CATfOOD: Counterfactual Augmented Training for Improving Out-of-Domain Performance and Calibration

Rachneet Sachdeva, Martin Tutek, Iryna Gurevych
Ubiquitous Knowledge Processing Lab (UKP Lab)
Department of Computer Science and Hessian Center for AI (hessian.AI)
Technical University of Darmstadt
www.ukp.tu-darmstadt.de

Abstract

In recent years, large language models (LLMs) have shown remarkable capabilities at scale, particularly at generating text conditioned on a prompt. In our work, we investigate the use of LLMs to augment training data of smaller language models (SLMs) with automatically generated counterfactual (CF) instances – i.e. minimally altered inputs – in order to improve out-of-domain (OOD) performance of SLMs in the extractive question answering (QA) setup. We show that, across various LLM generators, such data augmentation consistently enhances OOD performance and improves model calibration for both confidence-based and rationale-augmented calibrator models. Furthermore, these performance improvements correlate with higher diversity of CF instances in terms of their surface form and semantic content. Finally, we show that CF augmented models which are easier to calibrate also exhibit much lower entropy when assigning importance, indicating that rationale-augmented calibrators prefer concise explanations.[1]

1 Introduction

Ever since their introduction to the field of NLP, large language models (LLMs) have shown exceptional performance across a wide array of applications (Devlin et al. 2019; Brown et al. 2020; Wei et al. 2022b; *inter alia*). LLMs have frequently been utilized to enhance reasoning capabilities of smaller models (Li et al., 2022b), generate counterfactuals (CF) – minimally perturbed input instances – for data augmentation (Fryer et al., 2022; Paranjape et al., 2022), and have shown remarkable generalization capabilities, performing well on various tasks such as question answering (QA), complex reasoning, and code generation (Wei et al., 2022a; Black et al., 2022; Touvron et al., 2023). On the other hand, comparatively small language models (SLMs) such as BERT (Devlin et al., 2019) perform well on task specific data but their performance drops with a change in the data distribution (Koh et al., 2021; He et al., 2023) and they are frequently poorly calibrated, exhibiting under- or overconfidence in their predictions (Desai and Durrett, 2020; Kong et al., 2020; Guo et al., 2021; Jiang et al., 2021). In our paper, we examine how data augmentation with CFs of varying diversity improves out-of-domain (OOD) performance and model calibration of SLMs. For comparability to previous work, we perform our experiments in the extractive QA domain, but we believe our findings could generalize to other tasks given the remarkable versatility exhibited by LLMs (Wei et al., 2022a).

To alleviate the issue of poor OOD performance for QA, recent works have resorted to augmenting training data with *counterfactual* instances automatically generated by LLMs (Paranjape et al., 2022). Training on CF augmented data reduces model reliance on spurious features, which in turn improves generalizability (Sen et al., 2021). While Paranjape et al. (2022) fine-tune a T5-based model to generate

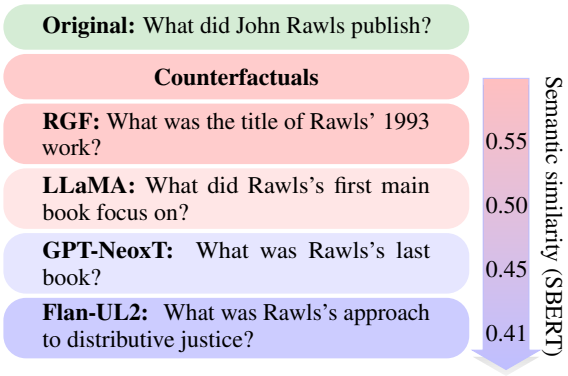

Figure 1: An illustration of the counterfactual samples (purple) for the input question (green) produced by the RGF baseline and our approaches using LLMs. While RGF produces a question closely related to the input, LLMs generate more diverse questions with respect to surface form and semantic content.

[1]We make our code available at: github.com/CATfOOD

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1876–1898

minimally different counterfactual instances with their Retrieve-Generate-Filter (RGF) approach, we leverage a range of more powerful LLMs such as Flan-UL2 (Tay et al., 2022) and LLaMA (Touvron et al., 2023). Owing to the extensive training of these LLMs on diverse data, coupled with their enhanced generative capabilities, we hypothesize they will produce counterfactual instances *more diverse* with respect to their surface form and semantic content, covering a broader part of the input space, further improving robustness and generalization. A sample of diverse CF instances is shown in Figure 1, highlighting variations in focus, temporality, specificity, and domain knowledge.

In other work, Ye and Durrett (2022) improve the calibration of SLMs by leveraging features from *rationales*, explanations of the inner decision making process of the model, to train a calibrator model – a simple classifier which predicts whether the base model is correct or not. We hypothesize that CF augmented models possess more precise explanations of their decisions, as they are forced to consolidate the more complex discrepancies between instances and their CFs, which should in turn provide better information to the calibrator model and improve calibration. To better investigate the connection between model explanations and calibrator performance, we introduce semantic features – dense representations of the most important tokens from explanations – to calibrator models, consider a wider range of explainability methods, and measure whether characteristics of explanations – such as *comprehensiveness* and *sufficiency* (Chrysostomou and Aletras, 2022) are indicative of the models' calibration performance.

In our work, we present the first systematic and comprehensive study on the effect of diverse CFs for augmenting SLMs with respect to their OOD performance, explanation quality and calibration performance. Our experiments show that: (1) more diverse CFs improve OOD performance and model calibration in extractive QA by a large margin; (2) introducing rationale semantics from CF augmented models to calibrators improves calibration performance; and (3) rationale augmented calibrators prefer concise and informative explanations.

2 Related Work

2.1 Counterfactual Generation

Counterfactual instances have demonstrated their importance in evaluating the OOD generalization capabilities of LLMs (Bowman and Dahl, 2021) and in augmenting training data (Longpre et al., 2021). One major downside of works which tackle CF generation (Kaushik et al., 2020; Khashabi et al., 2020; Ribeiro et al., 2020) has been the prohibitive requirement for human annotators, which would manually perturb data instances to generate CFs – a setup both expensive and difficult to scale.

With the improvements brought forward by LLMs, the idea of automatically generating CFs with generative models has gained significant traction. In the QA setup, Ye et al. (2021) and Longpre et al. (2021) generate counterfactuals by substituting entity names with other plausible entity names. However, this approach requires heuristic methods or human re-labeling to derive the resulting label changes. More recent work (Paranjape et al., 2022) focuses on creating fluent, and automatically labeled CFs with minimal human supervision. Their method requires fine-tuning models for both question generation and answering, which restricts the diversity of generated CFs to only what exists within the fine-tuning dataset. On the other hand, our methodology utilizes LLMs pretrained on a diverse array of datasets that enables us to generate CFs with a broader range of knowledge and linguistic nuances, surpassing the limitations posed by fine-tuning on specific datasets. Gat et al. (2023) prompt LLMs to generate CFs by altering a specific attribute conveyed in the input text while confounding attributes are fixed. In contrast, our work emphasizes on generating diverse CF instances without the constraint of changing a specific input attribute. In summary, our work investigates the previously unexplored relationship between CF diversity and OOD performance.

2.2 Model Calibration

Estimating the uncertainty of SLMs is challenging due to limited training data available, especially under OOD settings (Desai and Durrett, 2020; Guo et al., 2021). While prior approaches to model calibration have used "meta-features" based on model confidence (Kamath et al., 2020) and input representations (Zhang et al., 2021), these techniques do not incorporate features from explanations which is the central focus of our work. In the OOD calibration scenario, recent works have explored the use of explanations during training (Li et al., 2022a), and data augmentation (Park and Caragea, 2022). However, these works mostly focus on calibration techniques, whereas token importance scores from

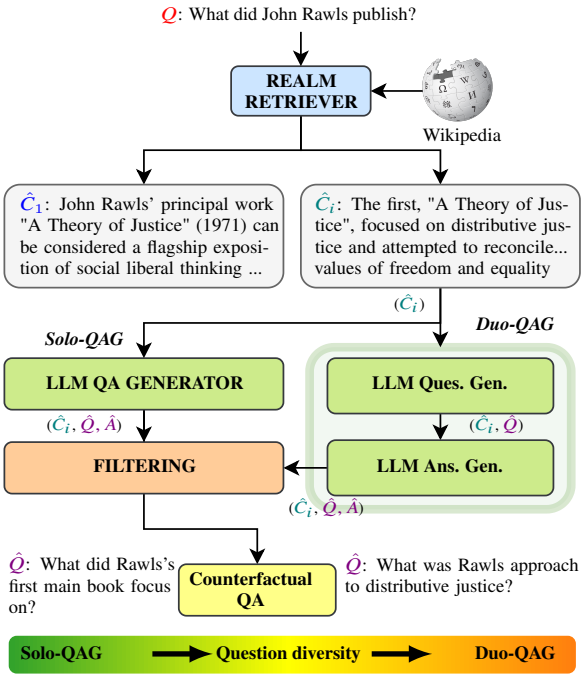

Figure 2: Our proposed methodology for generating counterfactual instances. The Solo-QAG approach (left) generates counterfactual QA pairs in a single pass while the Duo-QAG approach (right) first generates the question, and then the answer.

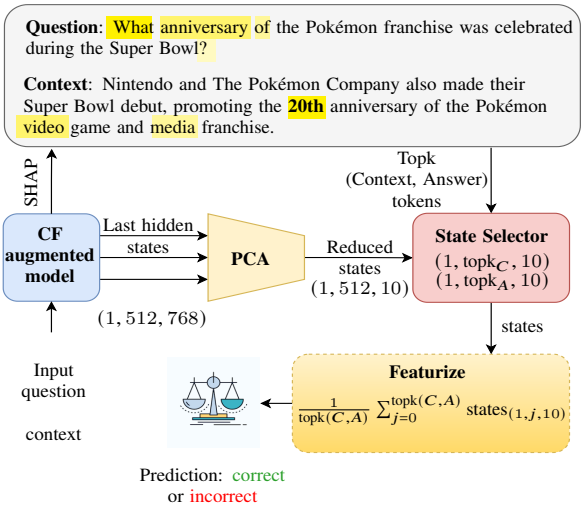

Figure 3: Our proposed calibration methodology. The dense representations of the highly important input tokens from the CF-augmented model are condensed and converted to semantic features to train a classifier that predicts if the model prediction is correct.

explanations are only used for selecting data samples that improve model generalization. More recently, Ye and Durrett (2022) studied how to improve a black box model's calibration in OOD settings by leveraging handcrafted features from explanations (Ribeiro et al., 2016; Lundberg and Lee, 2017). However, their method of computing handcrafted features maps tokens to linguistic features such as POS tags, a process in which the meaning of individual tokens is lost. In our work, we explore the connection between explanation content of CF augmented models and calibration performance. The questions we set out to answer are: (1) does the content of the explanation matter to the calibrator? (2) which explainer is the best at producing calibration features? and (3) which characteristics of explanations are important for calibration?

3 Methodology

3.1 Datasets

We evaluate our CF augmentation methods on seven extractive question answering datasets commonly used in related works: SQuAD (Rajpurkar et al., 2016), SQuAD-Adversarial (Jia and Liang, 2017), TriviaQA (Joshi et al., 2017), HotpotQA (Yang et al., 2018), Natural Questions (NQ) (Kwiatkowski et al., 2019), NewsQA (Trischler et al., 2017), BioASQ (Tsatsaronis et al., 2015). For all datasets except SQuAD, we directly use the pre-processed version of the dataset from the MRQA Shared Task (Fisch et al., 2019). We provide detailed descriptions of the datasets in Appendix A.

3.2 Setup and Base Models

Following the setup of Ye and Durrett (2022), we train a RoBERTa-base model (Liu et al., 2019) on the SQuAD dataset and evaluate its OOD performance on the remaining six datasets. To improve the generalization capabilities of our base model, we augment the SQuAD data with CFs automatically generated using the following LLMs: (1) GPT-JT (6B) and (2) GPT-NeoxT (20B), instruction tuned versions of GPT-J (Wang and Komatsuzaki, 2021) and GPT-Neox (Black et al., 2022); (3) LLaMA (13B) (Touvron et al., 2023), (4) Alpaca (Taori et al., 2023), (5) Flan-T5-xxl (11B) (Wei et al., 2022a), and (6) Flan-UL2 (20B) (Tay et al., 2022). We obtain the Alpaca model by Low-Rank Adaptation (LoRA) (Hu et al., 2022) fine-tuning the LLaMA (13B) model on the Alpaca dataset (Taori et al., 2023) for 10 epochs. These models are selected as they are publicly available, trained on varying data and representative of

both decoder-only and encoder-decoder families, as well as their instruction-tuned variants. We omit detailed model descriptions for brevity and refer the reader to Appendix B for more details.

3.3 Generating Counterfactuals

3.3.1 Retrieve-Generate-Filter

Introduced in Paranjape et al. (2022), retrieve-generate-filter (RGF) describes a framework used to create counterfactual instances with minimal human supervision. The *retrieval* step leverages the REALM retrieval augmented language model (Guu et al., 2020) to produce a ranked list of contexts and answers within those contexts, given a question as input. Based on this set of contexts and answers, RGF then *generates* question candidates using a T5-3B question generation model fine-tuned on the NQ dataset. These question candidates are then *filtered* to ensure quality. For the sake of space, we elaborate the details of all filtering steps in Appendix E. Each generated question, along with its corresponding context and answer, constitutes a *counterfactual* instance.

3.3.2 Solo-QAG and Duo-QAG

In our proposed approaches, we first use the REALM model to *retrieve* candidate contexts and then select the context for which the T5-large model generates the closest question based on the Levenshtein distance (Levenshtein, 1966). Then, given the chosen context $\hat{c}_i$, our *LLM QA generator* generates the counterfactual $(\hat{q}, \hat{a})$ pair using an LLM prompted in 1-shot manner with a prompt containing the original (q, a) pair. As our preliminary experiments have shown that some LLMs are better at jointly generating the question and answer, while others perform better at sequential generation, we propose the following two approaches to generating counterfactual instances:

Solo-QAG. For every chosen context $\hat{c}_i$, we prompt the LLM to produce a question-answer pair $(\hat{q}_i, \hat{a}_i)$ in a single generative step. We name this approach **S**ingle-**P**hase **Q**uestion-**A**nswer **G**eneration.

Duo-QAG. In this approach, we split LLM QA generation into two phases. We first generate the question $\hat{q}_i$ that can be answered based on the given context $\hat{c}_i$, and then use the question-context pair $(\hat{q}_i, \hat{c}_i)$ to generate an answer $\hat{a}_i$. We name this approach **D**ual-Phase **Q**uestion-**A**nswer **G**eneration.

We illustrate our proposed approaches in Section 3. In both generation approaches, we prompt the LLM a maximum of three times with different random seeds until a satisfiable instance is produced (e.g. one which is not empty or excessively short). We detail the prompts used for CF generation in Appendix D.

As the LLM-based CF generation approaches are still prone to generating open-ended questions which cannot be answered based on information provided in the input context, we introduce a filtering step designed to ensure high quality of generated CF instances. The first filtering step leverages *context relevance filtering* to identify CF questions where the corresponding input context does not provide sufficient information for an answer. Since context relevance filtering may also discard some complex, but answerable questions, we further employ the round-trip consistency approach (Alberti et al., 2019; Fang et al., 2020) to retrieve incorrectly discarded samples using an ensemble of three language models initialized with different seeds to answer the LLM generated questions. If answers from 2 or more language models agree with the LLM-generated answer, the CF sample is retained.

Intrinsic Evaluation. We evaluate the generated counterfactuals along two dimensions: *Fluency* and *Correctness*. Fluency measures whether the generated CF question is grammatically correct and semantically meaningful. Correctness measures the alignment between the generated question $\hat{Q}$, context $\hat{C}$, and answer $\hat{A}$, i.e., the question is answerable from the context and the answer is correct.

We perform a human evaluation on 50 CF instances sampled each from the RGF, LLaMA, GPT-NeoxT, and Flan-UL2 models and report our results in Figure 4. We find that over 90% of the generated questions from all the models are fluent, as the generation leverages high-quality pre-trained language models. We further quantify the correctness of the generated CF instances and find out that our methodology with LLaMA and Flan-UL2 models produces minimal (<5%) noisy data as compared to ~20% of RGF, stating that the Solo-QAG and Duo-QAG produce superior and answerable CF instances. We detail the annotation process in Appendix G.

3.3.3 Estimating Diversity of Counterfactuals

We quantitatively evaluate the *diversity* of generated counterfactual questions with respect to

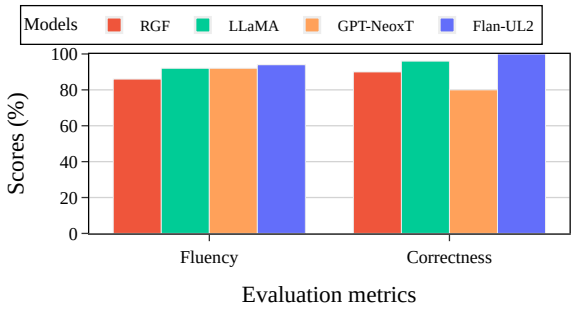

Figure 4: Quantitative evaluation of fluency and correctness of the CF instances generated by the RGF, LLaMA, GPT-NeoxT, and Flan-UL2 models.

the original questions along two axes: (1) *surface form variation*, measured by self-BLEU (Zhu et al., 2018) and Levenshtein edit distance, as proposed in Wu et al. (2021); and (2) *semantic variation*, measured by SBERT (Reimers and Gurevych, 2019) embedding similarity and semantic uncertainty (Kuhn et al., 2023). *Surface form variation* metrics quantify the surface form difference between the original question and its counterfactual counterpart through n-gram and character level overlaps. Lower self-BLEU and conversely, a higher edit distance indicate greater surface form diversity between a question and its corresponding CF. As surface form diversity does not necessarily imply semantic difference, we also estimate *semantic variation* through two methods. Our first method estimates semantic diversity through cosine similarity between the SBERT embeddings of a question and its counterfactual counterpart. We complement SBERT similarity by adapting a novel method measuring semantic uncertainty. Following Kuhn et al. (2023), we leverage a pretrained natural language inference model, in our case DeBERTa-large (He et al., 2021) and compute the bidirectional entailment (equivalence) probability between the original question and its corresponding CF. Herein, a lower equivalence score indicates lower confidence of entailment between the pair, which in turn corresponds to greater semantic variation. In Appendix F, we highlight semantic variations introduced in randomly sampled CFs generated by our approach.

3.4 Model Calibration

As prediction probabilities of SLMs are often poorly calibrated, practitioners frequently resort to model calibration – training simpler models to detect when the underlying model is faulty by producing a score which overrides the model confidence and conveys whether the original prediction is correct. The benefit of a good calibrator model is that faulty, but confident, predictions can be detected before a wrong answer is returned to an end-user. Apart from the base model confidence, these models usually leverage diverse heuristic features as additional inputs to a calibrator model. Following previous work (Kamath et al., 2020; Ye and Durrett, 2022), we use the random forest classifier as our calibration model. We train each calibrator model on 500 training samples, changing only the input feature sets, while the correctness of the base model prediction is used as the output label. To evaluate the quality of model calibration, we leverage the Macro-average Calibration Error, MacroCE (Si et al., 2022), a recently proposed enhanced version of the Expected Calibration Error (ECE) (Guo et al., 2017). We elaborate the model calibration procedure in Appendix I.

3.4.1 Baseline

Ye and Durrett (2022) focus on calibrating black-box models with explainers based on local perturbation techniques: LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017). Due to the large scale of our experiments and the high computational complexity of LIME, we only use their SHAP feature-based calibration technique. Ye and Durrett (2022) map input tokens to linguistic features, such as POS tags, and then aggregate importance scores across all tokens assigned specific feature values, e.g. nouns. These aggregated scores are used as input features of the calibrator model, augmenting it with the explanation information.

3.4.2 Improving Explanations for Calibrators

The calibration approach of Ye and Durrett (2022) has two main limitations: 1) they consider only black-box explainability methods, leaving uncertainty about calibrators' preferences for alternative explainability methods; and (2) input features to calibrator models are aggregated importance values of tokens from specific word categories (i.e. POS tags), a process where the token meanings are lost.

To tackle the first issue and account for the variation in quality of explanations generated by explainability methods (Jain and Wallace, 2019; Neely et al., 2022), we drop the restrictive black-box scenario and extend the scope of our evaluation to attention- and gradient-based white-box explainers, which provide a broader overview of how expla-

nations affect calibration performance. We employ normalized attention scores (α) (Jain et al., 2020) and gradient-scaled attention scores ($\alpha\nabla\alpha$) from the attention-based family, while we consider InputXGradients ($x\nabla x$) (Kindermans et al., 2016) and integrated gradients (IG) (Sundararajan et al., 2017) from gradient-based approaches. To address the second drawback, we augment calibrator models with semantic features computed from dense representations of input tokens assigned high importance by explanation methods. We select the top 10% and 20% most salient tokens from the *context* and *answer*, respectively, reduce their dimensionality to the top ten principal components using PCA (Shlens, 2014)[2], and then average their token representations. The resulting vector is then used as additional input to the calibrator model. We sketch our proposed calibration procedure in Figure 3.

We are also interested in the explanation characteristics indicative of the rationale-augmented calibrators' performance. To this end, we measure the *comprehensiveness* and *sufficiency* (DeYoung et al., 2020) of generated explanations, two metrics used to determine the influence of the rationale on a prediction. Given input tokens $\{x_i\}_{i=1}^{t}$, *comprehensiveness* masks $n\%$ input tokens assigned the highest importance scores. The comprehensiveness score is then determined as the change in the prediction probability of the model for the same answer, where a high difference in the prediction score indicates that the masked rationale tokens were influential for the prediction. To estimate the degree to which extracted rationales are *sufficient* for the models' prediction, given input tokens $\{x_i\}_{i=1}^{t}$, *sufficiency* retains only $n\%$ of tokens assigned the highest importance scores, masking out the rest. The sufficiency score is then determined as the change in prediction probability of the model for the same answer. Following Carton et al. (2020) and Chrysostomou and Aletras (2022), we constrain sufficiency between 0 and 1 and report $1 - suff$ so that higher is better.

In case of extractive QA, we do not mask (for comprehensiveness) and explicitly keep (for sufficiency) the question and answer tokens so that the model is able to answer the input question. We report average sufficiency and comprehensiveness scores when retaining (for sufficiency) or masking (for comprehensiveness) the top $n \in \{2\%, 10\%, 20\%, 50\%\}$ most important tokens.

[2]The number of principal components was determined through a series of non-exhaustive experiments. We experimented with 10 and 100 features, and found that using 10 features yields better results.

4 Experiments

4.1 Generating Counterfactual Instances

We report an overview of models used to generate CFs, their parameter sizes and the resulting number of generated (usable) CFs in Appendix F. The Duo-QAG approach yields a significantly higher number of usable samples (~70k) compared to Solo-QAG (~50k), indicating that the two-step approach produces higher fidelity CF instances. We hypothesize that the better generative abilities of the Duo-QAG approach arise from the extensive pre-training of FLAN-based LLMs on question generation and question-answering tasks.

In Table 1, we report the diversity of the generated CFs with respect to surface form and semantic variation. Our reference approach quantifies the upper bound of the SQuAD dataset diversity by comparing every data sample with another random sample from the dataset. The RGF approach produces the least diverse CFs, which is expected considering its methodology which aims to generate and select CFs which deviate minimally from the input samples. Contrary to RGF, our methodology utilizes capabilities of LLMs to produce CF instances that are semantically and contextually more diverse. We hypothesize that counterfactual instances more diverse from the original improve the models' input space coverage, which should in turn improve OOD performance and calibration. We verify this hypothesis in the following sections.

4.2 Generalization of CF Augmented Models

We report the exact-match scores of the CF-augmented RoBERTa-base model on six OOD datasets in Table 2.

Models augmented with CFs generated by our approach outperform all baselines across all OOD datasets, except NewsQA. We hypothesize that this is due to the complex reasoning required by NewsQA, involving synthesis of information from multiple sentences (Trischler et al., 2017) and that LLMs might not be able to generate diverse yet useful complex CF questions based on instances from the simpler SQuAD dataset. All CF-augmented models maintain a comparable performance on the in-domain SQuAD dataset, implying that training with diverse data improves OOD generaliza-

Approach	Model	Surface form variation		Semantic variation	
		Self-BLEU (↓)	Levenshtein (↑)	SBERT Sim. (↓)	Semantic Equivalence (↓)
Reference	-	0.11	1.00	0.11	0.54
RGF	T5-3B	0.31	0.61	0.56	0.52
SOLO-QAG	GPT-JT	0.26	0.67	0.48	0.46
	LLAMA	0.28	0.65	0.50	0.51
	ALPACA	0.27	0.67	0.50	0.55
	GPT-NEOXT	0.24	0.68	0.45	0.46
DUO-QAG	FLAN T5-XXL	**0.19**	**0.71**	**0.41**	0.41
	FLAN-UL2	**0.19**	**0.71**	**0.41**	**0.40**

Table 1: Quantitative evaluation of the diversity of generated counterfactuals with respect to the original questions. The metrics are complementary – diverse CFs are expected to be further away from original instances in both surface form and meaning. To contextualize semantic and surface form variation of CFs, we contrast them to a **reference** baseline – diversity of an instance compared to a randomly selected other instance from the dataset.

Exact Match	SQuAD	SQuAD$_{Adv.}$	TriviaQA	HotpotQA	NQ	NewsQA	BioASQ	G_{ood}
BASE	$84.98_{0.07}$	$66.60_{0.84}$	$39.09_{1.87}$	$48.16_{0.16}$	$41.94_{1.01}$	$42.21_{0.69}$	$47.93_{1.22}$	-
RGF	**$85.53_{0.04}$**	$65.97_{0.42}$	$44.98_{0.22}$	$52.88_{0.25}$	$46.22_{0.29}$	**$43.01_{0.32}$**	$50.50_{0.23}$	2.94
GPT-JT	$84.74_{0.18}$	$67.19_{0.42}$	$47.40_{0.33}$	$51.21_{0.50}$	$47.08_{0.84}$	$42.12_{0.59}$	$52.59_{1.13}$	3.61
LLAMA	$84.85_{0.31}$	$67.57_{0.42}$	**$48.13_{0.05}$**	$51.68_{0.86}$	$48.80_{0.68}$	$42.35_{0.47}$	$51.68_{1.25}$	4.05
ALPACA	$85.42_{0.23}$	$66.59_{0.98}$	$41.79_{1.35}$	$51.88_{0.55}$	$44.79_{2.22}$	$42.48_{0.55}$	$49.56_{0.50}$	1.86
GPT-NEOXT	$84.80_{0.25}$	$68.07_{1.30}$	$46.96_{0.46}$	$53.14_{0.67}$	$47.80_{1.83}$	$41.99_{0.73}$	**$53.19_{1.06}$**	**4.20**
FLAN-T5-XXL	$85.41_{0.28}$	$67.15_{0.59}$	$42.91_{0.53}$	$53.52_{0.95}$	$48.05_{0.86}$	$42.70_{1.18}$	$49.29_{0.52}$	2.95
FLAN-UL2	$85.38_{0.10}$	**$68.09_{0.82}$**	$45.40_{1.24}$	**$53.70_{0.56}$**	**$48.88_{0.91}$**	$42.99_{0.81}$	$51.33_{0.48}$	4.08

Table 2: EM results for RoBERTa-base model trained on the SQuAD dataset (BASE) and augmented with counterfactual data. We report the mean$_{std.}$ over 3 runs with different random seeds. The last column (G_{ood}) shows the average gain over the BASE model on OOD datasets. Numbers marked in **bold green**, and orange colours represent the highest and second highest scores. We also report the F1 scores, which follow a similar trend, in Appendix H.

tion while preserving in-domain performance. The spread of best-performing models shows that there is no *one-model-fits-all* strategy and that even less diverse CFs may be better suited for some OOD datasets. The size and training data of the CF generator LLMs may also play an influential role as the larger scale LLaMA, GPT-NeoXT, and Flan-UL2 models are also the best performers. However, this aspect should not limit the applicability of our approach since even the smaller GPT-JT model provides significant gains on OOD datasets.

Overall, the GPT-NeoxT CF augmented model has the highest average gain across all OOD datasets, with FLAN-UL2 and LLaMA closely behind. This is largely attributed to its strong performance on the BioASQ dataset, likely due to its pre-training on medical data from the large-scale PubMed Central dataset (Gao et al., 2021). Our findings show that although all CF-augmented models consistently outperform baselines, the best augmentation approach depends on the concrete OOD dataset, suggesting that alignment between domain expertise of LLMs used to generate CFs and the data distribution of OOD datasets is important.

4.3 Model Calibration

We report the model calibration results as % improvement over the base model in Figure 5. We compare our models against two baselines: (1) CONF, where the calibrator model only uses the thresholded probability of the predicted class to assess whether the prediction is trustworthy, and (2) SHAP. On the CONF baseline, when only the probability of the underlying model is used as input to the calibrator, our CF-augmented models improve calibration accuracy across all OOD datasets with an average increase of ~5%, and up to ~11% on the TriviaQA dataset. These results suggest that augmenting a model with counterfactual instances already improves the model's capability to capture nuanced shifts in the data distribution. Improved robustness of CF-augmented models is further evident from the high inverse MacroCE scores on the CONF baseline where even without features from explanations, CF-augmented models exhibit the best calibration scores (~ 6%) across all datasets.

When incorporating explanation features, on the SHAP baseline, the CF-augmented models improve calibration accuracy by an average of ~3% on two

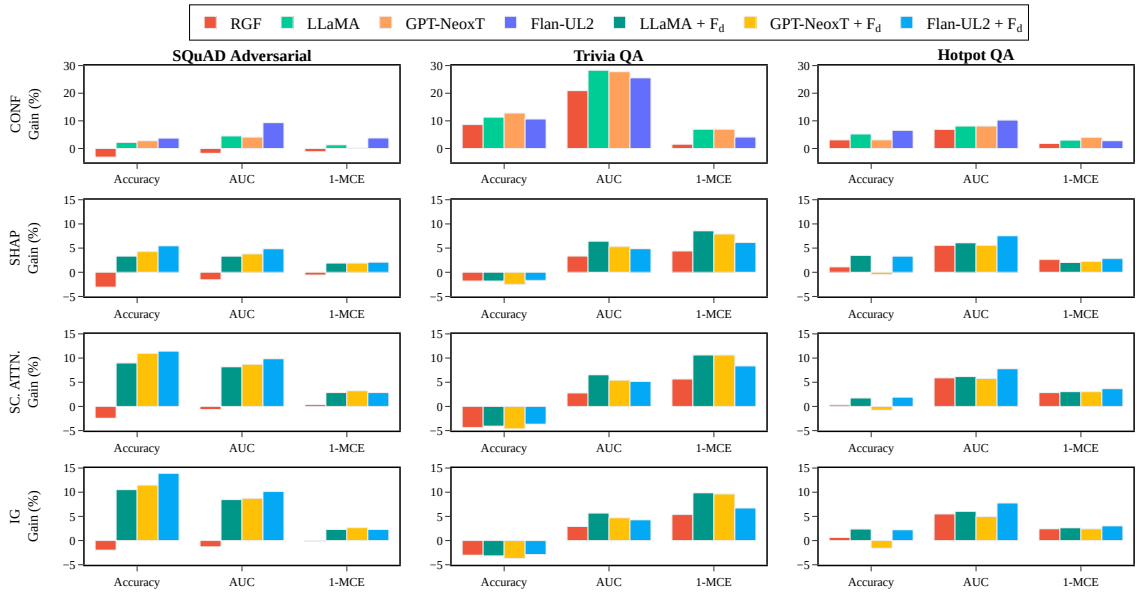

Figure 5: Percentage improvement of CF augmented models' calibration performance over the unaugmented RoBERTa-base model trained on SQuAD, using features based on probability (CONF) and rationales from SHAP, scaled attention and integrated gradients. The results for CONF (row #1) are reported on models which do not use explanation-based features. In the remaining experiments (other rows), along with BASE and RGF, we report the results of dense-feature augmented calibrators. We provide the complete results with other datasets and explanation methods in Appendix J.1.

Model	Comprehensiveness (↑)					Sufficiency (↑)				
	α	$\alpha\nabla\alpha$	x∇x	IG	SHAP	α	$\alpha\nabla\alpha$	x∇x	IG	SHAP
Base	0.33	0.34	0.36	0.38	0.35	0.51	0.51	0.52	0.52	0.51
RGF	**0.35**	**0.41**	**0.41**	**0.43**	**0.44**	0.41	0.43	0.42	0.43	0.41
LLaMA	0.32	0.34	0.33	0.34	0.32	0.54	0.55	0.54	0.54	0.54
GPT-Neox	0.29	0.31	0.31	0.33	0.30	**0.56**	**0.57**	**0.57**	**0.57**	**0.56**
Flan-UL2	0.33	0.36	0.37	0.38	0.36	0.47	0.48	0.48	0.48	0.48

Table 3: Comprehensiveness and sufficiency scores of explanations generated by baseline and counterfactual augmented models, averaged across the six OOD datasets. Numbers marked in **bold green**, and red represent the highest and lowest scores, respectively. We provide comprehensive results for each dataset in Appendix J.2.

out of three OOD datasets, the exception being TriviaQA, where the accuracy decreases marginally. Nevertheless, the CF-augmented models achieve superior AUC scores on all OOD datasets with an average improvement of ~5.5% compared to the SHAP baseline without CF augmentation. For completeness, we report results on the NQ, NewsQA, and BioASQ datasets, along with the results produced by α and $x\nabla x$ in the Appendix J.1.

Overall, the CF-augmented models coupled with dense rationale features improve calibration over all baselines, all explanation methods, and OOD datasets, specifically on the SQuAD adversarial dataset. Our results show that augmenting training data with CF instances improves model calibration and that calibrators benefit from the semantic content of the most salient tokens from explanations.

4.4 Desiderata of Rationales for Calibration

In this section, we explore whether underlying characteristics of explanations are indicative of their usefulness to calibrators. In Table 3, we report two metrics commonly used to estimate faithfulness of explanations – *sufficiency* and *comprehensiveness*. The RGF approach produces the most *comprehensive* explanations across all OOD datasets when compared to CFs generated by LLMs, while in terms of *sufficiency*, all CF-augmented models report higher scores compared to the RGF baseline. As *comprehensiveness* is higher when a larger number of tokens is salient for the prediction, while higher *sufficiency* means that the model relies on a smaller subset of tokens, the results imply higher sufficiency of explanations is indicative of calibrator model performance. This is intuitive as the RGF

approach generates minimally different counterfactuals with a large amount of feature overlap. On the contrary, we believe diverse CFs generated by our approach force the models to capture nuanced differences in explanations between instances.

5 Conclusion

In our paper, we present a novel approach for automatic data augmentation by LLM generated counterfactual instances diverse in surface form and semantic content. Our results show that augmenting training data of smaller models with LLM generated CFs consistently improves generalization capabilities of SLMs across six OOD extractive QA datasets. We further show that models trained on CF augmented data are easier to calibrate, both when considering the standard confidence-based setup as well as the explanation-augmented calibration setup. Finally, we show that rationale-augmented calibrator models prefer concise explanations, rather than comprehensive ones. By highlighting the fact that more diverse CF instances improve the quality of the models' internal representations we pave the way for future works exploring the relation between surface form and semantic diversity of data used for augmentation and the models' generalization performance.

Limitations

Our work only concentrates on the extractive QA task and can be extended to other generative tasks in the future. In addition, our approach of generating CFs can be computationally expensive for very large models and therefore we constrained ourselves to a maximum model size of 20B. In future, smaller and efficient LLMs can even make our methods better applicable. For model calibration, we utilize SHAP explanations as baselines from prior work which are also compute intensive since they need to compute many perturbations on the data. But these compute based limitations should not limit the applicability of our methods since we also show that efficient explanations based on attentions and gradients can also perform at par or sometimes even better than SHAP.

Ethics and Broader Impact Statement

The core of our work is based on the ability of LLMs to generate reasonable explanations but prior works have shown that these models hallucinate and are not free from biases captured from large-scale web data. These hallucinations and biases might trickle down to SLM as we augment them with LLM generated CF data. To overcome these issues, we design our approaches with hard and soft filtering stages that try to eliminate such noisy and biased data and still achieve significant improvements over existing baselines.

Acknowledgements

This work has been funded by the German Research Foundation (DFG) as part of the UKP-SQuARE project (grant GU 798/29-1), by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE, and by the European Union (ERC, InterText, 101054961). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. Gpt-neox-20b: An open-source autoregressive language model. *CoRR*, abs/2204.06745.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens

Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. Evaluating and characterizing human rationales. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9294–9307, Online. Association for Computational Linguistics.

George Chrysostomou and Nikolaos Aletras. 2022. An empirical study on explanations in out-of-domain settings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6920–6938, Dublin, Ireland. Association for Computational Linguistics.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4443–4458. Association for Computational Linguistics.

Yuwei Fang, Shuohang Wang, Zhe Gan, Siqi Sun, and Jingjing Liu. 2020. Accelerating real-time question answering via question generation. *CoRR*, abs/2009.05167.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering, MRQA@EMNLP 2019, Hong Kong, China, November 4, 2019*, pages 1–13. Association for Computational Linguistics.

Zee Fryer, Vera Axelrod, Ben Packer, Alex Beutel, Jilin Chen, and Kellie Webster. 2022. Flexible text generation for counterfactual fairness probing. *CoRR*, abs/2206.13757.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2021. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027.

Yair Ori Gat, Nitay Calderon, Amir Feder, Alexander Chapanin, Amit Sharma, and Roi Reichart. 2023. Faithful explanations of black-box NLP models using llm-generated counterfactuals. *CoRR*, abs/2310.00603.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330. JMLR.org.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2021. An overview of uncertainty calibration for text classification and the role of distillation. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 289–306, Online. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Guande He, Jianfei Chen, and Jun Zhu. 2023. Preserving pre-trained features helps calibrate fine-tuned language models. In *The Eleventh International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2021–2031. Association for Computational Linguistics.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5684–5696. Association for Computational Linguistics.

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

Daniel Khashabi, Tushar Khot, and Ashish Sabharwal. 2020. More bang for your buck: Natural perturbation for robust question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 163–170, Online. Association for Computational Linguistics.

Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. 2016. Investigating the influence of noise and distractors on the interpretation of neural networks. *CoRR*, abs/1611.07270.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran S. Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. 2020. Calibrated language model fine-tuning for in- and out-of-distribution data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1326–1340, Online. Association for Computational Linguistics.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

Dongfang Li, Baotian Hu, and Qingcai Chen. 2022a. Calibration meets explanation: A simple and effective approach for model confidence estimates. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2775–2784. Association for Computational Linguistics.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2022b. Explanations from large language models make small reasoners better. *CoRR*, abs/2210.06726.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguistics*, 19(2):313–330.

Michael Neely, Stefan F. Schouten, Maurits J. R. Bleeker, and Ana Lucic. 2022. A song of (dis)agreement: Evaluating the evaluation of explainable artificial intelligence in natural language processing. In *HHAI 2022: Augmenting Human Intellect - Proceedings of the First International Conference on Hybrid Human-Artificial Intelligence, Amsterdam, The Netherlands, 13-17 June 2022*, volume 354 of *Frontiers in Artificial Intelligence and Applications*, pages 60–78. IOS Press.

Bhargavi Paranjape, Matthew Lamm, and Ian Tenney. 2022. Retrieval-guided counterfactual generation for QA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1670–1686, Dublin, Ireland. Association for Computational Linguistics.

Seo Yeon Park and Cornelia Caragea. 2022. On the calibration of pre-trained language models using mixup guided by area under the margin and saliency. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5364–5374, Dublin, Ireland. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

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

Jonathon Shlens. 2014. A tutorial on principal component analysis. *CoRR*, abs/1404.1100.

Chenglei Si, Chen Zhao, Sewon Min, and Jordan Boyd-Graber. 2022. Re-examining calibration: The case of question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2814–2829, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, et al. 2022. Ul2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinform.*, 16:138:1–138:28.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5085–5109. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models. *CoRR*, abs/2206.07682.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022c. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

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.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.

Xi Ye and Greg Durrett. 2022. Can explanations be useful for calibrating black box models? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6199–6212, Dublin, Ireland. Association for Computational Linguistics.

Xi Ye, Rohan Nair, and Greg Durrett. 2021. Connecting attributions and QA model behavior on realistic counterfactuals. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5496–5512, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Knowing more about questions can help: Improving calibration in question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1958–1970, Online. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

A Datasets

We perform our experiments on English language datasets. SQuAD is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles. The SQuAD-adversarial dataset is an adversarial setting of SQuAD wherein automatically generated distractor sentences are inserted at the end of each example context to distract computer systems without changing the correct answer or misleading humans. TriviaQA comprises of QA pairs sourced from trivia and quiz-league websites. Similar to SQuAD, HotpotQA is also comprised of passages extracted from wikipedia but consists of questions requiring multiple reasoning steps. Natural Questions consists of questions collected from information-seeking queries to the Google search engine by real users under natural conditions. Answers to the questions are annotated in a retrieved Wikipedia page by crowdworkers. NewsQA is a challenging machine comprehension dataset of human-generated question-answer pairs based on a set of over 10,000 news articles from CNN. BioASQ is a large-scale biomedical semantic indexing and question answering dataset collected by domain experts. For evaluation, we use the pre-processed test sets from the MRQA shared task (Fisch et al., 2019).

B Models

We use the instruction tuned GPT models, namely GPT-JT and GPT-NeoxT. GPT-JT is a fork of EleutherAI's GPT-J (6B) model trained on diverse data such as chain-of-thought (Wei et al., 2022c), Public Pool of Prompts (P3) dataset (Sanh et al., 2022), and Natural-Instructions (NI) (Wang et al., 2022) with the UL2 training objective (Tay et al., 2022). GPT-NeoxT is based on ElutherAI's GPT-NeoX (20B) model fine-tuned on a set of 43M high quality dialog-style interactions spanning tasks such as QA, classification, extraction, and summarization. LLaMA models are a collection of foundation language models ranging from 7B to 65B parameters trained using publicly available datasets exclusively. In our work, we specifically utilize the LLaMA (13B) model, which aligns with the size of other language models we employ. We additionally use the Alpaca model, obtained through Low-Rank Adaptation (LoRA) (Hu et al., 2022) fine-tuning of the LLaMA (13B) model on the Alpaca dataset for 10 epochs. Lastly, we also experiment with the Fine-tuned LAnguage Net (FLAN) models. FLAN fine-tunes the model on a large set of varied instructions that use a simple and intuitive description of the task such as "Classify this movie review as positive or negative," or "Translate this sentence to Danish." Specifically, we use the FLAN versions of T5-11B (Raffel et al., 2020) and 20B UL2 (Unified Language Learner) (Tay et al., 2022) models.

C Training, Infrastructure and Runtime

We use a server with 8 NVIDIA A100 Tensor Core GPUs, each with 80GB VRAM to run all our experiments. Each individual experiment required at most one A100 GPU. LoRA fine-tuning of the Alpaca model took 10 hours using the refined Alpaca dataset. Generating counterfactual instances with LLMs, in total, took 24-72 hours per model and dataset. Training each base RoBERTa model augmented with CF instances took 4 hours, on average per dataset, while inference on OOD datasets required a few minutes per dataset. Training the calibrator random forest model took a maximum of five minutes across all models, datasets and input feature sets. Computing importance features of explanations for all methods except SHAP took 1-2 hours per experiment, while due to the computational complexity of SHAP, each experiment required 3-4 days.

We used the following hyperparameters to train the RoBERTa model used throughout our experiments: (1) learning rate: 1e-5; (2) batch size: 64; (3) epochs: 5; (4) warmup ratio: 0.06, (5) max input source length: 384. When generating counterfactual instances using LLMs, the hyperparameters used during inference were: (1) max new tokens: 50, (2) temperature: 0.7 for all considered LLMs.

D Prompt Details

In Table 4, we list the prompts used to generate CF data from the LLMs. Prompts for the *GPT* and *LLaMA* family are almost similar apart from minor tweaks according to the model type, e.g. *GPT-NeoxT* is a chat-based model so the instruction has to follow a chat style and *Alpaca* needs a specific instruction format based on its training. The *Flan* models follow a two-stage approach of generation: the question generation prompt asks for a question that can be answered from the context in a short span of 10 words (following SQuAD which has small answer spans) and the answer generation prompt asks for the answer to the generated question from the input context. If the question is not

answerable, we ask the model to give *I don't know* as output. Doing this maintains the fidelity of the answer generation capability of the model.

In Table 5, we list the prompt used for the context relevance filtering stage using the Flan-UL2 model. Given a generated QA pair from a context, we prompt the model to find if the context aligns with the answer and vice-versa. We ask the model to output a hard decision in terms of True or False.

E Filtering for Data Augmentation

Context Relevance Filtering. Despite their impressive generative capabilities, LLMs are still prone to generating open-ended questions that cannot be answered from the information provided in the input context alone. To account for such cases, in context relevance filtering, we use the FLAN-UL2 (Tay et al., 2022) model to filter samples where the context $\hat{c}$ does not provide sufficient information for a model to correctly answer the generated question $\hat{q}$ by discarding questions which the model labeled as unanswerable (see Appendix D). This approach may also discard some complex, but answerable questions, such as those based on chronological types.[3] An example of one such case that consists of multiple event dates which tends to confuse the context filtering model is given below:
***Question**: What year was the anniversary of the Cunard liner company?*
***Context**: In 2011, all three Cunard ships in service changed vessel registry to Hamilton, Bermuda, the first time in the 171-year history. On 25th May 2015, the three Cunard ocean liners sailed up the Mersey into Liverpool to commemorate the 175th anniversary of Cunard. The ships performed manoeuvres in the celebrations of the centenary of the Cunard Building on 2nd June 2016.*

Noise Filtering. To retain such answerable questions incorrectly discarded by context relevance filtering, we use the round-trip consistency approach (Alberti et al., 2019; Fang et al., 2020) which leverages existing QA models to answer the LLM generated questions, ensuring that the predicted answer aligns with the LLM generated target answer. During noise filtering, we employ an ensemble of three LLMs (the same ones used to generate CFs), initialized using different random seeds during inference to verify the generated answers. The generated CFs agreed upon by at least 2 models are kept, retaining 90%-95% data discarded by context relevance filtering in the DuoQAG approach.

[3]We sample 100 instances generated by LLMs and notice that such cases occur only in the Duo-QAG approach. We hypothesize that this is due to the generation methodology of the Duo-QAG approach, which, due to its looser coupling during CF generation, produces more diverse and complex questions than the Solo-QAG approach.

F Generated Counterfactuals

We report an overview of models used to generate CFs, their parameter sizes, and the resulting number of generated (usable) CFs in Table 6. The Duo-QAG approach yields a significantly higher number of usable samples (~70k) compared to Solo-QAG (~50k), indicating that the two-step approach produces higher fidelity CF instances. We hypothesize that the better generative abilities of the Duo-QAG approach arise from the extensive pre-training of FLAN-based LLMs on question generation and question-answering tasks.

In Table 7, we highlight semantic variations for randomly sampled counterfactuals generated by our approach. Even in the random sample, we can observe a variety of semantic changes such as metonymy, topic shift, clarification, reversal, and expansion.

G Human Evaluation of Generated Counterfactuals

To measure the fluency of generated CF questions, we score the question on a scale ranging from 1-5, see Table 8. A question with significant grammatical errors is assigned a low score whereas a well-written and comprehensible question is assigned a high score. We perform a small-scale human evaluation with one graduate candidate proficient in English and consider a question as fluent if it gets a score of three or above.

For measuring correctness, given a question, context, and answer pair, we set two criteria: (1) the question should be answerable from the given context, and (2) the answer should be a correct and a direct span from the context. If these criteria are met, we consider the CF instance to be correct. Similar to the fluency setup, we hire a graduate candidate proficient in English to perform this evaluation.

H Generalization of CF Augmented Models

In Table 9, we report the F1 results of the RoBERTa-base model trained on the SQuAD

Model	Prompt
GPT-JT & LLaMA	As a question generator, your task is to create a concise and clear question that can be answered by an answer span within a given context. The context should be a piece of text, such as a news article or historical document, and the question should require understanding and analysis of the information presented in the context. Your generated question should focus on key details or events described in the context and should demonstrate your ability to identify important information. Additionally, please ensure that your question is specific enough to have a single correct answer within the given context. Please note that you may need to read through the provided context multiple times to fully understand its contents before generating an appropriate question. Context: original context Question: original question Answer: original answer Context: context Question:
Alpaca	Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request. ### Instruction: As a question generator, your task is to create a concise and clear question that can be answered by an answer span within the given context. The context should be a piece of text, such as a news article or historical document, and the question should require understanding and analysis of the information presented in the context. Your generated question should focus on key details or events described in the context and should demonstrate your ability to identify important information. Additionally, please ensure that your question is specific enough to have a single correct answer within the given context. Please note that you may need to read through the provided context multiple times to fully understand its contents before generating an appropriate question. Context: original context Question: original question Answer: original answer Context: context Question:
GPT-NeoxT	Who wanted to take over North Korea? As a question generator, your task is to create a clear and concise question that can be answered by an answer span within a given context. The context should be a piece of text, such as a news article or historical document, and the question should require understanding and analysis of the information presented in the context. Your generated question should focus on key details or events described in the context, requiring readers to carefully read and analyze the provided text. Please ensure that your question is specific enough to have only one correct answer span within the given context. Please note that you should aim for clarity and concision while still maintaining accuracy and relevance to the provided text. Context: original context Question: original question Answer: original answer Context: context Question:
Flan-T5 & UL2	**Question generation:** Generate a fluent and answerable question from the given context. Ensure that the answer is a span in the context and is less than 10 words. Context: original context Question: original question Answer: original answer Context: context Question: **Answer generation:** Answer the question based on the context below. If the question cannot be answered using the information provided, then answer with "I don't know". Context: original context Question: original question Answer: original answer Context: context Question: generated question Answer:

Table 4: Prompts used to generate diverse counterfactual data.

Prompt
Given the question: {generated question} Decide if the following retrieved context is relevant to the {generated answer}: {retrieved context} Answer in the following format: Context is relevant: True or False.

Table 5: Prompt used for the context relevance filtering stage.

Approach	Model	Params	# CFs
RGF	T5-3B	3B	87k
Solo-QAG	GPT-JT	6B	73k (45k)
	LLaMA	13B	67k (47k)
	Alpaca	13B	63k (50k)
	GPT-NeoxT	20B	76k (44k)
Duo-QAG	Flan T5-xxl	11B	72k (66k)
	Flan-UL2	20B	74k (71k)

Table 6: General information about the counterfactual generation approaches, models used, their size and the total number of generated (selected) counterfactuals.

dataset (BASE) and augmented with CF data on six OOD datasets. The results are comparable to the exact match scores in Table 2, and models augmented with CFs outperform all baselines across all datasets. We find that LLaMA and GPT-NeoxT, based on the Solo-QAG approach, perform best on TriviaQA and BioASQ datasets, while the Flan-UL2 model, based on the Duo-QAG approach, performs best on SQuAD-adversarial, HotpotQA, NQ, and NewsQA datasets. These results strongly advocate the importance of diverse CFs in learning robust features for enhanced OOD performance.

Overall, the FLAN-UL2 CF augmented model also has the highest average gain across all OOD datasets, with GPT-NeoxT and LLaMA closely behind. This is largely attributed to its strong performance across most of our evaluated OOD datasets.

I Details of Model Calibration

I.1 Heuristic Properties for Calibration

Properties In the following paragraphs, we define the heuristic properties in V that are used for the calibration of QA models.

Input Segments The input in the QA task can be decomposed into two segments: *question* and *context*. Each individual token is assigned a corresponding segment name to yield features: Attributions to Question and Attributions to Context.

POS Tags We leverage POS tags from the English Penn Treebank (Marcus et al., 1993) to identify token tags. Ye and Durrett (2022) have already shown that some tags are more important in making the predictions eg. *proper nouns* in questions. Consequently, if a model fails to take into account the proper nouns in a QA pair, it may give incorrect predictions.

Conjunction of Groups We can combine the fine-grained features produced by input segments and POS tags to create high-level features that take into account the attributions of specific tags in the question or context of the QA pair. An example of this feature can be Attributions of NNP in Question.

I.2 Including Dense Features from Explanations

As stated in Section 3.4.2, when considering the most important tokens based on importance scores, we use a higher percentage of answer tokens. This is due to our initial experiments, where such a choice had a high correlation with calibration performance. Consequently, we exclude explanation-based features from the *question* as we observed diminishing calibrator performance on their inclusion. We hypothesize that this decrease is due to the noise or irrelevant information introduced by the *question* features.

I.3 Hyperparameters

We use the RandomForest implementation from Scikit-Learn (Pedregosa et al., 2011). Following the work of Ye and Durrett (2022), we choose a value of 300 for the *n_estimators* parameter and a value of 20 for the *max_depth* parameter for our experiments with different explanation methods. For further experiments with the addition of dense rationale features, we determine the hyperparameters through grid search using 500 training samples and 100 validation samples. The choices of *n_estimators* are [300, 400, 500] and *max_depth* is set to 20. Based on the results, we use *n_estimators* as 500 and *max_depth* as 20 for training the rationale-augmented classifier.

Model	Example (Question, Counterfactual)	Semantic change
RGF	Q: To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France? C: Who did the Virgin Mary appear to?	Paraphrase
	Q: What was Beyoncé's first fragrance called? C: What is the name of Rihanna's first fragrance for women?	Referential shift
	Q: How many UNESCO World Heritage Sites is Portugal home to? C: How many UNESCO World Heritage Sites does Italy have?	Referential shift
LLaMA	Q: What year did the legal importation of slaves end? C: What year was the Act Prohibiting Importation of Slaves enacted?	Metonymy
	Q: How many aircraft did Britain produce in 1940? C: How much food was produced in Britain during the war?	Topic shift
	Q: What was the name of the Luftwaffe plan to invade Britain? C: What was the codename of the German invasion of Britain?	Lexical shift
GPT-NeoxT	Q: What theater sits on Yale's campus? C: Who founded the Yale Repertory Theatre?	Referential shift
	Q: Which islands were a part of the Spanish East Indies? C: Who controlled the Spanish Empire?	Subject shift
	Q: What is the time period called from which no writing can be found? C: What is the time period called where early writing is not understood?	Clarification
Flan-UL2	Q: Which team did Barcelona beat to win the UEFA Super Cup? C: Which team won the 2015 UEFA Super Cup?	Reversal
	Q: What is Lionel Messi's goal total in all competitions? C: How many goals did Lionel Messi score in his career?	Expansion
	Q: What is regarded as the greatest literary work in Old English? C: What is considered the heart of Old English literature?	Reframing

Table 7: Randomly sampled counterfactual questions listed by the model used and the corresponding semantic change introduced. The counterfactual questions from RGF are closer to the original SQuAD samples as compared to the counterfactual questions generated by our approach using generative LLMs.

Category	Score	Description
Very poor	1	Incomprehensible with significant grammatical errors
Poor	2	Comprehensible but with several grammatical errors
Fair	3	Coherent and clear with minor grammatical errors
Good	4	Coherent, clear and grammatically correct
Excellent	5	Grammatically and semantically perfect

Table 8: Fine-grained scale for measuring the fluency of the generated counterfactual question. A low score indicates an incomprehensible text whereas a high score indicates a clear and coherent text without any grammatical errors.

I.4 Choice of calibration metric

Calibration metrics help assess the alignment between a model's predicted probabilities and it's observed predictions. When evaluating model calibration, the **Expected Calibration Error (ECE)** (Guo et al., 2017) has been widely used in prior works (Park and Caragea, 2022; Li et al., 2022a). The ECE can be computed by partitioning model predictions into K equal sized bins according to their model confidence scores. Mathematically, the calibration error can be written as

$$\hat{\mathcal{E}}_k = \frac{1}{|B_k|}\Big|\sum_{i\in B_k}[\mathbb{1}(\hat{y}_i = y_i) - \text{Conf}(x_i, \hat{y}_i)]\Big|,$$

where x is the input, y the ground truth, $\hat{y}$ the prediction, and $Conf(x, \hat{y})$ is the model confidence for i-th example, and B_k denotes the bin with prediction confidences bounded between l_k and u_k. The ECE can now be computed as the weighted average of all the bins such as

$$\text{ECE} = \sum_{k=1}^{K}\frac{|B_k|}{n}\hat{\mathcal{E}}_k,$$

where n is the number of model predictions. The goal is to minimize the ECE without diminishing accuracy. Though widely used, ECE has certain shortcomings. First, most instances are assigned similar confidence, which does not give a proper indication of correct or wrong predictions whereas an ideal calibration metric should be able to do so. Second, bucketing causes cancellation effects, ignoring instance-level calibration error as many predictions are clustered in the same buckets. As a result, there are many over-confident and under-confident predictions in the same bucket and are averaged to become closer to the average accuracy. Due to these issues, we use an enhancement of ECE called **Macro-average Calibration Error, MacroCE** (Si et al., 2022) which considers

F1	SQuAD	SQuAD$_{Adv.}$	TriviaQA	HotpotQA	NQ	NewsQA	BioASQ	G_{ood}
Base	$91.46_{0.05}$	$72.45_{0.95}$	$47.71_{2.11}$	$63.79_{0.41}$	$53.78_{2.04}$	$57.85_{0.89}$	$60.33_{0.99}$	-
RGF	$91.74_{0.09}$	$71.85_{0.30}$	$54.41_{0.15}$	$68.04_{0.23}$	$58.02_{0.27}$	$58.77_{0.17}$	$61.74_{0.18}$	2.82
GPT-JT	$91.30_{0.13}$	$72.83_{0.39}$	$56.67_{0.40}$	$66.60_{0.37}$	$59.79_{1.05}$	$58.14_{0.64}$	$63.45_{0.54}$	3.60
LLaMA	$91.40_{0.03}$	$73.72_{0.75}$	**$57.63_{0.15}$**	$66.65_{0.80}$	$61.65_{0.68}$	$58.28_{0.48}$	$62.65_{1.08}$	4.11
Alpaca	$91.73_{0.06}$	$72.69_{0.90}$	$51.22_{1.67}$	$67.49_{0.52}$	$57.00_{2.61}$	$58.28_{1.08}$	$61.27_{1.04}$	2.01
GPT-NeoXT	$91.31_{0.16}$	$74.09_{1.51}$	$56.42_{0.68}$	$68.28_{0.82}$	$60.41_{1.81}$	$58.10_{0.61}$	**$64.07_{0.70}$**	4.24
Flan-T5-xxl	**$91.85_{0.09}$**	$73.45_{1.04}$	$53.55_{0.91}$	**$69.29_{0.79}$**	$61.03_{1.01}$	$59.07_{1.75}$	$62.08_{0.16}$	3.76
Flan-UL2	$91.67_{0.15}$	**$74.25_{1.05}$**	$55.17_{1.17}$	$69.22_{0.47}$	**$61.88_{1.05}$**	**$59.17_{1.25}$**	$62.73_{0.18}$	**4.42**

Table 9: F1 results for RoBERTa-base model trained on the SQuAD dataset (Base) and augmented with counterfactual data. All the results are averaged over 3 runs with different random seeds. The last column (G_{ood}) shows the average gain of models over the Base model on out-of-domain datasets. Numbers marked in green, **bold** and orange colours represent the highest and second highest scores for the particular dataset and model, respectively.

instance-level errors, but it takes equal consideration of correct and wrong predictions made by the model. Specifically, it calculates the macro-average over calibration errors for positive and negative predictions:

$$\text{ICE}_{pos} = \frac{1}{n_p}\sum_{i=1}^{n_p}(1 - \text{Conf}(x_i, \tilde{y_i})), \forall \tilde{y_i} = y_i,$$

$$\text{ICE}_{neg} = \frac{1}{n_n}\sum_{i=1}^{n_n}(\text{Conf}(x_i, \tilde{y_i}) - 0), \forall \tilde{y_i} \neq y_i,$$

$$\text{MacroCE} = \frac{1}{2} \cdot (ICE_{pos} \quad ICE_{neg}),$$

where n_p and n_n are the correct and wrong predictions.

J Additional Results

J.1 Model Calibration

We report the model calibration results as % improvement over the base model in Figure 6 for the NQ, NewsQA, and BioASQ datasets. We compare our models against two baselines: (1) CONF, where the calibrator model only uses the thresholded probability of the predicted class to assess whether the prediction is trustworthy, and (2) SHAP. On the CONF baseline, when only the probability of the underlying model is used as input to the calibrator, our CF-augmented models improve calibration accuracy across all OOD datasets with an average increase of ~5%. These results suggest that augmenting a model with counterfactual instances already improves the model's capability to capture nuanced shifts in the data distribution. Improved robustness of CF-augmented models is further evident from the high inverse MacroCE scores on the CONF baseline where even without features from explanations, CF-augmented models exhibit the best calibration scores (~ 6%) across all datasets.

When incorporating explanation features, on the SHAP baseline, the CF-augmented models improve calibration accuracy by an average of ~3% on two out of three OOD datasets, the exception being NQ, where the accuracy decreases marginally. Nevertheless, the CF-augmented models achieve superior AUC scores on all OOD datasets with an average improvement of ~5.5% compared to the SHAP baseline without CF augmentation. For completeness, we report results produced by α and $x\nabla x$ in the Appendix J.1.

Overall, the CF-augmented models coupled with dense rationale features improve calibration over all baselines, all explanation methods, and OOD datasets. Our results show that augmenting training data with CF instances improves model calibration and that calibrators benefit from the semantic content of the most salient tokens from explanations.

In Table 10 and Table 11, we present the results of the calibration of LLM augmented models using features based on probability (CONF), and heuristics from three explanation methods: SHAP, scaled attention, and integrated gradients. Additionally, we include explanations based on *attention* and *inputXgradients*. For these additional explanation methods, we observe that the Flan-UL2 performs best on SQuAD adversarial, HotpotQA, and BioASQ datasets. On NQ and NewsQA datasets, our augmented models do not outperform the base model on accuracy but show significant improvements on AUC with a gain of ~+4 and ~+2 points, respectively using the Flan-UL2 CFs. Similarly, for TriviaQA, the LLaMA CFs improve the AUC by ~+5 points over the base model. These results indicate that our calibration methodology helps in calibrating models across a wide range of evaluated

	Approach	SQuAD Adv.			Trivia QA			Hotpot QA		
		ACC(↑)	AUC(↑)	MCE(↓)	ACC(↑)	AUC(↑)	MCE(↓)	ACC(↑)	AUC(↑)	MCE(↓)
CONF	Base	64.2	68.5	0.474	60.1	58.3	0.539	61.1	74.0	0.502
	RGF	62.2	67.3	0.480	65.3	70.5	0.532	63.0	79.1	0.493
	LLaMA	65.6	71.6	0.467	66.9	**74.8**	**0.507**	64.3	80.0	0.487
	GPT-NeoxT	66.0	71.3	0.473	**67.8**	74.5	**0.507**	63.0	80.0	**0.482**
	FLAN-UL2	**66.6**	**74.9**	**0.454**	66.5	73.2	0.520	**65.1**	**81.6**	0.488
SHAP	Base	75.0	84.3	0.471	**72.0**	71.8	0.545	63.3	75.5	0.504
	RGF	72.7	83.0	0.474	70.7	74.2	0.525	64.0	79.7	0.491
	LLaMA	73.4	84.0	0.468	70.7	76.1	**0.505**	65.3	80.0	0.493
	LLaMA + F_r	77.5	87.1	0.461	70.7	**76.4**	0.506	**65.5**	80.1	0.494
	GPT-NeoxT	74.5	84.7	0.470	70.3	75.5	0.508	62.7	79.6	0.493
	GPT-NeoxT + F_r	78.2	87.5	0.461	70.2	75.6	0.509	63.0	79.7	0.493
	FLAN-UL2	75.2	85.4	0.468	70.7	75.3	0.515	65.3	81.1	**0.489**
	FLAN-UL2 + F_r	**79.1**	**88.4**	**0.460**	70.8	75.3	0.517	65.4	**81.2**	0.490
SC. ATTN.	Base	65.7	78.1	0.476	**73.5**	71.9	0.558	63.1	74.5	0.509
	RGF	64.1	77.6	0.474	70.3	73.9	0.533	63.3	78.9	0.495
	LLaMA	65.4	78.1	0.475	70.5	76.3	**0.510**	64.0	78.9	0.493
	LLaMA + F_r	71.6	84.5	0.461	70.5	**76.6**	0.511	64.2	79.1	0.494
	GPT-NeoxT	67.0	79.0	0.475	69.9	75.6	0.511	62.5	78.7	0.492
	GPT-NeoxT + F_r	72.9	84.9	**0.459**	70.1	75.8	0.511	62.6	78.8	0.494
	FLAN-UL2	67.2	79.1	0.479	70.7	75.2	0.520	64.2	80.0	**0.491**
	FLAN-UL2 + F_r	**73.2**	**85.8**	0.461	70.8	75.6	0.521	**64.3**	**80.3**	**0.491**
IG	Base	65.5	77.9	0.476	**72.9**	72.2	0.554	62.9	74.6	0.507
	RGF	64.2	76.9	0.477	70.7	74.3	0.530	63.3	78.7	0.495
	LLaMA	66.5	79.1	0.474	70.5	76.0	**0.508**	64.1	78.9	0.493
	LLaMA + F_r	72.4	84.5	0.464	70.6	**76.3**	0.510	**64.4**	79.1	0.494
	GPT-NeoxT	66.6	79.2	0.474	70.2	75.2	0.511	62.0	78.5	0.493
	GPT-NeoxT + F_r	73.0	84.7	**0.462**	70.2	75.6	0.511	61.9	78.3	0.495
	FLAN-UL2	68.7	81.0	0.475	70.6	74.9	0.521	64.2	79.9	**0.492**
	FLAN-UL2 + F_r	**74.6**	**85.8**	0.464	70.8	75.3	0.524	64.3	**80.4**	**0.492**
ATTN.	Base	66.0	77.8	0.477	**71.7**	70.9	0.557	62.4	73.9	0.508
	RGF	63.7	76.0	0.479	70.5	73.8	0.532	63.2	78.7	0.494
	LLaMA	64.7	77.4	0.477	70.4	76.1	**0.508**	64.0	79.1	0.493
	LLaMA + F_r	71.4	84.0	**0.466**	70.4	**76.3**	0.510	**64.3**	79.4	0.493
	GPT-NeoxT	66.2	78.3	0.477	70.0	74.8	0.511	61.9	78.4	0.493
	GPT-NeoxT + F_r	72.2	84.4	**0.466**	69.8	74.9	0.511	62.0	78.4	0.494
	FLAN-UL2	66.3	78.6	0.478	70.4	74.9	0.519	64.0	79.9	**0.492**
	FLAN-UL2 + F_r	**72.5**	**84.9**	0.467	70.5	75.2	0.521	64.1	**80.2**	**0.492**
INPXGRAD	Base	65.1	78.1	0.474	**72.3**	71.4	0.556	63.0	74.2	0.508
	RGF	63.2	76.3	0.478	70.5	74.0	0.531	63.2	78.7	0.495
	LLaMA	65.7	78.5	0.475	70.5	76.2	**0.509**	64.2	78.8	0.494
	LLAMA + F_r	72.1	84.6	0.464	70.6	**76.3**	0.510	64.2	79.0	0.494
	GPT-NeoxT	66.3	78.9	0.475	70.1	75.2	0.511	62.1	78.5	0.493
	GPT-NeoxT + F_r	72.7	84.9	**0.461**	70.3	75.5	0.511	62.3	78.5	0.495
	FLAN-UL2	67.0	79.5	0.477	70.7	74.9	0.520	64.2	80.1	**0.492**
	FLAN-UL2 + F_r	**73.0**	**85.7**	0.462	70.7	75.1	0.522	**64.3**	**80.3**	0.493

Table 10: Calibration results for a Roberta-base model trained on SQuAD when transferring to out-of-domain settings using explanations based on attention and gradients.

	Approach	NQ			News QA			BioASQ		
		ACC(↑)	AUC(↑)	MCE(↓)	ACC(↑)	AUC(↑)	MCE(↓)	ACC(↑)	AUC(↑)	MCE(↓)
CONF	Base	68.6	70.1	0.531	68.0	71.3	0.535	63.3	72.1	0.499
	RGF	69.8	75.4	0.510	68.5	75.8	0.538	66.3	76.3	0.498
	LLaMA	**72.7**	80.5	0.490	69.9	75.1	0.530	66.8	79.5	0.477
	GPT-NeoxT	72.0	**81.3**	**0.476**	**70.3**	75.1	0.534	65.1	**80.2**	**0.465**
	FLAN-UL2	70.6	80.0	0.494	68.7	**76.7**	**0.527**	**67.5**	77.0	0.494
SHAP	Base	**76.9**	77.9	0.520	70.9	75.2	0.517	69.9	76.8	0.504
	RGF	75.3	79.8	0.507	69.5	75.2	0.512	69.9	78.8	0.495
	LLaMA	73.8	80.3	0.484	70.0	75.3	0.517	72.1	82.2	0.493
	LLaMA + F_r	73.8	80.5	0.486	70.0	75.3	0.516	72.0	82.2	0.493
	GPT-NeoxT	73.0	81.0	**0.478**	69.8	74.0	0.518	71.0	**82.4**	**0.485**
	GPT-NeoxT + F_r	73.1	81.2	0.480	69.8	74.0	0.517	69.7	82.3	**0.485**
	FLAN-UL2	73.9	81.4	0.493	71.1	77.4	**0.514**	**73.2**	81.0	0.497
	FLAN-UL2 + F_r	74.1	**81.5**	0.494	**71.2**	**77.5**	**0.514**	72.1	81.0	0.497
SC. ATTN.	Base	**76.6**	77.7	0.526	**72.0**	76.2	0.538	71.7	78.1	0.509
	RGF	75.2	79.7	0.511	69.8	75.8	0.523	70.8	79.0	0.503
	LLaMA	74.0	80.4	0.489	70.4	75.6	0.528	72.7	82.5	0.494
	LLaMA + F_r	74.1	80.5	0.490	70.5	75.8	0.529	72.4	82.3	0.495
	GPT-NeoxT	72.7	80.9	**0.481**	70.3	74.5	**0.526**	72.9	83.4	**0.486**
	GPT-NeoxT + F_r	72.9	81.0	0.482	70.2	74.8	**0.526**	72.0	83.0	0.487
	FLAN-UL2	74.5	81.3	0.495	71.3	77.7	0.527	**73.7**	81.0	0.502
	FLAN-UL2 + F_r	74.4	**81.4**	0.496	71.6	**78.1**	0.528	73.3	**81.3**	0.501
IG	Base	**76.7**	77.7	0.526	**72.0**	76.2	0.536	70.4	77.0	0.507
	RGF	75.3	79.7	0.511	69.9	75.9	0.519	70.3	78.7	0.498
	LLaMA	74.0	80.4	0.489	70.6	75.8	0.526	71.6	81.9	0.495
	LLaMA + F_r	74.0	80.5	0.490	70.5	76.0	0.527	72.0	82.2	0.495
	GPT-NeoxT	73.0	80.7	**0.482**	70.0	74.3	0.525	71.6	82.4	0.488
	GPT-NeoxT + F_r	73.0	80.8	0.483	70.2	74.6	**0.524**	70.3	**82.5**	**0.487**
	FLAN-UL2	74.2	81.3	0.496	71.4	77.9	0.528	**73.1**	80.7	0.502
	FLAN-UL2 + F_r	74.3	**81.4**	0.497	71.7	**78.3**	0.529	73.0	80.9	0.501
ATTN.	Base	**76.6**	77.5	0.527	**71.5**	75.4	0.536	70.9	77.3	0.509
	RGF	75.4	79.8	0.511	70.1	75.6	0.520	70.6	78.6	0.502
	LLaMA	73.9	80.4	0.490	70.2	75.5	0.526	72.6	82.6	0.494
	LLaMA + F_r	74.0	80.6	0.490	70.4	75.7	0.527	72.2	82.4	0.495
	GPT-NeoxT	72.9	80.8	0.481	69.7	74.1	0.524	71.7	**82.7**	**0.485**
	GPT-NeoxT + F_r	73.1	81.0	0.482	69.8	74.4	**0.523**	70.5	82.6	0.486
	FLAN-UL2	74.1	81.2	**0.496**	71.2	77.5	0.526	**73.0**	80.7	0.503
	FLAN-UL2 + F_r	74.2	**81.4**	0.497	71.4	**77.8**	0.528	71.9	80.2	0.504
INPXGRAD	Base	**76.9**	77.9	0.524	**71.9**	76.0	0.534	70.7	76.8	0.508
	RGF	75.7	79.9	0.510	70.1	75.7	0.520	70.4	78.5	0.500
	LLaMA	73.8	80.4	0.488	70.3	75.7	0.525	71.8	82.0	0.495
	LLAMA + F_r	73.8	80.5	0.490	70.5	76.0	0.526	71.7	81.9	0.495
	GPT-NeoxT	72.8	80.9	**0.481**	70.2	74.4	**0.523**	71.7	82.8	**0.487**
	GPT-NeoxT + F_r	72.9	81.0	0.482	70.2	74.6	0.524	70.9	**83.0**	**0.487**
	FLAN-UL2	74.2	81.4	0.494	71.2	77.7	0.527	73.0	80.7	0.503
	FLAN-UL2 + F_r	74.2	**81.5**	0.496	71.6	**78.1**	0.528	**73.1**	80.8	0.502

Table 11: *Cont.* Calibration results for a Roberta-base model trained on SQuAD when transferring to out-of-domain settings using explanations based on attention and gradients.

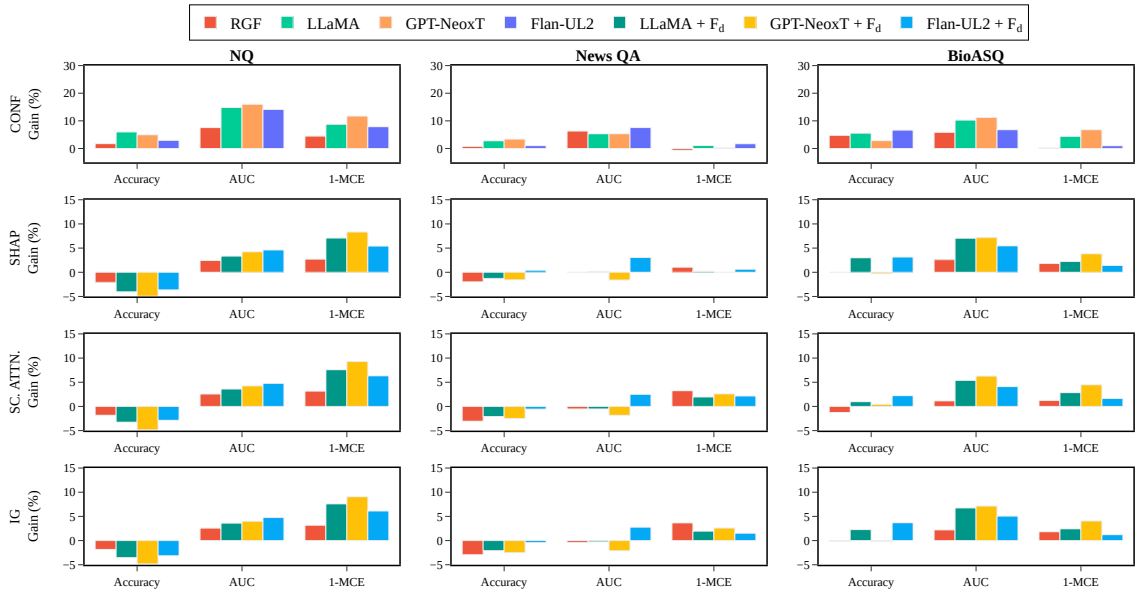

Figure 6: Percentage improvement of CF-augmented models' calibration performance over the unaugmented RoBERTa-base model trained on SQuAD, using features based on probability (CONF) and rationales from SHAP, scaled attention and integrated gradients. The results for CONF (row #1) are reported on models which do not use explanation-based features. In the remaining experiments (other rows), along with BASE and RGF, we report the results of dense-feature augmented calibrators.

explanations.

J.2 Desiderata of Explanations for Calibration

In Table 12, we report the comprehensiveness and sufficiency scores of explanations generated by the baseline and CF-augmented models for all the OOD datasets under evaluation. The RGF baseline provides the most comprehensive explanations across all datasets and explanation methods when compared to our CF-augmented models, while in terms of sufficiency, we observe an opposite trend with the RGF baseline performing worse than all the CF-augmented models. We hypothesize that the CF augmented models produce highly sufficient explanations i.e. assign more importance to a smaller subset of tokens since the diverse CFs help the base model in discerning specific key input features important for the prediction – resulting in better OOD and calibration performance.

	Model	Comprehensiveness (↑)					Sufficiency (↑)				
		α	$\alpha\nabla\alpha$	x∇x	IG	SHAP	α	$\alpha\nabla\alpha$	x∇x	IG	SHAP
Squad Adv	Base	0.33	0.35	0.37	0.38	0.34	0.39	0.39	0.39	0.39	0.40
	RGF	**0.34**	**0.39**	**0.38**	**0.40**	**0.37**	0.36	0.38	0.37	0.37	0.36
	LLaMA	**0.34**	0.36	0.36	0.36	0.33	0.40	0.41	0.40	0.40	0.41
	GPT-Neox	0.30	0.35	0.36	0.36	0.32	**0.41**	**0.42**	**0.42**	**0.42**	**0.43**
	Flan-UL2	0.32	0.36	0.37	0.36	0.33	0.39	0.39	0.39	0.39	0.40
Trivia QA	Base	0.35	0.35	0.37	0.39	0.35	0.56	0.55	0.56	0.56	0.54
	RGF	**0.37**	**0.43**	**0.45**	**0.47**	**0.44**	0.43	0.44	0.44	0.45	0.43
	LLaMA	0.29	0.33	0.33	0.34	0.29	0.57	0.59	0.58	0.59	0.58
	GPT-Neox	0.26	0.29	0.29	0.30	0.25	**0.63**	**0.64**	**0.64**	**0.64**	**0.63**
	Flan-UL2	0.32	0.36	0.37	0.39	0.33	0.51	0.51	0.52	0.53	0.51
Hotpot QA	Base	0.29	0.29	0.32	0.34	0.31	0.52	0.52	0.53	0.53	0.52
	RGF	0.33	**0.38**	**0.39**	**0.41**	**0.57**	0.37	0.38	0.37	0.37	0.37
	LLaMA	0.30	0.31	0.31	0.31	0.29	0.51	0.52	0.51	0.51	0.53
	GPT-Neox	0.28	0.29	0.28	0.30	0.27	**0.56**	**0.57**	**0.57**	**0.57**	**0.58**
	Flan-UL2	**0.35**	0.34	0.35	0.36	0.33	0.43	0.43	0.43	0.43	0.44
NQ	Base	0.35	0.34	0.36	0.37	0.34	0.55	0.55	0.55	0.55	0.55
	RGF	0.37	**0.39**	**0.43**	**0.43**	**0.38**	0.46	0.46	0.47	0.47	0.46
	LLaMA	0.36	0.35	0.34	0.35	0.32	**0.56**	**0.56**	0.55	0.55	**0.55**
	GPT-Neox	0.35	0.32	0.33	0.34	0.31	**0.56**	**0.56**	**0.56**	**0.56**	**0.55**
	Flan-UL2	**0.40**	0.37	0.38	0.38	0.35	0.49	0.50	0.49	0.49	0.49
News QA	Base	0.34	0.35	0.37	0.40	0.43	0.55	0.55	0.56	0.57	0.54
	RGF	**0.36**	**0.41**	**0.41**	**0.43**	**0.49**	0.45	0.48	0.46	0.46	0.45
	LLaMA	0.30	0.32	0.32	0.33	0.37	0.61	0.62	0.61	0.62	0.60
	GPT-Neox	0.28	0.29	0.30	0.31	0.35	**0.63**	**0.64**	**0.64**	**0.64**	**0.62**
	Flan-UL2	0.32	0.37	0.37	0.40	0.45	0.52	0.54	0.53	0.54	0.52
BioASQ	Base	**0.32**	0.38	0.38	0.40	0.35	0.50	0.51	0.51	0.51	0.50
	RGF	0.31	**0.44**	**0.41**	**0.42**	**0.38**	0.41	0.45	0.43	0.43	0.41
	LLaMA	0.30	0.34	0.33	0.34	0.32	**0.57**	**0.58**	**0.57**	**0.58**	**0.57**
	GPT-Neox	0.24	0.33	0.32	0.34	0.30	0.56	**0.58**	**0.57**	**0.58**	**0.57**
	Flan-UL2	0.29	0.38	0.38	0.39	0.35	0.49	0.50	0.50	0.50	0.49

Table 12: Comprehensiveness and sufficiency scores of explanations generated by baseline and counterfactual augmented models. Numbers marked in **bold** represent the highest scores for the particular dataset with a corresponding model and explanation.