Natural Language Counterfactual Explanations in Financial Text Classification: A Comparison of Generators and Evaluation Metrics

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 CE methods that apply minimal changes; however, the methods we analyze might not handle the domain-specific vocabulary well enough to generate plausible explanations. 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 (LLM) usage in specialist fields is growing. One specialist application of LLMs is the analysis of central bank monetary policy communications. Communications allow central banks to address factors such as inflation expectations that influence market growth (Rozkrut et al., 2007). In adjusting their own expectations, market participants closely monitor these communications for any signals that may indicate policy 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). The policy stance of a central bank can be broadly described as either *hawkish* (tighter policy) or *dovish* (looser policy). Since the bank's current stance is typically reflected in its communications, researchers have studied the use of LLMs to automatically classify press releases, meeting minutes and speeches as hawkish or dovish (Wang, 2023).

As with any use of black-box models in high-stakes domains, it is necessary to provide explainability and trustworthiness of these models. However, explaining predictions of an LLM can be difficult, especially when they operate in challenging domains. Counterfactual explanations (CE) (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. There are many methods to generate counterfactuals for LLM classifiers, but most have been trained and evaluated on generic tasks and datasets (Wu et al., 2021). In addition, the methods' evaluations often rely on imprecise quantitative and qualitative metrics.

In this paper, we evaluate CE generators for LLMs on a task from the financial domain. 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 perform 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 proposed in the literature, we consider a wide array of methods for generating text counterfactuals. We split the text CE generators into three categories based on how they produce counterfactual explanations.

The first category of generators, which we call *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 generator, including in this work.

The second category, *latent perturbation and decoding*, uses the latent representation of the factual sentence and perturbs it to generate a counterfactual embedding. The counterfactual embedding is then decoded into text. As a representative example for this category, we investigate PPLM (Dathathri et al., 2019), which uses a surrogate attribute model to optimize generation for a target class and 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 fill it with new tokens. In this work, we consider the RELITC generator (Betti et al., 2023) as a representative example. RELITC uses feature attribution to generate token masks. The tokens are then filled in with 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 the evaluation in line with the scope of this work.

The evaluation methods used in the literature on text CE generators are often related to the desiderata sought by the authors of the methods. Researchers often try to optimize for *minimality*, aiming for minimal perturbations that yield valid explanations. The size of the perturbations is typically measured using 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 *validity*, that is the success rate or accuracy of explanations (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 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 lead to adversarial-looking CEs (Altmeyer et al., 2023). Although commonly used, these metrics might be insufficient for assessing text CEs. To address this insufficiency, researchers have occasionally relied on qualitative evaluations performed by humans.

For example, human evaluators have been asked to judge the *fluency* of the CEs in numerous studies (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". In other works, humans have been asked to assess the *fidelity* or *content preservation* of explanations (Madaan et al., 2021; Betti et al., 2023; Wu et al., 2019) also referred to as *plausibility* and *reasonability* (Yang et al., 2020), to evaluate if they fall into the original topic.

These qualitative metrics are often not rigorously defined, if they are defined at all. Unclear definitions can confuse annotators, leading to incorrect annotations. We mitigate this issue by providing more precise definitions of *fluency* and *plausibility* to our evaluators (Appendix A) inspired by Ma and Cieri (2006) and Altmeyer et al. (2024).

3 Experiments

We use a dataset comprised of speeches, meeting minutes, and press conference transcripts from the Federal Open Market Committee (FOMC) (Shah et al., 2023). These texts are split into 1984 train and 494 test sentences and categorized into 3 classes: dovish, hawkish, and neutral. Shah et al. (2023) train a RoBERTa-large classifier on this dataset, which we use 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 CEs. For each generator, we generate several CEs, which are then classified by the classifier. As a final explanation, we select the text with the highest classification score if the class matches the assigned target class. Otherwise, a random counterfactual is chosen.

We perform three experiments using the FOMC dataset. In the first experiment, we use quantitative metrics for evaluation. We select the following metrics: perplexity, perplexity ratio, edit distance, semantic tree edit distance, embedding distance, implausibility, and faithfulness.

The second and third experiments involve human evaluations. For the first round of evaluations, we have recruited native English speakers via the Prolific platform. In this round, we ask the evaluators

Generator	Perplexity ↓	Perpl. ratio	Edit dist. ↓	Tree dist. ↓	Emb. dist. ↓	Implausib. ↓	Faithful. ↑	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)	0.18 (0.4)	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)	0.34 (0.6)	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)	**0.54** (0.6)	**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).

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 of 100 factual sentences, with each sentence receiving 5 evaluations, yielding 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 CEs for fluency and plausibility. With this expert evaluation, we aim to understand the properties of CEs sought after by experts, as well as the overall quality of these explanations in financial text classification.

We release additional information about the survey in Appendix B.

4 Results and Discussion

We present the results of the quantitative metrics in Table 1. The results do not point toward a method that performs best out of the three, although specific patterns emerge.[1] PPLM, which uses a GPT-2 model in its generation phase and optimizes for its fluency, performs best for perplexity-based metrics.[2] Similarly, RELITC, which tries to minimize the fraction of perturbed tokens, has the best results for the edit distance, flip rate and faithfulness metrics. Polyjuice achieves the best results solely for the embedding distance metric.

While the quantitative metrics capture characteristics of different CE generators, we are interested in understanding how the emerged patterns relate to human evaluations presented in Table 2.

Regarding fluency, experts' and non-experts' gradings are broadly aligned. The highest difference between the average grades in Table 2 is 0.22 for PPLM, while Polyjuice's fluency scores differ only by 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.

With the exception of distance-based metrics, quantitative metrics do not align with human evaluations for fluency. For example, even though the RELITC generator receives some of the worst results for the perplexity metrics, it produces the most fluent texts according to both groups of evaluators, while the opposite applies to PPLM.

Concerning plausibility, we find that counterfactuals receive less than sufficient expert ratings. Despite RELITC producing the most fluent counterfactuals, experts assign the highest plausibility scores to Polyjuice. This stems from the RELITC's misuse of domain-specific words, as reported in the experts' comments analyzed in Section 4.1.

Even though the expert and non-expert fluency scores are nearly the same and dictate the same hierarchy as the distance metrics, there is little apparent correlation between the qualitative and quantitative results.[3] Table 3 shows no strong correlation between plausibility and quantitative metrics. The correlation of fluency with both edit distance metrics shows low p-values, suggesting a significant (negative) correlation. This result is in line with our earlier findings, which suggest that methods that introduce fewer edits tend to be rated higher. We note that this result is different from the findings of previous work (Nguyen et al., 2024), which we attribute to the fact that we are investigating a specific

[3]We used the Pearson correlation coefficient to measure the dependence between metrics.

	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.

[1]Results are computed for all counterfactuals, including ones that do not succeed at flipping the label. We find no major differences when using only successful CEs (Table 6).

[2]The perplexity metric is highly dependent on the training data of the LLM used to compute it (Meister and Cotterell, 2021). Investigating whether the choice of models affects our results, we find no major differences between them (Table 7).

	Perplexity	Perp. ratio	Edit Dist.	Tree edit dist.	Emb. dist.	Implausib.
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.

domain. In more generic domains, a wider range of simple changes might still pass as plausible.

In summary, our findings indicate that many existing quantitative metrics are not reliable indicators for evaluating text counterfactual explanations.

4.1 Expert Insights on Counterfactuals

As 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*".

More than half of the comments regarding Polyjuice CEs relate to the lack of relevance of the introduced changes. Some comments address grammatical errors or an "... *entirely different subject*" that replaces the original in the Polyjuice CEs.

PPLM introduced errors in the sentences, too; however, unlike Polyjuice, PPLM's propensity to use domain-specific words introduces more room for errors in the usage thereof. The main critique of PPLM is unfinished CEs. PPLM generates tokens until reaching a fixed limit, making it possible that the generator does not finish a sentence. PPLM was also criticized for making the CEs 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, again introducing room for error. 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.

4.2 Faithfulness and Plausibility Trade-off

In our analysis, we take into account the trade-off in choosing faithfulness or plausibility as a main desideratum of a CE generator. We construct a simple counterfactual generator inspired by retrieve-and-generate (RAG) approaches (Dixit et al., 2022) using the `GPT-4o` model. The prompt of our *pseudo-RAG* generator includes few samples from the factual and target classes and the sample to generate a CE for (Appendix D). We rerun our quantitative metrics experiment including this generator. This method achieves the best success rate and produces seemingly plausible CEs, however, it performs worse than RELITC for the edit distance metrics. A plausible but unfaithful generator can be useful as a tool to generate high-quality text that changes the prediction of a model; however it does not contribute to gaining knowledge about the classifier (Agarwal et al., 2024; Altmeyer et al., 2024). An explanation with low plausibility and high faithfulness might not be realistic enough, especially in specialist domains. Thus, a balance between the two desiderata must be achieved (Lu and Ma, 2024). In CEs for LLMs this is not trivial – numerous approaches strive to increase the plausibility of their explanations and try to flip the label by producing a large number of CEs. Approaches like RELITC or PPLM take the important step towards faithfulness and introduce a link to the classifier in the process of generating a CE.

5 Conclusions

In this work, we evaluate a range of text CE generators on a financial dataset. We consider desiderata employed by the authors of the CE generators and aim to answer what qualities of these generators are the most sought after when applied to the financial domain. We run three experiments, one with quantitative and two with qualitative metrics using human evaluators. Our findings suggest that methods that apply minimal changes create counterfactuals that are more fluent than those that focus solely on CE validity. However, the plausibility of these explanations 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 CE generators that do not use specialist words might be preferable in specialist domains, suggesting that faithfulness can be as important as plausibility. In addition, we analyze a range of quantitative metrics used in evaluating CE generators in NLP. We find that most of them do not measure the generators' desiderata well and that they rarely agree with the expert ratings. 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 studied here 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 (Dixit et al., 2022). This limitation has also caused us to use a simplified measure of faithfulness (Zheng et al., 2024) instead of ones specifically developed for text counterfactuals (Atanasova et al., 2023).

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. Furthermore, one could gain more insight from performing similar evaluations on texts from other specialist domains, such as medicine or legal texts. By developing a more generalized benchmark, the applicability of counterfactual methods on specialist domains in general can be evaluated. The findings gathered from our work and a general analysis of CEs in specialist domains, can be leveraged to design a counterfactual generator better suited for this domain type.

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

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. Plausibility: On the (Un)Reliability of Explanations from Large Language Models. *arXiv preprint*. ArXiv:2402.04614.

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.

Andre Artelt, Valerie Vaquet, Riza Velioglu, Fabian Hinder, Johannes Brinkrolf, Malte Schilling, and Barbara Hammer. 2021. Evaluating Robustness of Counterfactual Explanations. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 01–09, Orlando, FL, USA. IEEE.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness Tests for Natural Language Explanations. *arXiv preprint*. ArXiv:2305.18029.

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.

Eoin M. Kenny and Mark T. Keane. 2021. On Generating Plausible Counterfactual and Semi-Factual Explanations for Deep Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11575–11585.

Xiaolei Lu and Jianghong Ma. 2024. Does Faithfulness Conflict with Plausibility? An Empirical Study in Explainable AI across NLP Tasks. *arXiv preprint*. ArXiv:2404.00140.

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].

Van Bach Nguyen, Christin Seifert, and Jörg Schlötterer. 2024. CEval: A Benchmark for Evaluating Counterfactual Text Generation. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 55–69, Tokyo, Japan. Association for Computational Linguistics.

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.

Arnaud Van Looveren and Janis Klaise. 2021. Interpretable Counterfactual Explanations Guided by Prototypes. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 650–665, Cham. Springer International Publishing.

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].

John S. White, Theresa A. O'Connell, and Francis E. O'Mara. 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, USA.

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].

Xu Zheng, Farhad Shirani, Zhuomin Chen, Chaohao Lin, Wei Cheng, Wenbo Guo, and Dongsheng Luo. 2024. F-Fidelity: A Robust Framework for Faithfulness Evaluation of Explainable AI. *arXiv preprint*. ArXiv:2410.02970.

A Improved Qualitative Metrics

We provide two qualitative metric definitions: fluency and plausibility. To establish them, we adapt existing metric definitions.

In designing a task for human evaluators, it is necessary to consider how they interpret the task's prompts. Especially in a field like text interpretation, non-experts can understand a value like fluency in many different ways. Not providing a definition or using a very broad one may lead to annotators essentially evaluating different qualities. It is thus crucial to establish a robust and detailed definition upfront.

The qualitative metric of fluency can be traced back to early works on machine translation that tried to unify what constitutes fluency in a machine-generated text. White et al. (1994) describe fluency measurement as determining whether a piece of text "reads like good English", disregarding the semantic correctness of the sentence and giving it a rating on a n-point scale. At the same time, longer and more defined definitions exist, such as "A fluent segment is one that is grammatically well formed; contains correct spellings; adheres to the common use of terms, titles and names; is intuitively acceptable; and can be sensibly interpreted by a native speaker of English." by Ma and Cieri (2006).

Many of the recent works on text CEs (Dathathri et al., 2019; Wu et al., 2021; Madaan et al., 2021; Ross et al., 2021; Betti et al., 2023) evaluate their texts using a very similar notion of fluency as that defined by White et al. (1994). However, the notion of fluency has been described vaguely or inconsistently. Other works use different names like *naturalness* (Robeer et al., 2021; Treviso et al., 2023) to measure essentially the same thing.

We derive a fluency definition by modifying one by Ma and Cieri (2006). The generators we use can produce texts where word capitalization is omitted or where the text changes abruptly. This impacts the quality of the generated text. To omit ambiguity in case a counterfactual contains these errors, we specify that they will also impact fluency. Our final definition is as follows:

> 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.

The definition of plausibility outside of counterfactual explanations for language models often refers to the explanation's similarity or closeness to the original data distribution (Kenny and Keane, 2021). Indeed, many approaches to generating counterfactual explanations that emphasize the interpretability (Van Looveren and Klaise, 2021) or the robustness (Artelt et al., 2021) of the explanations employ strategies that enhance the adherence of the counterfactual to a certain class.

Altmeyer et al. (2024) define plausibility as:

> Let $\mathcal{X}|y^+ = p(x|y^+)$ denote the true conditional distribution of samples in the target class y^+. Then for x' to be considered a plausible counterfactual, we need: $x' \sim \mathbf{X}|y^+$.

Some related works that evaluate counterfactual explanations for language models seemingly forgo the definition of the plausibility metric entirely (Madaan et al., 2021), or ask the annotators "*how plausible (mainly in terms of grammar and comprehension)*" (Yang et al., 2020), missing the definition of the metric. Gilo and Markovitch (2024) who generate counterfactuals for a movie review dataset, ask annotators to grade whether the CE is a movie review or not. While this definition considers the original data distribution, it does not include the adherence of the counterfactual to the target class.

We adapt the definition by Altmeyer et al. (2024) 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[4] 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[5] 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`.

D Pseudo-RAG Generator

The size of new LLMs, such as the GPT-4 or Mistral-7B, prevents these models from being used as part of counterfactual generators, such as the GPT-2 in the PPLM. Due to that, the quality of the contextual generators using older models might be lower compared to that possible with the use of new LLMs. The newer LLMs have been shown to perform even better than their predecessors on zero-shot tasks, so one might assume that their accuracy and their performance for a counterfactual generation task might also be good. We therefore performed an experiment using the `GPT-4o` model to create a counterfactual generator and tested it on the FOMC task.

In designing our proof-of-concept method, we take inspiration from the retrieval-augmented generation (RAG) technique. In RAG, an LLM is supplied with a number of texts or documents that the user's query relates to; the model is then tasked with answering the user's query using the contents of the documents. While several CE generators use RAG or similar concepts (Dixit et al., 2022), they all rely on data sets that contain factual-counterfactual pairs, pairs that the FOMC dataset, among many others, lacks. This is a severe limitation because the generators can only be applied to a handful of specific datasets. In view of this limitation, we decide to supply the LLM with several examples of factual sentences from both the factual class and the target class creating a pseudo-RAG generator. We then ask the model to create a new counterfactual that could be classified to the target class by making as few changes to the original sentence as possible.

Table 4 shows the results of the generation of text counterfactuals using our pseudo-RAG method.

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

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

A classification Machine Learning model classifies texts into three classes: DOVISH, HAWKISH and NEUTRAL. Your task is to transform a QUERY sentence that was classified as `{label}` into a COUNTERFACTUAL that should be classified as `{target}`. You can replace, remove or add words, but you should keep the amount of changes to minimum, only performing up to 5 changes. You can use the EXAMPLE `{factual label}` and EXAMPLE `{target label}` sentences as examples how sentences belonging to those classes might look like. You should generate only one COUNTERFACTUAL sentence.

EXAMPLE `{factual label}`:
`{factual class examples}`

EXAMPLE `{target label}`:
`{target class examples}`

`{factual label}` QUERY: `{factual}`

`{target label}` COUNTERFACTUAL:

Figure 1: Prompt of the proof-of-concept pseudo-RAG generator.

As in the previous experiments, we designed the experiment to use a reasonable number of generation attempts, generating five counterfactuals per factual text. Even with the small amount of counterfactuals generated, the method achieves the highest flip rate score of 0.88. Although the perplexity results for PPLM are still better than in this proof of concept, we get the second lowest perplexity out of the four generators. The results of the other metrics are comparable to the rest of the methods. A notable result is the implausibility metric, where this model receives the highest score, meaning that the embeddings of the counterfactuals generated by this model are furthest away from the factuals in our data set. A surprising result is that the pseudo-RAG method achieves the best result of the faithfulness metric, even though the method

has no input from the classifier. This result can be explained by the rather high reliance of the metric on the success rate of the CEs (Zheng et al., 2024) which likely causes the metric to be biased. On the other hand, the quality of the generated sentences, as shown in Table 5, is seemingly the best out of all generators. This is probably due to the complexity of the model and the higher quality of the outputs compared to the other models.

Similarly to Polyjuice, pseudo-RAG has no information about the classifier. However, similarly to PPLM, it has no restrictions with regard to the amount of tokens generated, so the changes it generates are not controlled, which can cause the counterfactuals to stray away from the factual sentences. The poor results of the implausibility metric, combined with the high accuracy and seemingly high quality of the counterfactuals, lead us to believe that involving the classifier and generating counterfactuals is important, especially for classification tasks. Although this model can be useful for generating new data sets or new training sets, it is unlikely to be used to generate useful explanations for classification tasks. It is hard to evaluate the faithfulness of the explanations generated using this method; however, it is likely to see the LLM introduce its own biases rather than explain our classifier.

D.1 Pseudo-RAG Generator Results

Generator	Perplexity ↓	Perpl. ratio	Edit dist. ↓	Tree dist. ↓	Emb. dist. ↓	Implausib. ↓	Faithful. ↑	Succ. rate ↑
Pseudo-RAG	74.00 (38.8)	1.37 (0.5)	0.29 (0.1)	19.40 (11.5)	24.86 (4.0)	32.39 (2.9)	**0.36** (0.5)	**0.88**
Polyjuice	86.49 (79.9)	1.58 (1.3)	0.26 (0.3)	17.36 (15.3)	**24.78** (3.5)	**31.56** (2.7)	0.00 (0.4)	0.36
PPLM	**37.11** (15.2)	**0.76** (0.4)	0.56 (0.2)	37.48 (7.3)	24.97 (4.4)	32.09 (4.5)	0.04 (0.7)	0.52
RELITC	86.72 (71.6)	1.54 (1.0)	**0.13** (0.1)	**11.00** (7.0)	25.83 (3.7)	32.18 (3.1)	0.32 (0.6)	0.80

Table 4: Results for the quantitative metrics including the Pseudo-RAG method. Averaged over 25 factual-counterfactual rows.

D.2 Sample Pseudo-RAG Explanations

Factual	Counterfactual
(Dovish) Moreover, inflation was running at a fairly low rate and quite possibly would edge down a little further over coming quarters.	**(Neutral)** Moreover, inflation was running at a **stable** rate and **might fluctuate somewhat** over coming quarters.
(Hawkish) In their discussion of the balance-of-risks sentence in the press statement to be issued shortly after this meeting, all the members agreed that the latter should continue to express, as it had for every meeting earlier this year, their belief that the risks remained weighted toward rising inflation.	**(Neutral)** In their discussion of the balance-of-risks sentence in the press statement to be issued shortly after this meeting, all the members agreed that the **statement** should continue to express, as it had for every meeting earlier this year, their belief that the risks **were balanced**.
(Neutral) The uncertainty about the threshold unemployment rate also suggests a differing degree of intensity in the response of monetary policy to deviations of inflation and output to their respective targets.	**(Dovish)** The uncertainty about the threshold unemployment rate **highlights the need for stronger and more accommodating** monetary policy to **address** deviations of inflation and output **from** their respective targets.

Table 5: Sample outputs of the pseudo-RAG generator. Changes introduced in the counterfactuals, except for word capitalization, are **highlighted**.

E Quantitative Results of Successful Counterfactuals

	Perplexity	Perp. ratio	Edit dist.	Tree dist.	Embedding dist.	Implausib.	Faithful.
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)	0.49 (0.5)
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)	0.63 (0.5)
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)	**0.71** (0.5)

Table 6: Quantitative results computer over results containing only successful counterfactuals.

F Alternative Models for Perplexity Calculation

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.

	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 7: 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).

G 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 8: 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**.