Natural Language Counterfactual Explanations in Financial Text Classification

Anonymous ACL submission

Abstract

The use of Large Language Model (LLM) classifiers in finance and other high-stakes domains calls for a high level of trustworthiness and explainability. We focus on Counterfactual Explanations (CE), a form of Explainable AI that explains a model's output by proposing an alternative to the original input that changes the classification. We use three types of CE generators for LLM classifiers and assess the quality of their explanations on a recent dataset consisting of central bank communications. We compare the generators using a selection of quantitative and qualitative metrics. Our findings suggest that non-expert and expert evaluators prefer counterfactual methods that apply minimal changes to the texts; however, the methods we analyze might not handle the domain-specific vocabulary well enough to generate plausible explanations for our task. We discuss shortcomings in the choice of evaluation metrics in the literature on text CE generators and propose refined definitions of the fluency and plausibility qualitative metrics.

1 Introduction

Large language models (LLMs) usage in specialist fields such as finance is growing. One specialist application of LLMs is the analysis of central bank monetary policy communications. Communications allow central banks to address factors like the inflation expectation which, in turn, influence the market's growth (Rozkrut et al., 2007). It is thus beneficial for stakeholders to analyze those communications and adequately react to incoming market changes. On the other hand, central bankers aim to communicate their policy stance to markets clearly, avoiding confusion in their interpretation, a difficult task considering the highly nuanced nature of these texts (Cieslak and Schrimpf, 2019). A common classification of these texts distinguishes between *hawkish* (tighter policy) or *dovish* (looser policy). As these different policies may be described with different languages, researchers have studied the use of LLMs to perform automatic hawkish/dovish classification from text (Wang, 2023).

However, as with any use of black-box models in high-stakes domains, it is necessary to provide means of explainability and trustworthiness of these models. However, explaining a complex LLM can be difficult, especially when the domain the model operates on is challenging. Counterfactual explanations (CEs) (Wachter et al., 2018) aim to explain a classification made by a machine learning model by perturbing the original input to generate a counterfactual that yields some desired model prediction. While many methods for generating text CEs for LLM classifiers exist, most are trained and evaluated on simple tasks and datasets (Wu et al., 2021). Additionally, the processes in which these methods are evaluated often rely on imprecise quantitative and qualitative metrics.

In this paper, we investigate the usability of CE generators for LLMs on a task from the financial domain and improve the evaluation thereof. We contribute to the field by: 1. Evaluating several categories of CE generators by comparing them from a quantitative and qualitative perspective, considering opinions from domain experts. 2. Showing that the state-of-the-art text counterfactual generators do poorly on texts from specialist domains. 3. Highlighting the need for human evaluation and improving the qualitative text CE evaluation metrics by providing more precise definitions.

2 Related Work

With the abundance of text CE techniques used in the literature, we consider a wide array of methods for generating LM counterfactuals. We split the text CE generators into three categories based on how they produce counterfactual explanations.

The first category of generators, *LLM-assisted generation*, contains generators that use another

LLM as a surrogate model to produce counterfactuals. Polyjuice (Wu et al., 2021), for example, uses a GPT-2 model fine-tuned for several counterfactual generation tasks. Polyjuice is often used as a baseline text CE generator, including in this work.

The second category, *latent perturbation and decoding*, uses the latent representation of the factual sentence as a starting point and perturbs that representation to generate a counterfactual sentence. The counterfactual sentence embedding is then decoded into a piece of text. As a representative example for this category, we investigate PPLM (Dathathri et al., 2019), which uses, firstly, a surrogate attribute model to optimize the generation for a target attribute; and then a fluency model (ex. GPT-2) to ensure high fluency.

In the third category, *sequential generation*, generators first mask a part of the input text, and then infill the masked area with new tokens. In this work, we consider the RELITC generator (Betti et al., 2023) as a representative example for that category. In RELITC, token masks are generated by using LM feature attribution methods. The tokens with the highest attributions are masked and infilled by a Conditional Masked Language Model (CMLM) one by one, conditioned on a target class.

This three-way split allows us to include the different characteristics of text counterfactual generators encountered in the literature while keeping our evaluation feasible. We perform quantitative and qualitative evaluations on explanations generated by a member of each category.

The evaluation methods used in the literature on text CE generators are often related to the desiderata sought by the authors of the methods. The researchers often try to optimize the *minimality* of the explanations, generating changes as minimal as possible while still explaining the model. The researchers measure the changes with distance metrics, such as edit distance (Gilo and Markovitch, 2024; Wu et al., 2021; Ross et al., 2021; Betti et al., 2023; Dixit et al., 2022), tree edit distance (Gilo and Markovitch, 2024; Wu et al., 2021; Madaan et al., 2021), embedding distance (Betti et al., 2023), or semantic measures of similarity (Robeer et al., 2021). Another desideratum is the *validity* of the CEs. Its measures include success rate or accuracy (Wu et al., 2021; Madaan et al., 2021; Ross et al., 2021; Betti et al., 2023; Robeer et al., 2021). A third popular choice is the *fluency* of the CE measured using model perplexity (Dathathri et al., 2019; Madaan et al., 2023; Treviso et al., 2023; Fern and Pope, 2021). Finally, numerous methods try to optimize the *plausibility* of the counterfactual (Gilo and Markovitch, 2024; Madaan et al., 2021; Yang et al., 2020) or its adherence to the underlying class conditional distribution.

The use of perplexity as a fluency metric has previously been criticized (Meister and Cotterell, 2021), and metrics like accuracy or distance might lead to adversarial-looking explanations (Altmeyer et al., 2023). While commonly used, these metrics might be insufficient for assessing text CEs, it is thus necessary to evaluate the CEs qualitatively.

To perform a qualitative analysis of CEs, authors usually employ human evaluation. Human evaluators are often asked to judge the *fluency* of the CEs (Dathathri et al., 2019; Wu et al., 2021; Madaan et al., 2021; Ross et al., 2021; Betti et al., 2023), frequently described as judging whether a sentence "reads like good English". Some works include metrics like the *fidelity* of the explanations or *content preservation* (Madaan et al., 2021; Betti et al., 2023; Wu et al., 2019), evaluating whether it does not stray from the original topic. Others opt to measure if the generated text explanations reflect the target domain well by measuring their *plausibility* and *reasonability* (Yang et al., 2020).

In the literature, those qualitative metrics often lack rigor in the definitions or are missing the definitions themselves. This looseness in definitions can confuse the annotators, leading to incorrect annotations. We mitigate that issue by giving more precise definitions of the fluency and plausibility metrics (Appendix A) inspired by previous works (Ma and Cieri, 2006; Altmeyer et al., 2024).

3 Experiments

We use a dataset comprised of speeches, meeting minutes, and press conference transcripts of the Federal Open Market Committee (FOMC) (Shah et al., 2023). These texts are split into 1984 train and 494 test sentences, and annotated into 3 classes: dovish, hawkish, and neutral. Shah et al. (2023) train a RoBERTa-large classifier on their dataset. We use this model in our experiments.

For each text in the dataset, we assign a random counterfactual label for which a CE should be generated. We use the three generators, Polyjuice, PPLM, and RELITC to generate counterfactual explanations. For each generator, we generate several CEs, which are then classified by the classifier. As a final explanation, we select the text with the

Generator	Perplexity ↓	Perpl. ratio	Edit dist. ↓	Tree dist. ↓	Emb. dist. ↓	Impl. ↓	Succ. rate ↑
Polyjuice	90.98 (172.1)	1.80 (4.6)	0.31 (0.3)	19.67 (24.0)	**20.32** (3.7)	33.64 (4.6)	0.34 (0.5)
PPLM	**36.97** (16.9)	**0.78** (0.5)	0.69 (0.5)	36.94 (10.3)	20.88 (3.7)	**32.18** (4.0)	0.51 (0.5)
RELITC	100.94 (125.2)	1.67 (1.2)	**0.14** (0.1)	**10.72** (12.2)	21.96 (3.9)	33.30 (3.9)	**0.74** (0.4)

Table 1: Averages and standard deviations of the quantitative metrics calculated for counterfactual explanations of texts in the test set. A perfect result for the perplexity ratio metric is thought to be 1 (Bhan et al., 2023).

highest score assigned by the classifier if the class matches the assigned random target class. Otherwise, a random counterfactual is chosen.

We perform three experiments using the FOMC dataset. In the first experiment, we use quantitative metrics to evaluate methods. We select the following metrics: perplexity, perplexity ratio, edit distance, semantic tree edit distance, embedding distance, and plausibility. The metrics are chosen to represent a wide array of desiderata of the generators.

The second and third experiments consist of human evaluations. The first round of evaluations consists of native English speakers recruited via the Prolific platform. In this round, we ask the evaluators to judge the fluency of the generated sentences on a scale ranging from 1 (poor) to 5 (good). This experiment allows us to perform a large-scale evaluation. We evaluate 100 factual sentences, with each sentence receiving 5 evaluations, which yields 1,500 non-expert human evaluations in total across all three generators.

In the second round of human evaluations, we ask central bank employees to evaluate the counterfactuals using the fluency and plausibility measures. This expert evaluation aims to give us insight into the usefulness of these explanations in the field of financial text classification. Additionally, we ask the experts for comments on the counterfactual text to learn the qualities that an expert might look for in those explanations.

We release additional information about the survey such as the participant recruitment, instructions, or the survey consent form in Appendix B.

4 Results and Discussion

The results of the quantitative metrics are presented in Table 1. The results do not point toward a method that performs best out of the three, although specific patterns emerge. PPLM, which uses a GPT-2 model in its generation phase and optimizes for its fluency, performs best for perplexity-based metrics. Similarly, RELITC, which tries to minimize the fraction of perturbed tokens in creating a counterfactual, has the best results for the Levenshtein distance and the tree edit distance. It also achieves the highest flip rate out of the three methods. Interestingly, Polyjuice achieves the best results solely for the embedding distance metric.

These results are computed for all counterfactuals, including ones that do not succeed at flipping the label. We observe no major differences when we instead used successful CEs only in Table 5. This result shows that the quantitative metrics capture characteristics of different CE generators, but are not precise enough to discern successful and unsuccessful explanations in this domain.

We present the results of the human-annotated qualitative metrics in Table 2. Even though the RELITC generator receives some of the worst results for the perplexity metrics, it produces the most fluent texts according to both the sampled groups, while the opposite applies to the PPLM method. Surprisingly, the only metrics that reflect the same ranking of the methods as the fluency are the Levenshtein and tree edit distances. These findings suggest that perplexity might not be a reliable metric for evaluating CEs.

All the methods receive rather low scores in the expert-judged plausibility metric, generally achieving lower than sufficient scores. Despite the RELITC method producing the most fluent counterfactuals, the experts assign the highest plausibility scores to Polyjuice. This stems from the RELITC's

	Annotators			
	Non-exp.	N-e. 5 CE	Expert	
Generator	Fluency	Fluency	Fluency	Plausibility
PPLM	2.86 (0.7)	2.48 (0.5)	2.26 (0.5)	1.83 (0.3)
Polyjuice	3.40 (0.9)	3.44 (0.7)	3.45 (0.9)	**2.45** (0.7)
RELITC	**3.43** (0.8)	**3.96** (0.5)	**3.90** (0.6)	2.12 (0.3)

Table 2: Results of the human annotation of the counterfactuals using the qualitative metrics. Each counterfactual receives five ratings which we average. We display the averages of those averages and their standard deviations. Since the expert evaluations are performed on a subset of five samples, we show the fluency scores the non-experts give on the same set of samples.

	Perplexity	Perp. ratio	Edit Dist.	Tree edit dist.	Emb. dist.	Impl.
Fluency (non exp.)	-0.06 (0.2)	-0.03 (0.5)	**-0.21 (0.0002)**	**-0.21 (0.0003)**	0.03 (0.7)	0.06 (0.3)
Fluency (exp.)	0.12 (0.6)	0.14 (0.6)	**-0.56 (0.016)**	**-0.56 (0.015)**	-0.25 (0.3)	0.13 (0.3)
Plausibility	0.32 (0.2)	0.02 (0.9)	-0.12 (0.6)	-0.28 (0.3)	-0.12 (0.6)	0.28 (0.3)

Table 3: Pearson correlation coefficients and p-values between the quantitative and qualitative metric results.

misuse of domain-specific words, as reported in the experts' comments analyzed in Section 4.1.

A surprising result is the closeness between the experts' and non-experts' grading of fluency. The highest difference between the average grades in Table 2 is 0.22 for PPLM, while Polyjuice's fluency scores differ by only 0.01. This indicates that the fluency metric might not depend on the annotator's background and that non-experts' ratings can give reliable results even in specialist domains.

Since the perplexity metric is highly dependent on the training data of the model used to compute it (Meister and Cotterell, 2021), we investigate whether the choice of models affects our results, however, apart from different scales of perplexities, we notice no major differences between the results.

Even though the expert and non-expert fluency scores are nearly the same and they dictate the same hierarchy as the distance metrics, we find that there is little apparent correlation between the qualitative and quantitative results. We analyze the agreement between our results performing statistical tests comparing the quantitative and qualitative metrics. We use the Pearson correlation coefficient to measure the dependence between metrics.

Table 3 shows that there is no strong correlation between plausibility and quantitative metrics. The correlation of fluency with both the edit distance metrics show low p-values, suggesting a significant (negative) correlation. This result is in line with the earlier findings, which suggest that the methods that introduce fewer edits tend to be rated higher.

4.1 Expert Insights on Counterfactuals

As a part of our expert evaluation questionnaire, we ask our respondents to elaborate on the shortcomings in "*the semantics of the [counterfactual] sentence, its structure, or content*".

Over half of the comments regarding Polyjuice counterfactuals relate to the lack of relevance of the introduced changes. A part of the comments addresses the grammatical errors or an "*... entirely different subject*" that replaces the original one in the Polyjuice counterfactuals.

PPLM generates tokens until reaching a fixed limit, therefore, it is possible that the generator does not finish a sentence within the given limit. Thus, the main critique is that the counterfactuals are unfinished. Similarly to Polyjuice, PPLM introduced errors in the sentences, however, unlike Polyjuice, PPLM's propensity to use domain-specific words introduces more room for errors in the usage thereof. PPLM was also criticized for making the counterfactuals too conversational.

RELITC is similar to PPLM in that it learns the domain-specific terms through its CMLM and then uses them to generate a counterfactual. This introduces room for error. The experts comment on sentences where RELITC introduces domain-specific terms that are factually incorrect, contradict the contents of the sentence, or make the tone of the counterfactual unclear or conversational.

5 Conclusions

We evaluate a range of text counterfactual explanation generators on a dataset stemming from a financial background. We consider desiderata employed by the authors of the CE generators and aim to answer what qualities of these generators are the most useful when applied to the financial domain. We run three experiments, one using quantitative and two with qualitative metrics using human evaluators. Our findings suggest that the methods that focus on the minimality of the changes, creating counterfactuals that are more fluent than those that solely focus on the CE validity. However, the plausibility of these counterfactuals is often low. With additional comments from domain experts, we find that an incorrect use of domain-specific terms can diminish the plausibility of the explanations. Surprisingly, using counterfactual generators that do not use specialist words might be preferable in specialist domains. Additionally, we analyze a range of quantitative metrics used in evaluating CE generators in NLP and find that many of them do not measure the generators' desiderata well. We emphasize the need to use human annotation when evaluating text CEs and provide more precise qualitative metric definitions.

Limitations

Our work is not without limitations. We select only 3 out of the multiple text counterfactual generation methods. While we attempt to consider a wide range of techniques used in the field, it is not feasible to evaluate all existing methods.

A limiting factor in using some methods is that some require additional data besides texts and labels for training purposes. PPLM's bag-of-words (BoW) attribution model requires a curated list of words for calculating the text generation direction (Dathathri et al., 2019). Similarly, the work by Yang et al. (2020) uses BoW for an infilling task similar to the one used in RELITC. Our work analyzes the feasibility of using text counterfactual methods in real-life applications where additional data might not be available. At the same time, we acknowledge that studying those methods might bring further insights into the field. Another limitation inherent to the FOMC dataset is the lack of ground-truth counterfactuals. We considered this in designing our study since datasets acquired from real-life data usually do not contain samples with exact semantic matches in their target classes. While this consideration makes our evaluation more realistic, it does not let us evaluate the results with machine translation metrics like BLEU or include the ground-truth counterfactuals in expert evaluation. Furthermore, one cannot use some of the retrieval-based generators without factual-counterfactual pairs.

Another limitation stems from the use of a single dataset in our evaluations. While we solely consider financial text classification, the texts in this field use specific terms that might or might not be present in the pre-training data for the foundational models used in the methods we evaluate. One could gain more insight from performing similar evaluations on texts from other specialist domains, such as medicine or legal texts.

Acknowledgments

The authors would like to thank the many central bank employees, including staff from the Federal Reserve Board of Governors, who devoted time to analyzing text counterfactuals in our experiments. The research responses, analysis, and conclusions set forth are those of the authors and do not indicate concurrence by other members of the research staff or the Federal Reserve Board of Governors.

References

Patrick Altmeyer, Giovan Angela, Aleksander Buszydlik, Karol Dobiczek, Arie Van Deursen, and Cynthia C. S. Liem. 2023. Endogenous Macrodynamics in Algorithmic Recourse. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 418–431, Raleigh, NC, USA. IEEE.

Patrick Altmeyer, Mojtaba Farmanbar, Arie van Deursen, and Cynthia C. S. Liem. 2024. Faithful Model Explanations through Energy-Constrained Conformal Counterfactuals. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10):10829–10837.

Lorenzo Betti, Carlo Abrate, Francesco Bonchi, and Andreas Kaltenbrunner. 2023. Relevance-based Infilling for Natural Language Counterfactuals. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, pages 88–98, New York, NY, USA. Association for Computing Machinery.

Milan Bhan, Jean-Noël Vittaut, Nicolas Chesneau, and Marie-Jeanne Lesot. 2023. TIGTEC: Token Importance Guided TExt Counterfactuals. In Francesco Bonchi, Elena Baralis, Manuel Gomez Rodriguez, Claudia Plant, and Danai Koutra, editors, *Machine Learning and Knowledge Discovery in Databases: Research Track*, volume 14171, pages 496–512. Springer Nature Switzerland, Cham.

Anna Cieslak and Andreas Schrimpf. 2019. Non-monetary news in central bank communication. *Journal of International Economics*, 118:293–315.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, J. Yosinski, and Rosanne Liu. 2019. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. *ArXiv*.

Tanay Dixit, Bhargavi Paranjape, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. CORE: A Retrieve-then-Edit Framework for Counterfactual Data Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2964–2984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiaoli Fern and Quintin Pope. 2021. Text Counterfactuals via Latent Optimization and Shapley-Guided Search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5578–5593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Gilo and Shaul Markovitch. 2024. A General Search-Based Framework for Generating Textual Counterfactual Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18073–18081.

Xiaoyi Ma and Christopher Cieri. 2006. Corpus Support for Machine Translation at LDC. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

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

Nishtha Madaan, Diptikalyan Saha, and Srikanta Bedathur. 2023. Counterfactual Sentence Generation with Plug-and-Play Perturbation. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 306–315, Raleigh, NC, USA. IEEE.

Clara Meister and Ryan Cotterell. 2021. Language Model Evaluation Beyond Perplexity. *arXiv preprint*. ArXiv:2106.00085 [cs].

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.

Marek Rozkrut, Krzysztof Rybiński, Lucyna Sztaba, and Radosław Szwaja. 2007. Quest for central bank communication: Does it pay to be "talkative"? *European Journal of Political Economy*, 23(1):176–206.

Agam Shah, Suvan Paturi, and Sudheer Chava. 2023. Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6664–6679, Toronto, Canada. Association for Computational Linguistics.

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.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *arXiv preprint*. ArXiv:1711.00399 [cs].

Yifei Wang. 2023. Aspect-based Sentiment Analysis in Document – FOMC Meeting Minutes on Economic Projection. *arXiv preprint*. ArXiv:2108.04080 [cs].

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.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Mask and Infill: Applying Masked Language Model for Sentiment Transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 5271–5277, Macao, China. International Joint Conferences on Artificial Intelligence Organization.

Linyi Yang, Eoin Kenny, Tin Lok James Ng, Yi Yang, Barry Smyth, and Ruihai Dong. 2020. Generating Plausible Counterfactual Explanations for Deep Transformers in Financial Text Classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6150–6160, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. *arXiv preprint*. ArXiv:2205.01068 [cs].

A Improved Qualitative Metrics

We provide two qualitative metric definitions: fluency and plausibility. To establish them, we adapt existing metric definitions. For fluency, we adapt a definition by Ma and Cieri (2006) by including common mistakes often made by a CE generator:

> A fluent segment is one that is grammatically well-formed; contains correct spellings; adheres to the common use of terms, titles and names; contains properly capitalized letters; and is intuitively acceptable. Unfinished sentences also impact the fluency of a segment.

For plausibility, we take the interpretation by Altmeyer et al. (2024) and adapt it to the text domain:

> A plausible counterfactual segment adheres well to samples seen in the real data distribution, and the target sentiment of the target class. The changes made to the factual, considering the meaning and context of the edited words, should also fit the target domain.

B Additional Survey Information

B.1 Participant Recruitment

We recruited the participants of our survey through the crowdsourcing platform Prolific. We recruit native English speakers from the UK and USA who have at least high-school level education. The participants were compensated with the standard for Prolific rate of 9 GBP per hour.

B.2 Informed Consent Form

You are being invited to participate in a [...] research study titled Evaluating Language Model Explanations in Specialist Fields. This study is being done by [the authors] from the [organization].

The purpose of this research study is to assess the usability of modern language model explainability tools in generating texts in specialist fields, such as finance. This study will take you approximately 15 minutes to complete. The data will be used for evaluating a counterfactual explanation method. We will be asking you to rate pieces of text on a number of criteria using a 1 to 5 scale, and describe your reasoning in open questions.

As with any online activity the risk of a breach is always possible. To the best of our ability your answers in this study will remain confidential. We will minimize any risks by only collecting your personal information for the purpose of verification of the identity of the respondents. In our research we will pseudonymize your identity and solely use the answers to the questions relating to text assessment. The survey data will be stored on a [...] drive at [the organization] and all personal information will be destroyed after the end of the thesis project.

Your participation in this study is entirely voluntary and you can withdraw at any time. You are free to omit any questions.

Contact details for the corresponding researcher: [the details]

By submitting a response to this survey you agree to this Opening Statement and to your response being used for the research described above, and for your de-identified answers to be included in the final data set that will be publicly available when the research is published. I understand that once my response has been submitted my data will have been processed in such a way that it is no longer possible for it to be withdrawn.

B.3 Survey Topic Introduction

Counterfactual Explanations are a form of explainable AI aiming to explain a classification made by a Machine Learning model by proposing an alternative to the original input. Imagine you write a text that you intend to be perceived as positive, but a sentiment analysis Language Model doesn't find it quite convincing. Through a counterfactual explanation, we can generate a text which could better reflect the intended tone.

Your task:

We will present you with several counterfactual sentences generated via different means. On each page, we will show you an original (factual) sentence and three variants of counterfactuals. We will ask you to **grade the sentences** you see using the following criteria:

Fluency: A fluent segment is one that is grammatically well-formed; contains correct spellings; adheres to the common use of terms, titles and names; contains properly capitalized letters; and is intuitively acceptable. Unfinished sentences also impact the fluency of a segment.

Please rate the texts using this definition of fluency. A text should receive a score of:

- 5/5 if it follows this definition completely.
- 3/5 if there are several mistakes but the text still is interpretable.
- 1/5 if it is not fluent or grammatically correct English.

`For expert evaluation only:`

Plausibility: A plausible counterfactual segment adheres well to samples seen in the real data distribution, and the target sentiment of the target sentence class. The changes made to the factual, considering the meaning and context of the edited words, should also fit the target domain.

Please rate the texts using this definition of plausibility. A text should receive a score of:

- 5/5 if it follows this definition completely.
- 3/5 if there are several mistakes but the text reflects the right sentiment.
- 1/5 if the changes are nonsensical.

These criteria will also appear at the end of each page.

In an **open question**, we will ask you to describe what qualities that you might look for in a text

like this are missing. Your comment can refer to the semantics of the sentence, its structure, or its contents. If you do not have any comments you can also leave the answer empty.

The order of the methods used for each question will be randomized.

B.4 Sample Non-Expert Question

Grade the following sentences using the Fluency criterion. You can find the grading criterion at the bottom of the page.

Sentence 1

For equities, a stock's price-earnings ratio is a standard benchmark used to measure how well a company's financials compare to its peers. for the sake of comparison, a company can be

Fluency

- Very bad (1/5)
- Bad (2/5)
- Sufficient (3/5)
- Good (4/5)
- Very good (5/5)

The participants were shown the definition of fluency introduced in Appendix A

B.5 Sample Expert Question

Consider the following segment originally classified as **neutral**:

This lack of congressional momentum could be interpreted as lack of congressional support for inflation targeting, or it could merely reflect a more neutral absence of strong opinions.

Please rate the counterfactuals aiming to rewrite the segment with **dovish** as target class. You can find the grading criteria at the bottom of the page.

Neutral Factual

This lack of congressional momentum could be interpreted as lack of congressional support for inflation targeting, or it could merely reflect a more neutral absence of strong opinions.

Dovish Counterfactual 1

This lack of congressional momentum could be interpreted as lack of congressional support for the president's executive orders. as the president himself has said he will not be issuing a single executive order during his first 100

Fluency

- Very bad (1/5)
- Bad (2/5)
- Sufficient (3/5)
- Good (4/5)
- Very good (5/5)

Plausibility

- Very bad (1/5)
- Bad (2/5)
- Sufficient (3/5)
- Good (4/5)
- Very good (5/5)

Considering the counterfactual from the previous question, describe what qualities that you might look for in a text like this are missing. Your comment can refer to the semantics of the sentence, its structure, or contents. If you do not have any comments you can also leave the answer empty.

The participants were shown the definitions of fluency and plausibility introduced in Appendix A

C Scientific Artifacts and Licensing

As described in Section 3, we use the FOMC communications dataset[1] by Shah et al. (2023). The authors' original license is `cc-by-nc-4.0`, which we fully adhere to. For the purpose of our experiments, we generate a dataset with counterfactual labels and release it in the *Hugging Face* platform[2] under the `cc-by-nc-4.0` license. We share our codebase used to generate the data and evaluate the models under the `MIT License`.

[1] huggingface.co/datasets/gtfintechlab/fomc_communication

[2] huggingface.co/datasets/TextCEsInFinance/fomc-communication-counterfactual

D Alternative Models for Perplexity Calculation

	facebook/opt-125m		gpt2		lxyuan/distilgpt2-finetuned-finance	
	Perplexity	Perpl. ratio	Perplexity	Perpl. ratio	Perplexity	Perpl. ratio
Polyjuice	107.06 (291.9)	1.90 (7.9)	90.98 (172.1)	1.80 (4.6)	104.06 (150.3)	1.62 (3.84)
PPLM	**36.07** (15.9)	**0.68** (0.4)	**43.90** (23.5)	**0.78** (0.5)	**43.89** (23.5)	**0.69** (0.4)
RELITC	108.86 (153.8)	1.52 (0.8)	100.95 (125.2)	1.67 (1.2)	111.99 (142.0)	1.52 (1.0)

Table 4: Comparison of perplexity-based metrics computed using three language models. The base GPT-2, an Open Pretrained Transformer (OPT) (Zhang et al., 2022) opt-125m (https://huggingface.co/facebook/opt-125m), and a GPT-2 model fine-tuned on four financial datasets (https://huggingface.co/lxyuan/distilgpt2-finetuned-finance).

The PPLM generator includes a GPT-2 in its fluency optimization and decoding steps. Due to the fact that we use the same model for calculating our main results in Table 1, we want to test whether the choice of the model for calculating perplexity affects the resulting perplexity scores substantially. We analyze the effect a LM has on the resulting perplexities by calculating the average perplexity achieved by each of the three generators when using different models for perplexity.

E Quantitative Results of Successful Counterfactuals

	Perplexity	Perp. ratio	Edit dist.	Tree dist.	Embedding dist.	Impl.
Polyjuice	99.64 (227.0)	1.91 (4.6)	0.36 (0.3)	22.10 (21.7)	**20.35** (4.1)	**29.06** (3.4)
PPLM	**36.64** (16.2)	**0.77** (0.4)	0.76 (0.6)	36.25 (6.7)	20.69 (3.7)	29.56 (2.9)
RELITC	104.04 (130.2)	1.68 (1.3)	**0.12** (0.1)	**9.90** (13.2)	21.84 (3.8)	33.35 (3.5)

Table 5: Quantitative results computer over results containing only successful counterfactuals.

F Sample Expert Comments

	Text	Expert comments
Factual	At the conclusion of this discussion, the Committee voted to authorize and direct the Federal Reserve Bank of New York, until it was instructed otherwise, to execute transactions in the System Account in accordance with the following domestic policy directive: The information reviewed at this meeting suggests that the expansion in economic activity is still robust.	
Polyjuice	At the conclusion of this discussion, the committee voted to authorize and direct the federal reserve bank of new york, until it was instructed otherwise, to execute transactions in the system account in accordance with the following domestic policy directive: the information **was not** suggests that the expansion in economic activity is still robust.	**1:** "Language is off. The negation at the end makes the statement unclear.", **2:** "Again all capital letters are missing. This time, the last sentence is also incorrect"was not suggests" is clearly a mistake". This mistake makes the whole message impossible to understand.", **3:** "The last clause is not grammatically correct. Otherwise it does come across a bit more dovish."
PPLM	At the conclusion of this discussion, the committee voted to authorize and direct the federal reserve bank of new york, until it was instructed otherwise, to execute transactions in **securities that are not covered by the exchange act**.	**1:** "There now is a completely different meaning at the end of the statement.", **2:** "Again capital letters are missing, and the second sentence is incomplete. But at least the first sentence can be understood and sounds dovish (execute transactions in additional securities)", **3:** "There is an incomplete sentence at the end of the excerpt. It also loses the link to the current state of the economy and so isn't more dovish"
RELITC	At the conclusion of this discussion, the committee voted to authorize and direct the federal reserve bank of new york, until it was instructed otherwise, to execute transactions in the system account in accordance with the following domestic policy directive : the information reviewed at this meeting suggests that the **impact of the response** is still robust.	**1:** "There is a change of meaning in the last sentence which makes it less clear.", **2:** "All capital letter are missing, but the rest of the text seems to be correct. In terms of content, it is not clear at all, in particular the sentence "the impact of the response is still robust".", **3:** "The vagueness of 'impact of the response' makes it difficult to extract the message or signal this would try to send."

Table 6: Sample counterfactuals and the expert comments regarding them. Factual label: *neutral*, target label: *dovish*. Changes introduced in the counterfactuals, except for word capitalization, are **highlighted**.