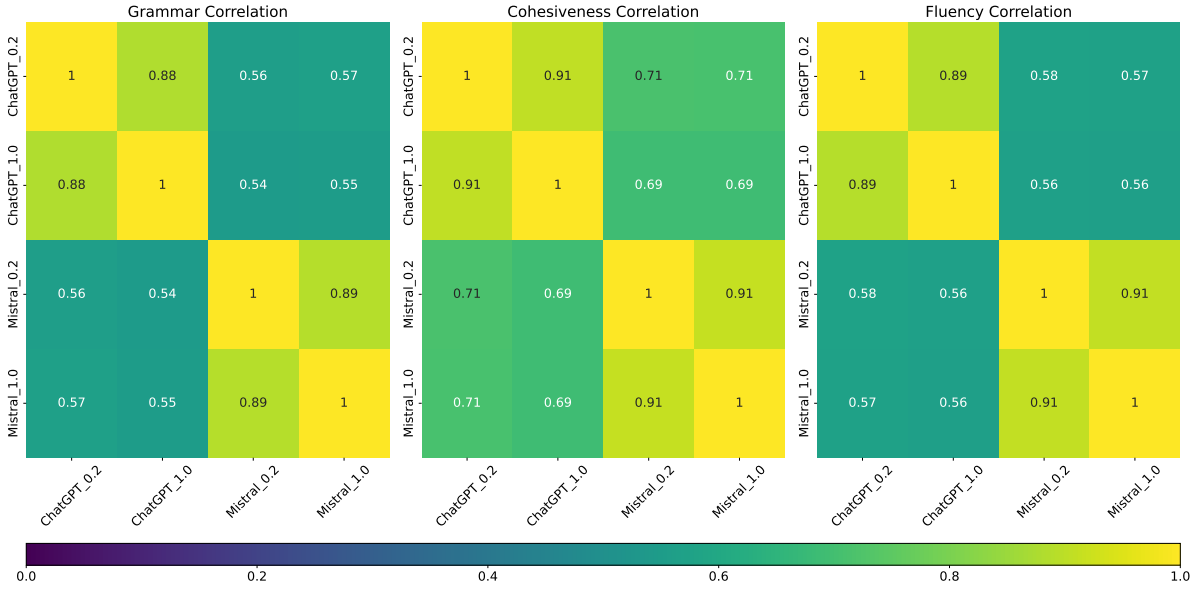

Figure 5: Pearson correlation between Mistral and ChatGPT in text quality evaluation with different temperatures (0.2 and 1.0) on the SNLI dataset. Text quality evaluation results of the same model with the different temperatures are strongly correlated; results from different models are moderately correlated.

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