Causal Reasoning through Conceptual Explanation Generation

Athiya Deviyani
Carnegie Mellon University
adeviyan@cs.cmu.edu

Mehak Malik
Carnegie Mellon University
mehakm@cs.cmu.edu

Prasoon Varshney
Carnegie Mellon University
pvarshne@cs.cmu.edu

Abstract

Understanding causality has the potential to improve robustness, fairness, and interpretability of Natural Language Processing (NLP) models. In this work, we focus on the task of model-based causal reasoning (CR) and conceptual explanation generation (EG) for causal facts. We train and evaluate numerous baseline large language models for both tasks using the recently developed human-annotated **e**xplainable **CA**usal **RE**asoning (e-CARE) dataset. We also find that multitask learning on the above two related tasks of CR and EG improves performance on both tasks. In future work, we aim to improve the current models using methods such as using a special BERT-based framework for causal reasoning, abductive commonsense reasoning, prompt-based fine-tuning, and question generation and answering. Our code is available on GitHub[1].

1 Introduction

The field of Natural Language Processing (NLP) has been observing remarkable growth due to the introduction of several high-capacity neural architectures such as BERT (Devlin et al., 2019), which are able extract correlations from large-scale datasets. However, these models make no distinction between causes, effects, or confounders, and they make no attempt to identify causal relationships. This may lead to these largely correlational models to be untrustworthy in their predictions (Jacovi et al., 2021). By being heavily reliant on spurious correlations, these models may perform poorly across different groups of users (Zhao et al., 2017) or in out-of-distribution (OOD) settings (McCoy et al., 2019). Feder et al. (2022) suggested that these shortcomings can be addressed by the causal perspective.

Causal reasoning is central to human intelligence (Waldmann and Hagmayer, 2013). By reasoning about the observed facts around them, humans are able to use causal knowledge as the basis of predictions, decision making, problem solving, and more. Understanding this reasoning capability is key to allowing complex models to reason like humans, and make robust and explainable decisions.

There have been multiple attempts to build causal reasoning models for specific tasks, such as controllable text generation (Hu and Li, 2021), named entity recognition (Zeng et al., 2020), and information extraction (Nan et al., 2021), and uncovering biases in visual question answering (Niu et al., 2021). However, their performances still lag far behind humans, are susceptible to adversarial attacks (McCoy et al., 2019).

Du et al. (2022) speculated that causal reasoning models lag behind humans because humans naturally have a deep conceptual understanding of causality and can explain observed causal facts based on world knowledge, while most causal reasoning models only learn to induce empirical causal patterns predictive to a specific label (such as *cause-effect*, *entailment*, *contradiction*, etc.). On the other hand, conceptual explanations of causal patters can help a model in the reasoning process, much like chain of thought prompting has been shown to elicit reasoning capabilities (Wei et al., 2022). To this extent, they introduced the **e**xplainable **CA**usal **RE**asoning (e-CARE) dataset, which contains over 21K multiple-choice causal reasoning questions and over 13K unique conceptual explanations about the deep understanding of the causal facts.

In this work, we reproduce the current state-of-the-art models on this dataset and thoroughly evaluate their performance. Further, based on our error analysis and evaluation of previous literature, we identify some methods to address the limitations presented by the models and plan to attempt these in a future work. These methods include using CausalBERT, abductive commonsense reasoning, prompt-based fine-tuning, and question answering.

[1] https://github.com/fly-back/e-CARE

2 Related work

2.1 Causal reasoning in NLP

The main goal of causal reasoning is to understand the general causal dependency between common events or actions. This understanding is essentially equivalent to measuring the *plausibility* of one event statistically leading to another.

For this, Luo et al. (2016) proposed a framework to deduce causality by harvesting a causality network (CausalNet) from a cause-effect sentence pairs dataset (Roemmele et al., 2011). Their method was quite simple, to build a graph with nodes representing unigrams and edges representing directed co-occurrences of the two words in a cause-effect sentence pair. Thus, the graph encodes how many times a word w_i in `cause` causes a word w_j to be in the `effect`.

Ning et al. (2018) suggests that identifying both temporal and causal relations between events is a fundamental natural language understanding task. They propose a novel Temporal and Causal Reasoning (TCR) framework which jointly extracts temporal and causal relations, which involves a constrained conditional model (CCM) (Chang et al., 2012) and an integer linear programming (ILP) objective (Roth and Yih, 2004) to enforce declarative constraints, such as how a cause must temporally precede its effect, during the inference phrase.

The Choice of Plausible Alteratives (COPA) dataset (Roemmele et al., 2011) propose a causal inference task formulated closely to a multiple choice question-answering, where the question is a premise and the choices are two hypothesis, one being more plausible than the other. This dataset has been a widely used benchmark for causal reasoning models.

Since causal reasoning is widely used to understand and explain model decisions, they are commonly found in models used in critical decision making settings. De Choudhury et al. (2016) used the propensity score matching to understand the causal relationship between linguistic and social interaction-based measures on Reddit text and suicide attempt. Finally, the randomized controlled trial (RCT) method (McGovern, 2001) was used to understand how the gender or racial identity of the judge affects the text of legal rulings (Gill and Hall, 2015). Therefore, improving the reasoning ability of causal models will not only benefit the NLP community, but also encourage the progress of other intersectional fields as well.

2.2 Explanation generation of causal facts

Motivated by the fact that humans do not learn solely from supervised labeled examples supplied by a teacher, but by seeking conceptual understanding of a task through both demonstrations and explanations, Camburu et al. (2018) collected e-SNLI, a large corpus of human-annotated explanations for the Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015). In addition to providing explanations, the annotators also highlighted words which are considered to be essential for the label. These highlighted words in the e-SNLI dataset are also used as a part of the Evaluation Rationales And Simple English Reasoning (ERASER) benchmark proposed by DeYoung et al. (2019), which contains a unified set of diverse NLP datasets containing human rationales for decisions.

Camburu et al. (2018) trained models on the e-SNLI dataset and gauge for their ability for multiple tasks, such as the ability to predict a label and generate an explanation for the predicted label (PREDICTANDEXPLAIN). For this task, they have used the InferSent architecture and conditioned the explanation on the label, and prepend the label as a word at the beginning of the explanation. Although they achieved a reasonable performance, we can notice from figure 1 that the gold-standard explanations mainly contain words from the premise and hypothesis, and do not reason about the label conceptually or beyond how the premise implies/does not imply the hypothesis. Therefore, the generated explanations would most likely be unable to generate conceptual explanations of the causal relationship between the premise and hypothesis.

Premise	A man in an orange vest leans over a pickup truck.
Hypothesis	A man is touching a truck.
Label	Entailment
Explanation	Man leans over a pickup truck implies that he is touching it.

Figure 1: An example instance from e-SNLI with human-annotated explanations. The highlighted words are words annotators considered essential for the label.

One might argue that to generate conceptual explanations, we will need to imbue external knowledge to the model to be used to reason about how a causal relationship is established. Inspired by the concept of abductive reasoning, or inference to the most plausible explanation, Bhagavatula et al. (2019) introduced a challenge dataset, ART, which consists of over 20k commonsense narrative contexts and 200k human explanations. They also introduced two subtasks related to abductive com-

monsense reasoning, namely (1) Abductive Natural Language Inference (aNLI), which is a multiple-choice question answering task for chooisng the more likely explanation, and (2) Abductive Natural Language Generation (aNLG), which is a conditional generation task for explaining given observations in natural language. For the latter task, they used ATOMIC (Sap et al., 2019) as their knowledge base for commonsense reasoning.

3 Methodology

3.1 Dataset

In this work, we use the e-CARE (Du et al., 2022) dataset, which is the largest human-annotated causal reasoning dataset containing over 21K pairs of causal reasoning questions and their corresponding natural language explanations. Each instance of the e-CARE dataset consists of two components: (1) a multiple-choice causal reasoning question which contains a premise and two hypotheses, with one of the hypotheses forming a valid causal fact with the premise, and (2) free-text-formed conceptual explanations to explain why the causation exists. An example of an instance from the e-CARE dataset is shown in figure 2. Additionally, the instance also contains an ask-for indicator which decides whether the premise or the candidate hypothesis to be the cause or effect, respectively.

Premise	Tom holds a copper block by hand and heats it on fire.
Ask-for	Effect
Hypothesis 1	His fingers feel burnt immediately.
Hypothesis 2	The copper block keeps the same.
Explanation	Copper is a good thermal conductor.

Figure 2: An example instance from the e-CARE dataset. The correct hypothesis is highlighted.

3.2 Task description

3.2.1 Causal reasoning task

The causal reasoning task is formulated as a multiple-choice task to choose the hypothesis which forms a valid causal fact with the premise. For example, in figure 2, the hypothesis *"His fingers feel burnt immediately"* forms a valid causal fact with the premise *"Tom holds a copper block by hand and heats it on fire."*, as observing the aforementioned premise causes the corresponding hypothesis, and the ask-for indicator *"effect"* signifies that the hypothesis is an effect of the premise not the cause. In our case, causal reasoning task is casted as a prediction problem, where the input of the model is candidate causal fact containing a premise and hypothesis pair, and the output is a score measuring the reasonableness of the candidate causal fact.

3.2.2 Explanation generation task

Given a premise and the correct hypothesis, the model will generate an explanation in natural language to highlight why a causal relationship exists between the premise and the correct hypothesis, and finally reach a plausible conceptual explanation which goes beyond the isolated facts and reveal the principle of the causal mechanism. In figure 2, we want to find an explanation that connects the premise “Tom holds a copper block by hand and heats it on fire." to the effect “His fingers feel burnt immediately". The corresponding explanation points out the nature of copper which causes anyone holding heated copper to feel their fingers burnt immediately.

3.3 Models

The causal reasoning task is framed as a prediction task: given a premise and a choice of two hypotheses, the hypothesis with the highest reasonableness score will be chosen as the correct one. The authors evaluated the performance of several state-of-the-art discriminative language models on the causal reasoning task, namely BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019b), and ALBERT (Lan et al., 2019), as well as autoregressive generative pretrained language models adapted for the predictive causal reasoning task such as GPT2 (Radford et al., 2019) and BART (Lewis et al., 2020).

For the explanation generation task, the authors trained a GRU-based Seq2Seq model (Chung et al., 2014) and finetuning GPT2 (Radford et al., 2019). Given a premise and the correct hypothesis, the ask-for indicator denotes which of the premise or the hypothesis is the cause or the effect. From this information, we are able to construct the input to the models in the form of the concatenation of the cause and effect from the premise and hypothesis.

3.4 Metrics

We will employ accuracy to evaluate the performance of the causal reasoning models, where a correctly matched premise and hypothesis would be classified as one correct prediction instance. To evaluate generated explanations, there are a number of metrics that are commonly in use such as BLEU (Papineni et al. (2002)) and ROUGE (Lin

(2004)). We will be evaluating our models on BLEU, ROUGE and perplexity.

4 Experimental setup

For baseline reproduction, we very closely followed the setup presented in (Du et al., 2022) for both the causal reasoning and explanation generation tasks. It is important to note that while the authors published their code repository, it had bugs and was not in a runnable state. The baseline reproduction required us to fix their implementation for all tasks.

We'd like to note here that the test set is blind, i.e. it is not publicly available. Benchmarking on the test set requires additional author permissions to submit to their task leaderboard. As such, we leave submission to this leaderboard to future work, once we have substantial improvements. We report the relevant dataset splits in table 1. For both the tasks, we used a `g4dn.2xlarge` AWS instance with a 16GB Nvidia Tesla T4 GPU.

4.1 Causal reasoning

For the causal reasoning task, we finetuned all pretrained large language models for 5 epochs with a batch size of 64 and learning rate of 2e-5. Note that while the authors present baseline results with a learning rate of 1e-5, we empirically found a learning rate of 2e-5 to work better consistently for all 8 pretrained models tested.

Ask-for	Train	Dev	Test	Total
Cause	7,617	1,088	2,176	10,881
Effect	7,311	1,044	2,088	10,443
Total	14,928	2,132	4,264	21,324

Table 1: e-CARE dataset split distribution by question type

4.2 Explanation generation

For the explanation generation task, we finetuned GPT2 for 10 epochs with a batch size of 32 and learning rate of 2e-5. We ran multitask learning with GPT2 to generate cause-effect explanations and then perform the reasoning task. The training/development/test split consists of 10,491/2,012/3,814 explanation sentences respectively.

5 Results

Table 2 presents our results on the causal reasoning task. We benchmark a total of 8 models, 5 discriminative models pretrained with a masked language modeling objective, and 3 generative autoregressive models with a sequence classification head. As discussed in section 4.1, on optimizing the learning rate, we are able to marginally exceed the baseline performance numbers presented by Du et al. for all models except XLNet.

Table 3 shows the results of baseline models over the explanation generation task. Our baseline implementation outperforms the reference implementation in almost all performance metrics except perplexity where it closely resembles the reference implementation.

6 Analysis

6.1 Quantitative analysis

6.1.1 Causal Reasoning

In line with the findings of e-CARE authors, we find that the vanilla BERT model (Devlin et al., 2019) performs better than its variants. In general, the masked language models perform better than the autoregressive models on the reasoning task. We hypothesize that BERT outperforms the other models because its pretraining is based on Wikipedia and the BooksCorpus. These datasets encode a lot of concepts, properties, and relationships between entities and concepts like copper, thermal conductance, etc. On the other hand, models like GPT2 are trained on large-scale social media data which is full of real and fake news, opinions, toxicity, jokes, etc. that are largely irrelevant to reasoning between a cause and its effect.

Finally, from a cursory look of the dataset, we noticed that the premise and the two hypotheses sentences are usually short, and often contain repeating entities. For instance, in the cause-effect pair <"Adding rock into acid.", "Rock dissolved.">, the entity rock repeats. However, the case of the first letter 'r' is different in the two sentences. Given the reasoning task is happening between entities in the two sentences, we hypothesized that it's better to use a model that's agnosting to case instead of being sensitive. Therefore, we tried `bert-base-uncased` in addition to `bert-base-cased`. In line with our hypothesis, we saw a performance increase of +1.12% (75.66% to 76.78%), which is a significant improvement over the best results presented in the baseline.

	Our Implementation	Reference[2]Implementation(Du et al., 2022)	
Model	**Dev Set**	**Dev Set**	**Test Set (publicly unavailable)**
Masked Language Models			
BERT (base,uncased)	**76.78%**	NR	NR
BERT (base,cased)	75.66%	75.47%	75.38%
AlBERTa (base,v2)	74.25%	73.97%	74.6%
XLNet (base,cased)	74.2%	75.61%	74.58%
RoBERTa (base)	71.34%	70.64%	70.73%
Causal/Autoregressive Language Models			
BART (base)	73.83%	73.03%	71.65%
GPT2	70.64%	70.36%	69.51%
GPT	69.75%	67.59%	68.15%

Table 2: Accuracy for various pretrained large language models on the Causal Reasoning task. NR ≡ Not Reported.

Model	Accuracy	BLEU-1 ↑	BLEU-4 ↑	ROUGE-1 ↑	ROUGE-l ↑	Perplexity ↓
Our Implementation (Dev)						
$GPT2_{CR}$	70.64%	-	-	-	-	-
$GPT2_{EG}$	-	54.26	18.56	33.55	32.41	6.72
GPT_{CR-EG}	**71.63%**	**56.42**	**23.06**	**37.15**	**36.29**	**6.49**
Reference[1] Implementation (Du et al., 2022) (Test Set; publicly unavailable, NR on Dev Dataset)						
$GPT2_{CR}$	69.51%	-	-	-	-	-
$GPT2_{EG}$	-	55.17	18.79	33.17	32.05	6.87
GPT_{CR-EG}	71.58%	56.32	22.36	35.70	34.88	6.64

Table 3: Results for GPT2 on the Explanation Generation task with and without multitask learning. Up arrow ≡ higher is better. Down arrow ≡ lower is better. NR ≡ Not Reported.

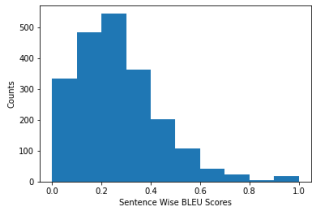

Figure 3: Histogram of Sentence-wise BLEU scores of Explanations Generated using GPT2 based Model

6.1.2 Explanation Generation

As we can see in figure 3, majority of explanations generated by the GPT2 based model have a BLEU score between 0.00-0.50. There is a higher concentration between 0.00-0.25. This shows that majority of sentences have a very low score. This correlates to whether or not the BLEU score is an effective metric for measure explanations of causal facts. We can see via qualitative examples in error analysis that there are instances where even though generated explanation is reasonably correct, metrics such as BLEU and ROUGE award poor scores since it is different or a paraphrasing of the gold standard explanation.

6.2 Error analysis

6.2.1 Causal Reasoning

Upon examining the results of causal analysis we see that the best MLM-based model, i.e, BERT-base often makes errors when it comes to premises that have more ambiguity in terms of subject or use pronouns instead of common nouns. For example, for the premise, "It is difficult for him to get mind equilibrium now." the model incorrectly chooses a completely irrelevant hypothesis, "He suffers from epilepsy." as opposed to the correct hypothesis "Henry always runs into hindrances in doing business." This could possibly be due to the ambiguity presented by the word "it". There are also other errors where it seems like the model does not have enough external information to select the correct premise. For example, consider the premise "Human might get infected owing to sick mosquitoes." Deciding that "Virus attacked the mosquitoes." is the correct premise requires understanding the relationship between viruses and mosquitoes. Ultimately, from the error analysis, we hypothesize that we require language models to make connections between entities and use the connections to make decision regarding which hypothesis is plausible.

6.2.2 Explanation Generation

Upon examining our results we can see that there are numerous different kinds of errors in the explanation generation model. On particular class of errors is a set of explanations that seem slightly relevant to the premise and hypothesis, but they seem to be a circular definition of a relevant term. Examples of such "circular" explanations is shown in table 4. In example 2, we see friction being referred

[2]Reference implementation results available on Du et al.'s official Github repository.

to as a cause of friction on roads which is simply not true and does not convey information. Another interesting point here is that friction is relevant to neither the premise nor the hypothesis.

Another class of errors involves instances where BLEU or ROUGE penalize the model even though the explanations are reasonably correct from a human evaluation perspective. Such examples are shown in table 5. For example, in the first example, the generated explanation reasonably demonstrates a strong understanding of why the hypothesis is valid, it receives a BLEU score of 0.07. This occurs because while the generation is a strong explanation for causality, it is worded differently than the gold standard. Similarly, the second example is a little more verbose than the ground truth but effectively the same semantically but recieves a BLEU score of 0.34. The final class of errors is where the explanation semantically makes sense and are relevant. They are shown in table 6. We can see that these are semantically correct and relevant statements but not the best explanations or sufficiently justify the hypothesis.

7 Future work

7.1 CausalBERT

For the causal reasoning task, we are interested in exploring CausalBERT (Li et al., 2021b) and its extensions (Li et al., 2021a). CausalBERT is a three-stage sequential transfer learning framework (Li et al., 2019): (1) large-scale unsupervised pre-training tasks with language modeling objective, (2) self-supervised pre-training with the different causal pairs, and (3) direct causal pair classification or further fine-tuning. The second stage involves two different pre-training tasks, namely causal pair classification or ranking. For future work, we will explore the performance of the CausalBERT framework fine-tuned on the e-CARE dataset on the causal reasoning task.

7.2 Abductive commonsense reasoning

Bhagavatula et al. (2019) proposed a dataset for the abductive commonsense reasoning, split into two subtasks: abductive natural language inference (aNLI) and abductive natural language generation (aNLG). We are mainly interested in their methodology towards the aNLG task, where given two observations, the task is to generate a valid hypothesis (explanation). They have utilized a GPT-2 model (Radford et al., 2019) conditioned on the tokens of the two observations to generate the hypothesis. Additionally, both observations are enclosed with field-specific tags, and they have used ATOMIC (Sap et al., 2019), a repository of inferential if-then knowledge as a natural source of background commonsense to reason about the narrative context in the ART dataset. The knowledge from ATOMIC is not directly compatible with a neural model, therefore they have utilized COMeT (Bosselut et al., 2019), which is a transformer model trained on ATOMIC that generates nine commonsense inferences of events in natural language. The information from COMeT is then integrated to GPT2 as either textual phrases or as embeddings.

In our case, we can use this task by assigning a premise and valid hypothesis pair from the e-CARE dataset as the two observations, and generate the explanation. Given the promising results reported by the authors, we would like to explore this methodology to improve the explanation generation on the e-CARE dataset.

7.3 Prompt-based attribute conditioning

Let C denote the cause sentence, E the effect sentence, and Exp the explanation. The models presented in Section 4 use a naive concatenation of the cause and effect sentence pairs as C <SEP> E (with Exp being the target generation) and do not use any special tokens or prompts to let the model learn the structure underneath. We plan to modify the entire dataset to have the following structure for training: `For <cause>` C, `<effect>` E is because `<explanation>` Exp.

At inference time, the utterances would not include Exp at the end, as the model would be expected to generate the explanations. The tokens `<cause>`, `<effect>`, `<explanation>` will be added as special tokens to GPT-2 vocabulary such that the model can learn to enter an Exp generation model when it encounters `<explanation>` as a prompt.

7.4 Question generation and answering

Another way we can formulate the explanation generation task is to view it as a two-part open-domain question-answering task: (1) question generation and (2) question answering. We describe this process at a high-level in figure 4. In this figure, the question was generated manually by a human by looking at the cause-effect pair, however the answer was generated automatically by the `text-davinci-002` GPT3 model (Brown

		Explanations	
Premise	**Hypothesis**	**Ground Truth**	**Generated**
The worker fremented some sugar cane with yeast.	He got some rum.	Rum is made from sugar cane using yeast fermentation.	Rim is a type of sugar cane made from the sugar cane.
There are many obstacles on the track.	Racing drivers will encounter common dangers from time to time.	Obstacles present common danger.	Friction is one of the most common causes of friction in the road.

Table 4: "Circular Explanation" Type Error in explanation generation using GPT2

		Explanations	
Premise	**Hypothesis**	**Ground Truth**	**Generated**
Simon's mother often let Simon eat soys.	Simon grows healthy.	Soy offers health benefits.	Soy is a plant-based food source of vitamins A, C, D, E, etc.
My car is exposed to the acid rain.	The paint on my car is clipped.	Acid rain could effect a car's color.	Acid rain is a common cause of damage to the car's paint.

Table 5: Example explanations generated using GPT2 demonstrating the failure of BLEU and ROUGE as evaluation metrics.

et al., 2020) which was not fine-tuned for a QA task. This example provides us a motivation to move forward with this proposed methodology.

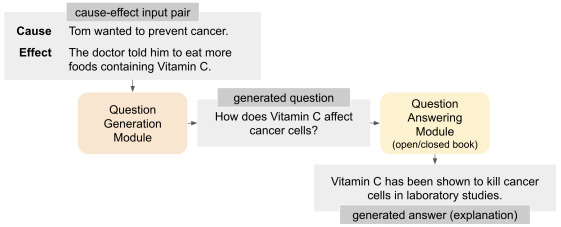

Figure 4: Explanation generation through question generation and answering.

7.4.1 Question generation

The question generation task is formulated as follows: given a premise and the correct hypothesis as a cause and effect pair, generate a question such that the answer would form an explanation for the causal relationship. Stasaski et al. (2021) has built a pipeline which extracts causal relations from passages of input text, retrieve cause and effect pairs from the passage, and feed these pairs to a neural question generator. Their work results in a novel and publicly available collection of cause-and-effect questions. They have used a ProphetNet model (Qi et al., 2020) fine-tuned on SQuAD 1.1 (Rajpurkar et al., 2016) to generate their questions. We can easily adopt their methodology to solve our question generation task, given that we can skip the causal relationship extraction as we already have the cause and effect (premise and correct hypothesis) pairs.

7.4.2 Open-book QA

In an open-book QA system, the QA module is paired with a knowledge base to identify relevant documents as evidence of answers. Yang et al. (2019a) introduced BERTserini, an end-to-end question answering system that integrates BERT with the open-source Anserini (Yang et al., 2017) information retrieval toolkit. The retrieved information serves as the context, and this will be fed to the fine-tuned BERT along with the question. We have tried this methodology with a manually generated question based on a cause-effect input pair, and we have found that the answers obtained through BERTserini are appropriate for our use case. An example is shown in figure 6 in the appendix.

7.4.3 Closed-book QA

Large language models are sometimes able to encode a surplus of factual knowledge, which allows them to perform question-answering without explicit context. Roberts et al. (2020) fine-tuned the T5 language model (Raffel et al., 2019) to answer questions without inputting any additional information or context. They performed continual pre-training with salient span masking over the Wikipedia corpus, and fine-tuned the model on specific QA datasets. Although this methodology successfully obtained competitive results in closed-book open-domain QA, the GPT3 model (Brown et al., 2020) performs comparatively well without any gradient updates or fine-tuning. An example generated answer from GPT3 is shown in figure 4.

We would like to evaluate the answer quality from several generative models for generating explanations and compare them with the baselines on the e-CARE dataset. We would also like to explore whether fine-tuning these models on CommonsenseQA (Talmor et al., 2018), which is a dataset aimed to capture common sense beyond associations for commonsense question answering. This will promote the language models to generating explanations that encompasses conceptual understanding of the causal relationship.

References

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

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

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Ming-Wei Chang, Lev Ratinov, and Dan Roth. 2012. Structured learning with constrained conditional models. *Machine learning*, 88(3):399–431.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110.

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. 2019. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-CARE: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446, Dublin, Ireland. Association for Computational Linguistics.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Michael Gill and Andrew Hall. 2015. How judicial identity changes the text of legal rulings. *Available at SSRN 2620781*.

Zhiting Hu and Li Erran Li. 2021. A causal lens for controllable text generation. *Advances in Neural Information Processing Systems*, 34:24941–24955.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 624–635.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Zhongyang Li, Xiao Ding, Kuo Liao, Bing Qin, and Ting Liu. 2021a. Causalbert: Injecting causal knowledge into pre-trained models with minimal supervision. *arXiv preprint arXiv:2107.09852*.

Zhongyang Li, Xiao Ding, and Ting Liu. 2019. Story ending prediction by transferable bert. *arXiv preprint arXiv:1905.07504*.

Zhongyang Li, Xiao Ding, Ting Liu, J Edward Hu, and Benjamin Van Durme. 2021b. Guided generation of cause and effect. *arXiv preprint arXiv:2107.09846*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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. *arXiv preprint arXiv:1907.11692*.

Zhiyi Luo, Yuchen Sha, Kenny Q Zhu, Seung-won Hwang, and Zhongyuan Wang. 2016. Commonsense causal reasoning between short texts. In *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Dermot PB McGovern. 2001. Randomized controlled trials. *Key topics in evidence based medicine. Oxford: BIOS Scientific Publishers*, pages 26–9.

Guoshun Nan, Jiaqi Zeng, Rui Qiao, Zhijiang Guo, and Wei Lu. 2021. Uncovering main causalities for long-tailed information extraction. *arXiv preprint arXiv:2109.05213*.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288, Melbourne, Australia. Association for Computational Linguistics.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model?

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95.

Dan Roth and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, pages 1–8, Boston, Massachusetts, USA. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Katherine Stasaski, Manav Rathod, Tony Tu, Yunfang Xiao, and Marti A Hearst. 2021. Automatically generating cause-and-effect questions from passages. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 158–170.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Michael R Waldmann and York Hagmayer. 2013. Causal reasoning.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 1253–1256.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019a. End-to-end open-domain question answering with bertserini. *arXiv preprint arXiv:1902.01718*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019b.

Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. 2020. Counterfactual generator: A weakly-supervised method for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7270–7280.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

A Appendix

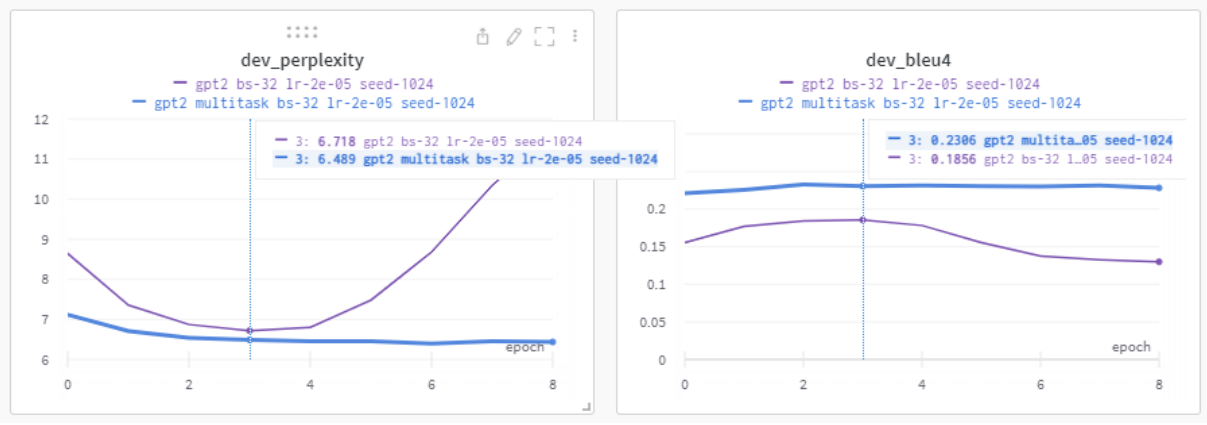

Figure 5: Validation perplexity and BLEU-4 charts for GPT-2 with Multitask Learning. Jointly performing causal reasoning and explanation generation not only increases performance on both tasks, but also mitigates overfitting

		Explanations	
Premise	**Hypothesis**	**Ground Truth**	**Generated**
The lecturer is talking about t lymphocytes.	I can hear t-cells sometimes.	T lymphocytes are also called t-cells.	T-cells are the most common type of lymphocytes in the body.
The financial crisis left many people homeless.	After the financial crisis, the suicide rate increased significantly.	Homelessness greatly increases the likelihood of a suicide attempt.	Suicide is a leading cause of death among young adults in the United States.

Table 6: Example explanations generated using GPT-2 that are coherent and relevant but not explanatory

Cause **Effect**	Tom wanted to prevent cancer. The doctor told him to eat more foods containing Vitamin C.
Generated Q	How does Vitamin C affect cancer cells?
Generated explanation	Neutralize free radicals in the body and thus prevent cell damage and oxidative damage to DNA.

Figure 6: Generated explanation from BERTserini.