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 language models for both tasks using the recently developed human-annotated **e**xplainable **CA**usal **RE**asoning (e-CARE) dataset. However we focus more on explanation generation and explore techniques such as prompting, multitask learning, question generation and answering. We found that neural knowledge graph based approach COMET results in significant improvement in causal explanation generation. 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 commonsense 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, a repository of inferential if-then knowledge as a natural source of background commonsense to reason about the narrative context in the ART dataset. ATOMIC is not directly compatible with a neural model, therefore they utilize COMET (Bosselut et al., 2019), a transformer model trained on ATOMIC that generates nine commonsense inferences of events in natural language.

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. 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.

3.2 Task description

In this work, we will attempt to improve the benchmarks on the tasks introduced by the authors of the e-CARE dataset, namely **causal reasoning** and **explanation generation**. An overview of the tasks and desired results from an instance of the e-CARE dataset is shown in figure 2.

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.

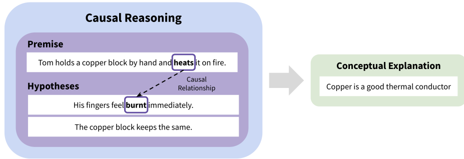

Figure 2: An example of causal reasoning and conceptual explanation generation from an instance of the e-CARE dataset

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 Baseline 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.

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.

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

Further, for generation, while Du et al. use a repetition penalty of 1.5, we hypothesized that since the model needs to reason about entities present in the premise and hypothesis, it at least needs to repeat the entities that are causally linked in these sentence pairs. Based on this hypothesis, we reduced the repetition penalty to 1.2, and saw slightly better results. For consistency, all results in the rest of this work are reported with these modified hyperparameters.

The training/development/test split consists of 10,491/2,012/3,814 explanation sentences respectively.

5 Experiments

In this section, we describe the techniques explored for the two tasks on the e-CARE dataset. We primarily focus on the explanation generation task (approaches detailed in sections 5.2, 5.3 and 5.4). For the causal reasoning task, we implement several baselines from (Du et al., 2022) and verify if the CausalBERT (section 5.1) model can yield improvements over them.

5.1 CausalBERT

For the causal reasoning task, we explore 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 finetuning. The second stage involves two different pre-training tasks, namely causal pair classification or ranking. The architecture of CausalBERT is highlighted in figure 3.

5.2 Prompting

For the explanation generation task, the first idea we explore is prompting GPT-2 by giving a semantic structure to the input sentence pairs and ending them with a prompt that is finetuned to elicit an

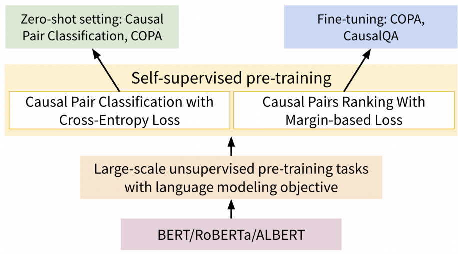

Figure 3: The CausalBERT architecture

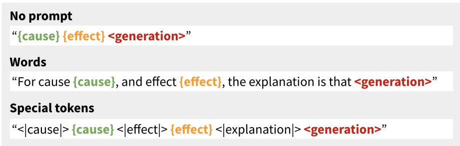

Figure 4: Prompt templates for explanation generation

explanation. This is carried out in two ways, using: (1) English words, and (2) Special tokens. The modifications to the input are described in Figure 4. Here, *{cause}* denote the cause sentence, *{effect}* the effect sentence, and *{generation}* is the placeholder for ground truth or model output. The models presented in Section 4 use a simple concatenation of the cause and effect sentence pairs. We hypothesize that this would make it difficult for the model to relate them as a cause-effect pair as it is not inherently implied by the structure underneath.

Note that prompting with special tokens requires adding the tokens <|cause|>, <|effect|>, <|explanation|> to the tokenizer vocabulary, which is not required when prompting with words. On fine-tuning on the augmented dataset, we expect the model to enter an *"explanation generation mode"* after encountering <|explanation|> in case of special tokens, and *the explanation is that* in case of prompting with words.

5.3 Common sense knowledge injection

Following the approach in Bhagavatula et al. (2019), we use ATOMIC$^{20}_{20}$ (Bosselut et al., 2019) as our large external knowledge base to inject real-world common sense from a knowledge graph of nodes describing entities and edges describing relationships that links the entities. For instance, "node(*money*)-relationship(*has property*)-node(*earned by working*)", or "node(*a mechanic*)-relationship(*is located at*)-node(*garage*)". This knowledge is transferred to existing language models by training them using the COMET transformer which trains on tuples from the knowledge graph to predict the target phrases in the graph given source/head phrases.

We use a pretrained COMET(BART) model and fine-tune it on the task of explanation generation. We use a concatenation of premise and hypothesis from the e-CARE dataset as input to generate an explanation for causality.

5.4 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 5.

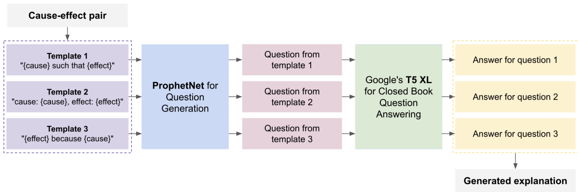

Figure 5: Question generation and question answering pipeline for conceptual explanation generation

5.4.1 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 adopt their methodology to solve our question generation task, given that we can skip the causal relationship extraction (since we have the cause-effect sentence pair). As shown in 5, we have concatenated the cause and effect pairs using various templates to evaluate how to best present these pairs such that the question generation network outputs the most relevant questions.

5.4.2 Question answering

For our task, we use a closed-book T5 XL (Raffel et al., 2019) pretrained question answering model (`google/t5-xl-ssm-nq`), primarily because time and space constraints presented by open-book question answering frameworks such as BERTserini (Yang et al., 2019a) which integrates BERT with the open-source Anserini (Yang et al., 2017) information retrieval toolkit. 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 6.

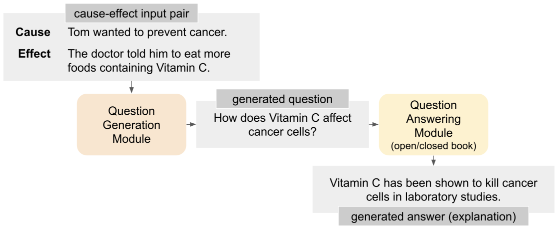

Figure 6: Explanation generation through question generation and answering.

6 Results and Discussion

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 auto-regressive 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 for BART and COMET-BART models outperforms the baseline GPT-2 implementation by a large margin. Our GPT-2 implementation also slightly outperforms the reference implementation using prompting and hyperparameter tuning.

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 auto-regressive models on the reasoning task. We hypothesize that BERT outperforms the other models because its pre-training 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.

Finally, we observe that CausalBERT (Li et al., 2021b) did not improve the causal reasoning ability, even though that it is trained to make distinctions between causes, effects, and confounders. We hypothesize that this is because of lack of relevant knowledge being injected, and explore this in more detail in qualitative analysis in Section 6.2.1

6.1.2 Explanation Generation

For explanation generation, we tried multiple approaches quantitatively compared in Table 3. BART significantly outperforms all GPT-2 based models on both BLEU and ROUGE metrics, achieving an average-BLEU score of 47.46, and COMET-BART further improves for knowledge injection in BART and results in average-BLEU score of 52.52. We can see that our best performing model in terms of all metrics is COMET-BART. This could be indicative of the fact that the models required external facts to generate explanations closer to the ground truth.

Multitask learning was another successful approach that not only yielded improved performance

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

	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%
CausalBERT (Li et al., 2021b)	73.45%	NR	NR
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 $\equiv$ Not Reported.

Model	Accuracy	BLEU-1 $\uparrow$	BLEU-4 $\uparrow$	AVG-BLEU $\uparrow$	ROUGE-l $\uparrow$	Perplexity $\downarrow$
Our Implementation for GPT-2 with ablations for prompting and multitask learning (Dev Set)						
GPT2$_{CR}$	70.64%	-	-	-	-	-
GPT2$_{EG}$	-	53.79	18.2	31.74	35.23	6.69
+ Prompting (Words)	-	54.64	19.91	33.26	36.14	6.59
+ Prompting (ST)	-	54.16	16.52	30.69	31.55	7.99
GPT2-large$_{EG}$	-	53.90	**24.24**	35.96	**42.46**	**4.73**
+ Prompting (Words)	-	52.41	**24.24**	34.69	40.47	4.95
+ Prompting (ST)	-	56.08	22.37	36.09	40.49	4.86
GPT2$_{CR-EG}$	72.62%	55.06	23.37	35.63	35.93	6.44
+ Prompting (Words)	**72.81%**	**57.14**	23.92	**36.73**	36.15	6.41
+ Prompting (ST)	72.05%	56.47	22.57	35.53	35.27	6.62
Our Implementation for BART and COMeT-BART (Dev Set)						
BART$_{EG}$	-	62.68	37.59	47.46	39.75	8.42
COMeT-BART$_{EG}$	-	**67.55**	**42.82**	**52.52**	**46.25**	**3.92**

Table 3: Results and ablations for GPT2 and BART-based models on the Explanation Generation task. The *CR-EG* subscript denotes multitask learning for causal reasoning and explanation generation. Up arrow $\equiv$ higher is better. Down arrow $\equiv$ lower is better. NR $\equiv$ Not Reported. ST $\equiv$ Special Tokens.

Model	Accuracy	BLEU-1 $\uparrow$	BLEU-4 $\uparrow$	AVG-BLEU $\uparrow$	ROUGE-l $\uparrow$	Perplexity $\downarrow$
Reference 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 4: Results for reference baseline implementation[1] (Du et al., 2022) on the Explanation Generation task. Up arrow $\equiv$ higher is better. Down arrow $\equiv$ lower is better.

for GPT-2 on the explanation generation task, but also on the causal reasoning task. For instance, GPT-2 with prompting achieves an accuracy of 72.81% compared to 70.64% when only the causal reasoning task is performed in isolation.

We also explored prompting techniques that improved the performance of baseline GPT-2 model. Prompting with English words to give the cause effect pair some semantic structure consistently performs better for GPT-2 than without any prompt.

For prompting with special tokens, we observe that the model gets confounded in the initial epochs with a very high perplexity. With more epochs, while it slowly reaches close to the performance achieved without prompting. This points to the possibility that fine-tuning with special tokens, while potentially promising, requires more data and training epochs than prompting with words already in model vocabulary.

For the large version of GPT-2 `gpt2-large`, we observe that prompting techniques had a smaller impact on its performance, this could possibly be because having a larger capacity makes it insensitive to addition of a few special tokens. Further, we noticed that while GPT2-large BLEU scores are similar to GPT2-base, its rouge (recall) scores are higher, again indicating that the larger model is able to recollect more words from its pretraining than the smaller model.

6.2 Error analysis

6.2.1 CausalBERT

We qualitatively examined the results of the CausalBERT model and tabulated examples in table 8. We can see from the first example that choosing the correct hypothesis from "There is gravity among planets" and "There is magnetic force among planets", external knowledge of gravitational force is required. Similarly, in the second example there is no causal deduction required to choose the correct hypothesis. The correct hypothesis can only be chosen by awareness of the fact that there are two types of polypropylene and not five. Lastly in the final premise, "he" could refer to "Tom" or the "worker" which makes this example much more confusing. The correct hypothesis being chosen requires the connection between rum and sugarcane to be apparent. Ultimately, from our qualitative analysis of the CausalBERT results we can observe there might be a lack of external knowledge in the model. This is substantiated by the fact that CausalBERT was fine-tuned on Choice Of Plausible Alternatives (COPA) which consists of only 1000 questions. It may not have been possible to inject knowledge relevant to e-CARE dataset with a corpus of this size. This could be a reason why fine-tuning CausalBERT did not improve causal reasoning performance.

6.2.2 COMET-BART

Upon examining the explanations generated by COMET-BART we can see that the quality of the generated explanations is very high.

There are a very high number of instances where the generated explanations sufficiently explain the relation between the premise and the hypothesis while being syntactically and semantically sound. However, since our metrics BLEU and ROUGE focus more on the similarities between the gold standard and generated explanations, these explanations are rejected as they differ from the gold standard in terms of vocabulary, tense and number. Such examples are displayed in table 6. We can see in the final example in the table that the generated explanation differs from the ground truth only in the word "extremely" which is a synonym of the word "intensely" in the ground truth. The generated explanation in this case should receive a full score since it is semantically equivalent to the ground truth. However, since our metrics BLUE and ROUGE do not consider semantics and are focused on matching n-grams, this sentence achieves lower BLEU and ROUGE scores.

There is another kind of error apparent in the generated explanations. On certain occasions, the model links the premise in hypothesis with a statement that is technically true but not the underlying explanation. These errors can be better explained with the examples in the table 5. In the first example, while it is probably implied from true that the keepers encourage reproduction of the animals, it is not a sufficient explanation of why the hypothesis is implied. Similarly, in the second example, it is true that "Re-settlements take place" but that is a generic statement that is true. It does not explain why if Jack's country was at war, he was resettled. These errors could possibly be attributed to how the premise and hypothesis are passed to the model which is via a simple concatenation which does not necessarily require the model to sufficiently explain why the second statement is implied if the first is true.

There are also cases where the ground truth does not sufficiently explain causality but the generated explanations do. Such examples are displayed in table 7. Consider the first example. Given the premise and hypothesis, the generated explanation "Ponds occur in suitable areas" is intuitively a better explanation than "Areas provide water". Similarly, in the next explanation while the ground truth "Cigarettes have significant effects" is a true a statement it does not explain the reason why the individuals fingers are stained. A sufficient explanation is provided by the model which is that "cigarettes can stain a finger". These examples show that even though there are certain inconsistencies and inaccuracies in dataset ground truths, the model is able to generate fairly logical and appropriate explanations.

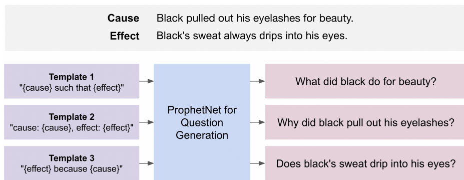

Figure 7: Three different templates used for Question Generation using ProphetNet (Qi et al., 2020) from a causal pair and the corresponding generated questions

6.2.3 Question Generation and Question Answering

For the question generation and answering approach, we generated questions using three different templates used to combine the premise and hypothesis. This process is explained. in figure 7. We qualitatively analyzed the generated questions and answers (generated explanations) for causality. We have tabulated some examples in table 10. We can see for the first example that the question generated using template 1 is not syntactically correct and while the answer is a relevant statement, there is loss of context while converting to a question and then generating the answer for that incorrect question. This trend is observed throughout the results. It also seems that the model generates very specific question instead of generating a general question regarding the hypothesis. The answers generated to the questions are also sometimes in-

		Explanations	
Premise	**Hypothesis**	**Ground Truth**	**Generated**
Spring is the season for animals to reproduce.	The keepers put them in contact with each other.	Reproduction requires contact.	Keepers encourage reproduction.
Jack's country is at war.	He was resettled in Russia.	Resettlement occurs when the refugee has no hope of returning safely to the home country.	Resettlements take places.
Tom followed the flamingo to go back their habitat.	He found that there are a lot of flamingos.	Flamingos live in groups.	Flamingos live in habitats.

Table 5: Examples of Insufficient Explanations Generated by COMET-BART

		Explanations	
Premise	**Hypothesis**	**Ground Truth**	**Generated**
Black pulled out his eyelashes for beauty.	Black's sweat always drips into his eyes.	Eyelashes keep sweat out of the eye.	Eyelashes help to control the amount of sweat dripping into eyes.
Mary read some papers.	She knew lots of details.	Paper gives details.	Details appear in papers.
Jack added nitrites into the water.	The catfish died earlier than the scalefish in the water.	Nitrites are more toxic to catfish than scalefish.	Nitrites kill catfish faster than scalefish.
Jack's interest is to study human species.	He decides to choose anthropology as his major in college.	Anthropology is the discipline devoted to the study of the human species.	Anthropology is the scientific study of human species.
Madame Curie studied radium all her life.	Her body had excessive radiation levels.	'Radium is intensely radioactive.	Radium is extremely radioactive.

Table 6: Example explanations generated using COMET-BART demonstrating the failure of BLEU and ROUGE as evaluation metrics.

correct and irrelevant as seen by the answer to the question "What did Black do for beauty?". The answer "Made up to be beautiful? Asked Dixon's wife" is completely unrelated to the premise and hypothesis and this is due to lack of context i.e., if somehow the fact that this situation was focused on eyelashes and sweat was included in the question, a better question could have been generated. Since there is loss of information while transforming the combination of premise and hypothesis into a an "incorrect" question, the generated answer seems random in such cases. Ultimately, we observe that in many cases the model is unable to generate a question that would be conducive to explanation generation.

7 Conclusion

Given the task of causal reasoning and explanation generation on the e-CARE dataset, we were able to exceed baseline performance in causal reasoning and explanation generation using multiple techniques like prompting and knowledge graph based injection (COMET).

For future directions, an interesting technique followed in the field of pragmatic reasoning for language models is sampling and re-ranking a generative model's outputs based on an independent and separate re-ranking model that evaluates an objective closer to causal strength.

Also, noting the poor performance of our question generation and answering for causal explanation generation, we must explore other ways of generating questions from premise and hypothesis. This could be by providing additional context while generating questions or while using a different model to decide what kind of question would best combine the premise and the hypothesis.

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.

		Explanations	
Premise	**Hypothesis**	**Ground Truth**	**Generated**
This farmer wants to build a fish pond.	The farmer found a suitable area.	Areas provide water.	Ponds occur in suitable areas.
He has been smoking cigarettes for four years.	His finger has been stained with the cigarette.	Cigarettes have significant effects.	Cigarettes can stain a finger.
Tom's heart beat faster because of smoking.	The doctor injected Tom with amiodarone.	Amiodarone decreases the effects of chemicals on the heart.	Amiodarone decreases the urge to smoke by inhibiting blood coagulation.

Table 7: Example generations using COMET-BART that explain causality better than the ground truth

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.

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

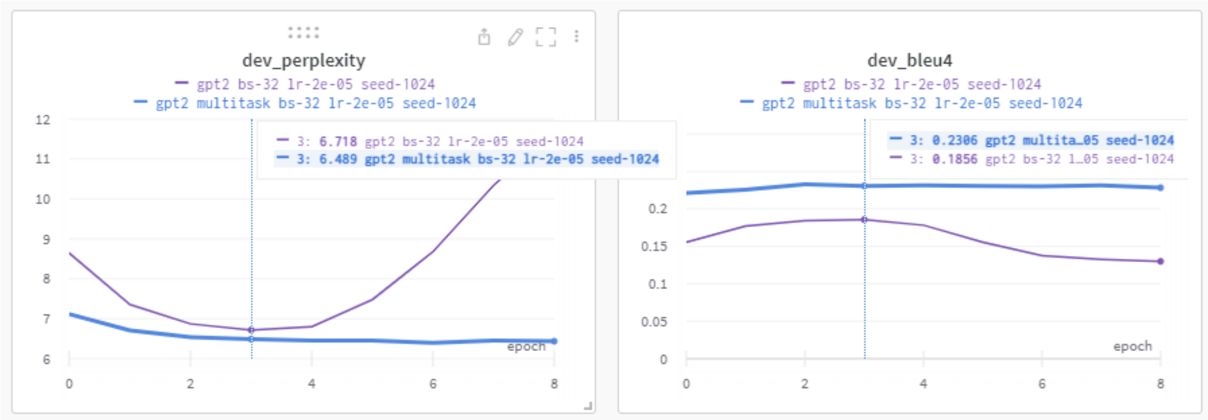

Figure 8: Validation perplexity and BLEU-4 charts for GPT-2 with Multitask Learning. Jointly performing causal reasoning and explanation generation not only increases performance on both tasks, but also mitigates overfitting

Premise	**Ask-For**	**Hypothesis 1**	**Hypothesis 2**
The major planets interact.	Effect	There is gravity among the planets.	There is magnetic force among the planets.
Tom studied the types of polypropylene.	Effect	He found that they came in five types.	He found that they came in two types.
He got some rum.	Cause	The worker fermented some sugar cane with yeast.	Tom went out and want to hunt some cottontails

Table 8: Examples of Incorrect Hypothesis Selected by CausalBERT

Cause	Tom wanted to prevent cancer.
Effect	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 9: Generated explanation from BERTserini.

Model	BLEU-1 ↑	BLEU-4 ↑	AVG-BLEU ↑	ROUGE-l ↑
Question Template 1	42.15	4.29	17.10	4.17
Question Template 2	44.74	4.72	17.89	4.62
Question Template 3	44.06	4.62	18.22	4.82

Table 9: Results for our question generation followed by question answering approach.

Test Data			Question Template 1		Question Template 2		Question Template 3	
Premise	**Hypothesis**	**Ground Truth**	**Question**	**Answer**	**Question**	**Answer**	**Question**	**Answer**
Tom has good eyesight.	Tom found the poisonous snake in time.	Eyesights play roles.	Tom has good eyesight such that he found the poisonous snake in time?	Tom's good eyesight and keen sense of smell	What is the cause of tom finding the poisonous snake?	His father's illness episodes of tuberculosis	How was tom able to find the poisonous snake?	His nose, or on his shoulder, etc
Black pulled out his eyelashes for beauty.	Black's sweat always drips into his eyes.	Eyelashes keep sweat out of the eye.	What did black do for beauty?	Made up to be beautiful? Asked Dixon's wife	Why did black pull out his eyelashes?	To reduce eye bleeding episodes)seen	Does black's sweat drip into his eyes?	Glasses are never worn on black days" (page 2

Table 10: Example generations using Question-Generation Question-Answering Approach following 3 Different Templates