CONCR: A Contrastive Learning Framework for Causal Reasoning

Yinghuan Zhang* and **Yufan Bao*** and **Jiayi Shen***
{yinghuan, yufanb, jiayi2}@andrew.cmu.edu
Carnegie Mellon University

Abstract

This paper presents CONCR, a CONtrastive learning framework for Causal Reasoning that advances state-of-the-art causal reasoning on the e-CARE dataset. CONCR is a model-agnostic framework and works better with better sentence encoders. It discards the projection in previous contrastive learning frameworks and uses cosine similarity to score the causal relationship between one premise and one hypothesis. CONCR achieves 77.58% accuracy on BERT-base-uncased and 78.75% on RoBERTa-base, improving previous work by 2.40% and 4.08% respectively. Our code is available at: https://github.com/sherryzyh/CONCR.git.

1 Introduction

Causal Reasoning is defined as the process of identifying causality. For various Natural Language Processing (NLP) applications, identifying causality in an important topic. Even though recent causal reasoning models are able to achieved good performance on certain datasets, they are still not comparable to human performance and cannot achieve stable performance across different datasets. The problem behind it is that most current causal reasoning models are trained to learn the empirical causal patterns and predict the labels on the test datasets, while in reality, the deep and conceptual understanding of the causality is needed.

Recently, Du et al. (2022) proposed a new benchmark called e-CARE, which is used to explore the explainable causal reasoning. Specifically, they claim that conceptual explanations of the causality can be very helpful. The reasons are, the explanations can not only help to check whether the relationship between cause and effect has been understood correctly, but also be used to support the causal reasoning process. The explainable CAusal REasoning dataset (e-CARE) dataset contains over 21K multiple-choice causal reasoning questions, together with natural language formed explanations for each causal question to explain why the causation exists. e-CARE is also the largest human-annotated commonsense causal reasoning dataset.

We present CONCR, a CONtrastive learning framework for Causal Reasnoing, which leads to a great jump in performance on causal reasoning task. CONCR is a model-agnostic framework for contrastively learn the causality-embedded representation. It encodes the sentence separately unlike the previous two-sentence encoding in e-CARE. With the sentence representation, CONCR discards the projection which is widely used in the contrastive learning but use a simple cosine simlarity scorer to calculate the causal score between given premise-hypothesis pair. In the training, the positive samples are constructed by pairing the premise with its correct hypothesis and the negative samples are contructed by pairing premise with any other hypothesis within the same mini-batch. A contrastive cross-entropy learning objective is used to enforce the model to learn the causality-embedded representation.

To better understand each component's contribution to CONCR, we conduct ablation studies on each components and run comprehensive fine-tuning experiments on 6 pre-trained language models.

To conclude, our main contributions include:

1. We propose a new contrastive learning framework targeting at causal reasoning task, call CONCR. It reaches the SOTA performance on the e-CARE benchmark causal reasoning task. Fine-tuned with CONCR, the BERT based and RoBERTa based model can achieve an accuracy of 77.58% and 78.75%, with the increase of 2.40% and 4.08% from previous work, respectively.

*Everyone Contributed Equally

2 Related Work

2.1 Causal Reasoning

Benchmarks There are several causal reasoning in NLP benchmarks. COPA (Roemmele et al., 2011) is a widely used benchmark for causal inference. Causal inference task requests the system to determine either the cause or effect of a given premise from two candidate answers. Huang et al. (2019) proposed a new benchmark called CosmosQA where machine reading comprehension is done with contextual commonsense reasoning. Recently, Du et al. (2022) introduced a new benchmark called e-CARE, where explanation is provided in addition to the premise and two candidate hypotheses. They proposed a new explanation generation task and concluded that explanation of the causal relationship help the causal reasoning.

Causal Reasoning in NLP Many researchers work on causal reasoning in NLP field. Sharp et al. (2016) firstly train causal embeddings in question answering when there exists a causal relationship. Hassanzadeh et al. (2019) studied the problem of answering questions about relationship between two general phrases without any constraints. Xie and Mu (2019) focused on building up distributed representation of words in cause and effect spaces. They proposed three different causal word embedding models to map the labels of sentence level cause-effect pairs to word level.

As large language models become more popular, more work that take advantage of the text representation capabilities with the help of pre-trained language models to do causal reasoning task is gradually emerging. Wu et al. (2021) show that it is beneficial to encourage the student to imitate the causal dynamics of the teacher through a distillation interchange intervention training objective. Zhang et al. (2022) work on commonsense causality reasoning which aims at identifying plausible causes and effects in natural language descriptions.

Explainable Causal Reasoning Structured data like large-scale knowledge graph, event graph, or causal relationship corpus are good source for model to learn about causal relationship. Li et al. (2021b) proposed CausalBERT which can generate multiple semantically distinct possible causes or effects. ExCAR(Du et al., 2021) acquires additional evidence information from a large-scale causal event graph as logical rules and aggregate the chain-level information to predict the causal strength.

Compared to the structured database, unstructured data is more common and easily accessible. CausalNet (Luo et al., 2016) is proposed to automatically capture a network of causal relationships between cause-effect terms from the unstructured web corpus. Li et al. (2021a) investigate the problem of injecting the causal knowledge into pre-trained language models to improve the system's causal reasoning ability.

2.2 Contrastive Learning

Self-supervised learning reached huge success in the recent years. MoCo (He et al., 2020) and SimCLR (Chen et al., 2020), pioneering this area, largely minimized the gap in performance between self-supervised learning and fully-supervised methods in Computer Vision. Several works have applied this learning paradigm to Natural Language Processing as well.

Sennrich (2016) and Pan et al. (2021) applied contrastive learning to cross-language representation learning. In natural language understanding, Gunel et al. (2020) proposed an auxiliary task where supervised contrastive learning loss is used to improve the pre-training language model. Later, SimCSE (Gao et al., 2021) takes an input sentence and predicts itself in a contrastive objective, reaching a better sentence-level representation.

3 Problem Statement

We work on the e-CARE benchmark which explores the explainable causal reasoning. e-CARE includes two tasks, i.e., Causal Reasoning and Explanation Generation.

Causal Reasoning Task Given one premise, denoted as P, and two hypotheses candidates, denoted as H_0 and H_1, this task is formulated as a two-stage task: Firstly, the model takes premise and one hypothesis as the input, and predict its causal score. With these two scores S_0 and S_1, the predictor select the hypothesis with a higher causal score as the output.

Explanation Generation Task Given one premise P and the correct hypothesis H, this task is asking the model to take P and H as the input and generate a free-text-formed explanation E for this cause-effect pair.

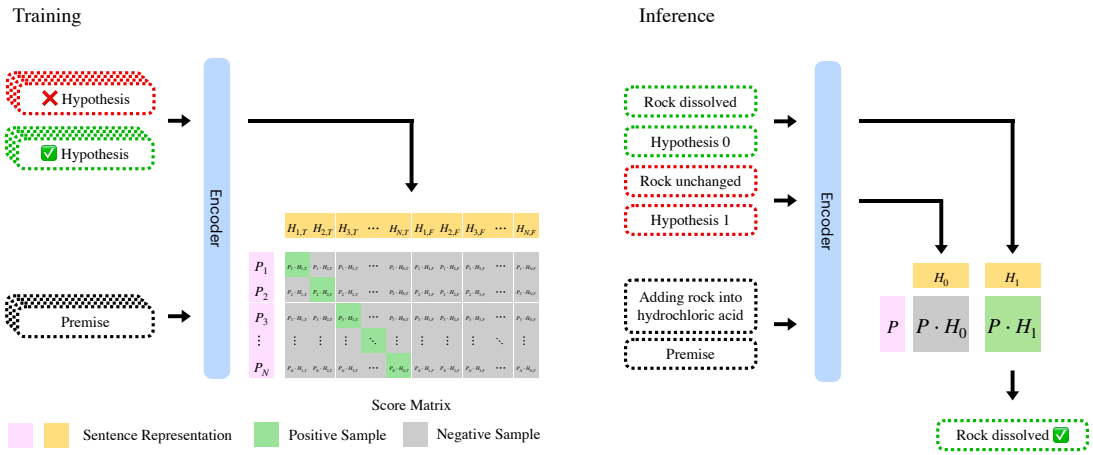

Figure 1: A CONtrastive learning framework for Causal Reasoning (CONCR). In the CONCR framework, an encoder is trained to learn a causality-embedded sentence representation. The causal score between a given premise and a hypothesis is calculated by cosine similarity, which is used to measure the possibility of one premise being causally paired with one hypothesis. Here we use the $P \cdot H$ to denote tye causal score between a Premise-Hypothesis pair. (left) In the training, CONCR treats the premise as the anchor, the true hypothesis of the given premise as positive sample, and hypotheses of other in-batch premises as negative samples. Green blocks are scores of positive samples and grey blocks are scores of negative samples, including hard negative ones. (right) In the inference, CONCR calculates the score between premise and two candidate hypotheses and return the one with higher score as the prediction.

4 CONCR: A Contrastive Learning Framework for Causal Reasoning

Prior work (Sennrich, 2016; He et al., 2020; Chen et al., 2020; Gao et al., 2021; Pan et al., 2021) has widely applied contrastive learning to representation learning. Inspired by supervised version SimCSE (Gao et al., 2021) where model is trained by predicting whether the relationship between two sentences is entailment, neutral or contradiction, we propose CONCR, a contrastive learning framework for causal reasoning, as illustrated in Figure 1. CONCR learns causality-embedded representations by predicting whether there is a causal relationship between them via a contrastive loss in the latent space.

Positive and Negative Samples CONCR takes a batch of premises $\{P_i\}_{i=1}^{N}$ and hypotheses $\{H_{i,\mathrm{T}}, H_{i,\mathrm{F}}\}_{i=1}^{N}$ where $H_{i,\mathrm{T}}$ and $H_{i,\mathrm{F}}$ are the true and false hypothesis of P_i respectively as input. A key component in the contrastive learning framework is the definition of positive and negative samples. Premise as anchor, we use the true hypothesis of this given premise as the positive, all other hypotheses from the same mini-batch as negative samples. With a batch size of N, we have N positive pairs and $2N^2 - N$ negative pairs.

Cosine Similarity Causal Scorer Unlike prior work focusing on learning the semantic representation of sentences, CONCR focuses on learning causality-embedded sentence representation. We propose to simply apply a cosine similarity scorer which takes the encoded sentence representation directly as the input and discarding the projection. Though representative work in contrastive learning like SimCLR (Chen et al., 2020) and SimCSE (Gao et al., 2021) widely use a projection layer after the encoder and calculate the similarity of projected sentence representation, our simple cosine similarity causal scorer without projection helps model learn the causality-embedded representation. Detailed comparison and discussion is presented in Section 5.5.

No Hard Negative Sample CONCR discards the hard negative sample design. In other words, given one premise, apart from the only positive sample which is the paired correct hypothesis, we treat all other $2N - 1$ hypothesis in the $N-$sized batch equally as negative samples. We also compare No-Hard-Negative design with Hard-Negative-Working design. Detailed results and discussion are in Section 5.6.

Contrastive Causal Learning Objective CONCR follows the contrastive framework in

Chen et al. (2020); Gao et al. (2021) and take a cross-entropy objective with in-batch negatives (Chen et al., 2017) with a replacement of representation similarity with causal score: let $\mathbf{p}_i$ denotes the causality-embedded representation of premise i and the $\mathbf{h}_i^+$ denotes the representation of the correct hypothesis of premise i, working as the positive sample. The representation of the false hypothesis of premise i, denoted as $\mathbf{h}_i^-$, and the representation of hypotheses of other premises, denoted as $\mathbf{h}_j^{+/-}, j \neq i$, are all the negative samples. The training objective is defined as:

$$L = -\log \frac{e^{\text{cs}(\mathbf{p}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} \left(e^{\text{cs}(\mathbf{p}_i, \mathbf{h}_j^+)/\tau} + e^{\text{cs}(\mathbf{p}_i, \mathbf{h}_j^-)/\tau} \right)} \quad (1)$$

where τ is the temperature and we set it to a constant 0.05 following SimCSE. $\text{cs}(\mathbf{p}, \mathbf{h})$ represents the causal score between one premise and one hypothesis, calculated by the cosine similarity scorer.

Causal Reasoning Inference In the inference, CONCR's encoder extract the causality-embedded representation of the premise and two candidate hypotheses. Scorer calculates score for each pair and the hypothesis with a higher paired causal scorer is predicted as the correct hypothesis given this premise.

5 Experiment

In this section, we first introduce the experiment setup (Section 5.1) and evaluation metric (Section 5.2). Then we present comprehensive experiments on CONCR framework as well as each component. The main results of CONCR is shown and discussed in Section 5.3. We include an ablation study on CONCR framework in Section 5.4. Experiments on each components are discussed in Section 5.5, 5.6, and 5.7.

5.1 Experimental Setup

Causal Reasoning CONCR is a contrastive learning framework which can adopt any pre-trained language model as the encoder. We experiment on different pre-trained language models and fine-tune the encoder with CONCR's constrastive learning objective. For all experiments, we use the batchsize of 64, encoder's learning rate of $1e-5$, and run 50 epochs. We select the best model based on the loss on the dev set.

For all detailed experiments (Section 5.4 $\sim$ Section 5.6), we use the BERT-base-cased model as our encoder.

5.2 Evaluation Metric

Causal Reasoning We simply use **Accuracy** to measure the model's performance on causal reasoning task.

5.3 Main Results

The main results of CONCR are shown in Table 1. On top of any pre-trained language model, CONCR performs stablely better than current state-of-the-art e-CARE.

5.4 Ablation Study on Framework Design

In previous discussion, we don't highlight the difference between the input format of CONCR and that of e-CARE. CONCR follows a siamese encoder architecture in which the same encoder generates the single-sentence representation for premise and hypothesis. The scorer then calculates the causal score based on the two representations of a premise-hypothesis pair. On the contrary, e-CARE concatenates the premise and hypothesis into one sentence, separated by the `[SEP]` special token and denoted by different token type ids. Only one representation is returned by the encoder with a two-sentence input; a linear layer, taking one representation vector as the input, is used to predict the causal score.

To figure out what makes CONCR successful, we design another experiment on the siamese encoder architecture. We name it Siamese-CR. In the Siamese-CR model, following the CONCR design, we select cosine similarity as the causal scorer. Meanwhile, the learning objective of the Siamese-CR model is binary cross entropy, following e-CARE. Based on the result comparison in Table 2, we conclude that the success of CONCR is not merely coming from the siamese architecture. The CONCR architectural design and learning objective design together contribute to its success.

5.5 Scorer - Simple is the Best

CONCR works without a projection. Specifically, cosine similarity as the causal scorer is calculated on representations from the encoder. We compare our design with three other methods: Projection + Cosine Similarity (the classic contrastive learning framework), Dot Product Scorer, and MLP Scorer. The classic contrastive learning framework

Language Model	Accuracy (%)	
	e-CARE (Du et al., 2022)	CONCR
BERT-base-uncased	77.25†(✗)	**78.52**
BERT-base-uncased	75.18†	**77.58**
RoBERTa-base	74.67†	**78.75**
XLNet-base-cased	73.73†	**77.49**
sup-SimCSE-BERT-base-uncased	(✗)	**78.71**
sup-SimCSE-RoBERTa-base	(✗)	**79.27**

†: Our reproduced results. (✗): Not reported in e-CARE paper.

Table 1: Results of CONCR fine-tuned on different pre-trained language model as the sentence encoder.

Model	Siamese†	CL††	Accu. (%)
e-CARE	✗	✗	75.18
Siamese-CR	✓	✗	63.70
CONCR	✓	✓	**77.58**

†: Siamese encoder design.
††: Contrastive Learning Objective.

Table 2: Results of ablation study on architectural design and learning objective design.

includes a projection of representations from the encoder, and the score (e.g. cosine similarity) is calculated on the projections. Following this approach, we carefully align the cause and effect in each pair and use dual-projectors for cause and effect projections separately. That said, different from a semantic representation encoder, CONCR aims to train a causality-embedded representation encoder. We thus try to modify the classic approach. Specifically, we eliminate the projection layer and explore different scorers: cosine similarity, dot product, and multi-layer perceptron. For the MLP scorer, we experiment with different learning rates for the scorer. The results in Table 3 prove the success of the simple cosine similarity (without projection).

Scorer	Accuracy (%)
Cosine Similarity⋆	**77.58**
Projection + Cosine Similarity	76.27
Dot Product Scorer	74.72
MLP Scorer (lr $1e-3$)	65.06
MLP Scorer (lr $1e-1$)	73.21

⋆: CONCR's scorer.

Table 3: Comparison of scorers.

5.6 Hard Negative (HN) Samples or Not?

Previous supervised contrastive learning frameworks like SimCSE (Gao et al., 2021) select the negative-labeled sentence in the same pair as the hard negative (HN) sample while other non-paired sentences serve as normal negative samples. Looking into the nature of the HN samples, we hypothesize that HN samples help because they usually share many tokens with the positive samples. The projection helps extract the most important features from sentences, and including HN samples enforces encoder to learn the slight difference. In CONCR, encoder learns a causality-embedded sentence representation. The false hypothesis, like other non-paired sentences in the same batch, has no causal relationship with the given premise.

We compare No-HN-Sample design to HN-Sample design with different HN weights. SimCSE (Gao et al., 2021) concludes that HN weight of 1.0 pushes the encoder to extract the best representations. The learning objective including the hard negative weight could be written as:

$$L = -\log \frac{e^{\text{cs}(\mathbf{p}_i,\mathbf{h}_i^+)/\tau} + \lambda \cdot e^{\text{cs}(\mathbf{p}_i,\mathbf{h}_i^-)/\tau}}{\sum_{j=1}^{N}\left(e^{\text{cs}(\mathbf{p}_i,\mathbf{h}_j^+)/\tau} + e^{\text{cs}(\mathbf{p}_i,\mathbf{h}_j^-)/\tau}\right)} \quad (2)$$

We hypothesize that, adding the score of HN samples into the score of positive samples effectively makes it easier for the model to predict the correct class during training, thus mitigating the learning. Therefore, we experiment with three HN weights: $+1$, the best weight from SimCSE; $+0.1$, a smaller pushing of HN samples; and -1 which is used to verify our hypothesis. The results are shown in Table 4.

We conclude from the results that, in the causal reasoning task, a negative HN weight does better than a positive one. In fact, it is better to not differ-

Hard Negative Weight	Accuracy (%)
No HN Samples*	**77.58**
$+1$	77.44
$+0.1$	76.50
-1	77.53

*: CONCR's design.

Table 4: Comparison of Hard-Negative designs.

entiate HN samples from normal negative samples in the causal reasoning setting.

5.7 A Better Sentence Encoder Matters

One of the important observations in our baseline analysis (included in Appendix B) is that some hypotheses are highly similar in sentence representations. In these cases, the encoder can hardly distinguish the two hypotheses, leading to similar causal scores. We hypothesize that, with more distinguishable sentence representations, the performance will improve. Here we include results of different sentence encoders as well as their sentence embedding performance on STS tasks (reported in SimCSE (Gao et al., 2021)), shown in Table 5.

Model†	Avg. STS $\uparrow$	Accu. (%) *
BERT	66.28	77.58
RoBERTa	61.73	78.75
(S)BERT††	81.57	78.71
(S)RoBERTa††	82.52	79.27

*: Fine-tuned on e-CARE dataset with CONCR.
†: BERT refers to the BERT-base-uncased model and RoBERTa refers to the RoBERTa-base model.
††: (S) Represents the model pre-trained in supervised setting with SimCSE (Gao et al., 2021).

Table 5: Comparison between sentence embedding performance and causal reasoning performance of different pre-trained language model.

It is clear that the pre-trained models which generate better sentence embeddings (as measured by STS task performance) can reach higher accuracy, after fine-tuning, in the causal reasoning task.

6 Conclusion and Future Work

CONCR, the contrastive learning framework for causal reasoning, is a big success and can be applied to any other dataset. One of our future work is to evaluate this framework on other causal reasoning tasks like COPA (Roemmele et al., 2011).

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. 2017. On sampling strategies for neural network-based collaborative filtering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 767–776, New York, NY, USA. Association for Computing Machinery.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2021. Excar: Event graph knowledge enhanced explainable causal reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2354–2363.

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.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

Oktie Hassanzadeh, Debarun Bhattacharjya, Mark Feblowitz, Kavitha Srinivas, Michael Perrone, Shirin Sohrabi, and Michael Katz. 2019. Answering binary causal questions through large-scale text mining: An evaluation using cause-effect pairs from human experts. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5003–5009. International Joint Conferences on Artificial Intelligence Organization.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *arXiv preprint arXiv:1909.00277*.

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

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*.

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*.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. *arXiv preprint arXiv:2105.09501*.

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.

Rico Sennrich. 2016. How grammatical is character-level neural machine translation? assessing mt quality with contrastive translation pairs. *arXiv preprint arXiv:1612.04629*.

Rebecca Sharp, Mihai Surdeanu, Peter Jansen, Peter Clark, and Michael Hammond. 2016. Creating causal embeddings for question answering with minimal supervision. *arXiv preprint arXiv:1609.08097*.

Zhengxuan Wu, Atticus Geiger, Josh Rozner, Elisa Kreiss, Hanson Lu, Thomas Icard, Christopher Potts, and Noah D Goodman. 2021. Causal distillation for language models. *arXiv preprint arXiv:2112.02505*.

Zhipeng Xie and Feiteng Mu. 2019. Distributed representation of words in cause and effect spaces. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Jiayao Zhang, Hongming Zhang, Dan Roth, and Weijie J Su. 2022. Causal inference principles for reasoning about commonsense causality. *arXiv preprint arXiv:2202.00436*.

A Causal Explanation Quality (CEQ) Score

Causal Explanation Quality (CEQ) score is proposed by Du et al. (2022), which is used to measure the causation in the generated explanation. This metric is based on the word-pair-based causal strength (cs) score as described in Luo et al. (2016), where the causal strength is a score in [0, 1]. Ideally, the causal strength for a valid causal fact should be equal to 1. Therefore, CEQ score is a metric to measure the increase of causal strength brought by the explanation. The CEQ score is defined as:

$$CEQ = \Delta_{cs} = cs(C, E|X) - cs(C, E)$$

where C, E and X denote the cause, the effect and the generated explanation, $cs(C, E)$ is the original causal strength between C and E, $cs(C, E|X)$ is the causal strength after adding the explanation information. $cs(C, E|X)$ is defined by:

$$cs(C, E|X) = max[cs(C+X, E), cs(C, E+X)]$$

where "+" represents for the string concatenate operation.

The calculation of causal strength is proposed by Luo et al. (2016), which is defined as:

$$cs(C_A, E_B) = \frac{\Sigma_{w_i \in C_A, w_j \in E_B} cs(w_i, w_j)}{N_{C_A} + N_{E_B}} \quad (3)$$

where (C_A, E_B) is an arbitrary causal fact, N_{C_A} and N_{E_B} are the number of words within C_A and E_B respectively, $cs(w_i, w_j)$ is the causal strength between word w_i and w_j, which is estimated from a large corpus as:

$$cs(w_i, w_j) = \frac{Count(w_i, w_j)}{Count(w_i)Count(w_j)^{\alpha}}$$

where α is a penalty coefficient and Luo et al. (2016) empirically set $\alpha = 0.66$.

B Baseline Analysis

B.1 Sources of Errors in Causal Reasoning

We first consider causal reasoning errors made by both GPT2$_{CR-EG}$ and GPT2$_{CR}$. We hypothesize that not all examples in the dataset are equally difficult, and one determinant might be how similar the two given hypotheses are. To measure the similarity between two hypotheses, we use a sentence transformer *all-MiniLM-L12-v2*[1] from HuggingFace to obtain the embeddings of each hypothesis, and compute the cosine similarities between them. We visualize the distribution of hypotheses' cosine similarity and accuracy in Figure 3. The result aligns with our expectation, that the higher the hypotheses' cosine similarity is, the more likely the model makes mistakes. That said, we recognize that the vast majority of errors lie in examples where the cosine similarity between two hypotheses is not very high (lower than 0.6). Therefore, we look at both error cases that have highly similar hypotheses and error cases that have lower hypotheses' similarities.

For cases where two hypotheses have a cosine similarity above 0.9, usually only one word is different and that word makes the two hypotheses have opposite meanings. We hypothesize that given the extremely high similarity in the hypotheses' sentence embeddings, it is very difficult for the model to distinguish the two sentences from each other. We propose to use prompt-based sentence representation to improve the quality of sentence embeddings, so that hypotheses which differ in few important words look more different to the model. Here we include a few example error cases and make the word that sets hypotheses apart bold:

Premise: He wasn't thrown from the bus in the car accident.
Ask-for: Cause
Correct hypothesis: He **fastened** seatbelts when getting on the bus.
Wrong hypothesis: He **unlocked** seatbelts when getting on the bus.

Premise: They easily damage their spines.
Ask-for: Cause
Correct hypothesis: Dachshunds have **long** backs.
Wrong hypothesis: Dachshunds have **short** backs.

Premise: John explored the altitude variation of swamp.
Ask-for: Effect
Correct hypothesis: He found very **low** topographic relief.
Wrong hypothesis: He found very **high** topographic relief.

For error cases that have hypotheses' cosine

[1] https://huggingface.co/microsoft/MiniLM-L12-H384-uncased

	avg_bleu	rouge1	rouge2	rougel	avg_rouge	cos_sim	correct	ask_for_cause	hypo_sim
avg_bleu	1.000000	0.815827	0.665500	0.812003	0.812772	0.671610	0.019955	-0.017886	-0.001544
rouge1	0.815827	1.000000	0.760931	0.992075	0.979106	0.609272	0.018344	-0.005131	0.004634
rouge2	0.665500	0.760931	1.000000	0.771091	0.873485	0.409612	0.027723	0.000762	0.001015
rougel	0.812003	0.992075	0.771091	1.000000	0.982006	0.605090	0.019803	-0.008222	0.001493
avg_rouge	0.812772	0.979106	0.873485	0.982006	1.000000	0.581092	0.022566	-0.004849	0.002637
cos_sim	0.671610	0.609272	0.409612	0.605090	0.581092	1.000000	0.039139	-0.002814	0.026346
correct	0.019955	0.018344	0.027723	0.019803	0.022566	0.039139	1.000000	0.026074	-0.177581
ask_for_cause	-0.017886	-0.005131	0.000762	-0.008222	-0.004849	-0.002814	0.026074	1.000000	-0.031006
hypo_sim	-0.001544	0.004634	0.001015	0.001493	0.002637	0.026346	-0.177581	-0.031006	1.000000

Figure 2: Correlation matrix of the GPT2$_{CR-EG}$ performance metrics and task metrics. avg_rouge is the average of rouge1, rouge2, and rougel scores; cos_sim is the cosine similarity between the generated explanation embedding and the provided explanation embedding; hypo_sim is the cosine similarity between the embeddings of the two hypotheses, as introduced in Section B.1.

similarity lower than 0.6, the two hypotheses are noticeably more different, but the model often fails due to the lack of commonsense knowledge or domain knowledge. In the examples below, the model fails to recognize that anthropology, not nanotechnology, studies human species; that gamma, not feet or foot, is the unit for measuring magnetism; that baptism is an identification with the community of believers. We plan to look into the possibility of leveraging an external knowledge base to equip the model with such knowledge. It is possible that even the aforementioned cases associated with highly similar hypotheses can benefit from a knowledge base, which may inform the model that swamps have low topographic relief and spines in long backs damage more easily.

Premise: Jack's interest is to study human species.
Ask-for: Effect
Correct hypothesis: He decides to choose anthropology as his major in college.
Wrong hypothesis: Jack's company mainly uses nanotechnology.

Premise: Tom has measured the magnetism of the material.
Ask-for: Effect
Correct hypothesis: It is 15 gammas.
Wrong hypothesis: The result reached nearly four feet.

Premise: Jerry was baptized.
Ask-for: Effect
Correct hypothesis: He was identified by the believers.
Wrong hypothesis: The public saw baptism.

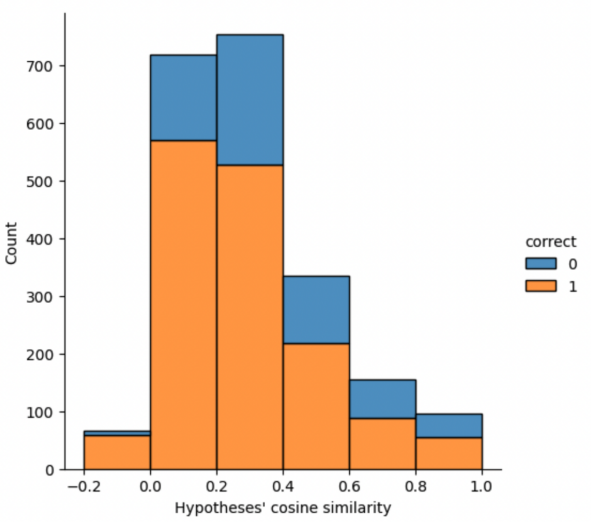

Figure 3: Distribution of cosine similarity between two hypotheses and accuracy

B.2 Role of Explanation Generation in Causal Reasoning

Although GPT2$_{CR-EG}$ has a slightly higher accuracy in the causal reasoning task than that of GPT2$_{CR}$, many questions remain in how the explanation generation task contributes to the causal reasoning task in GPT2$_{CR-EG}$, and by generating explanations, whether the model is reasoning about causation or correlation only. We examine the explanations generated by GPT2$_{CR-EG}$ when it succeeds in the causal reasoning task, while GPT2$_{CR}$ does not. In the first example below, the generated explanation is contrary to reality. It is possible that the model uses the correlation between "donate" and "help" to choose the correct hypothesis, without understanding the actual mechanism that donation helps relieve, rather than cause, distress. In the second example below, the

generated explanation is completely irrelevant. It is possible that the model leverages the correlation between "solve the problem" and "understanding" to choose the correct hypothesis, but the model does not seem to understand that the act of problem solving contributes to cognition.

Premise: This charity donates money to Asian areas with famine.
Ask-for: Effect
Correct hypothesis: Some people western india were helped.
Wrong hypothesis: Some people western india were hungry.
Explanation generated by GPT2$_{CR-EG}$: Donations help cause famine in Asian countries.

Premise: He successfully solved the problem.
Ask-for: Effect
Correct hypothesis: He has a clear understanding of this problem.
Wrong hypothesis: He cannot solve the problem from start to finish.
Explanation generated by GPT2$_{CR-EG}$: Philosophy is the study of the mind.

We also investigate the explanations generated by GPT2$_{CR-EG}$ on examples where it fails in the causal reasoning task, to better understand the failures. For the first error case mentioned in Section B.1, the model generates "Seatbelts provide protection." as the explanation. While this explanation seems very reasonable at first glance, looking at it side by side with the two hypotheses, we find the explanation too general. The model may have found some correlation between seatbelts and protection, but the failure to reach a concrete understanding of how seatbelts provide protection (whether "fastening seatbelts" helps or "unlocking seatbelts" helps) shows the model's lack of ability to perform causal reasoning. Therefore, it is not surprising that adding the explanation generation task does not contribute to the causal reasoning task in cases like this one.

B.3 Limitations in Current Metrics and Points of Comparison

On the metrics used to evaluate the explanation generation task, Du et al. (2022) points out that average-BLEU and ROUGE-1 hardly have any correlation with the results of human evaluation. In the same vein, we find that there is barely any correlation between the various BLEU/ ROUGE metrics and whether GPT2$_{CR-EG}$ succeeds in the causal reasoning task, as shown in Figure 2.

As another attempt to measure the similarity between the generated explanation and the provided explanation, we use a sentence transformer *all-MiniLM-L12-v2*[2] from HuggingFace to obtain the embeddings of generated explanations and provided explanations, and compute the cosine similarities between them. We denote this metric as cos_sim in Figure 2. Compared to the BLEU and ROUGE metrics, it has higher correlation with whether GPT2$_{CR-EG}$ correctly chooses the hypothesis, but this correlation is not meaningful, either. We hypothesize that if there were ten explanations provided for each example, and we measure the average similarity between the generated explanation and each of the provided explanation, the similarity metrics like BLEU, ROUGE, and cos_sim may be meaningful. However, given that only one explanation can be compared with for each example, any similarity score is likely not a good indicator of the quality of the generated explanation. Consequently, we plan to rely more on metrics that measure causation directly, such as the CEQ score, in future experiments.

On the causal reasoning task, Du et al. (2022) states the 92% human evaluation accuracy as the ceiling. However, looking carefully at the dataset, we question whether this is a valid point of comparison, if we ask the model to directly reason about the causality in each cause-effect pair and then choose the one with the higher score. This process is very different from the human evaluation process, which treats each example as a multiple-choice question, and the human annotator can compare the two hypotheses side by side, instead of giving them two scores at different times. For examples like the one below, if the human annotator is not presented the two hypotheses together, they are likely to make more mistakes, because they could not focus on the key phrases that set the two hypotheses apart.

Premise: She got tired of announcing similar result involving prizes.
Ask-for: Cause
Correct hypothesis: The hostess hosted so many different contests.

[2] https://huggingface.co/microsoft/MiniLM-L12-H384-uncased

Wrong hypothesis: The hostess joined in so many host contests.

While it may be easier to improve the accuracy of the model by presenting it with two hypotheses together, like in a multiple-choice question, we intentionally avoid this strategy because we aim to improve the model's causal reasoning ability that is more comparable to and useful in human's everyday life. Humans do not face multiple-choice causal reasoning questions apart from in exams; instead, we think about how likely one thing may lead to the other thing.

C KER Failure Case Analysis

Here we include a few examples on which the e-CARE model with vanilla sentence embeddings (BERT-base-cased baseline) succeeds and the e-CARE model with KER fails, with injected knowledge in parentheses:

Premise with knowledge: Various features show that this is a dicot (is a angiosperm) in the family Piperaceae.
Ask-for: Effect
Correct hypothesis with knowledge: Researchers suspected it is a peperomia (is a herb).
Wrong hypothesis with knowledge: This seed is a monocot (is a angiosperm) rather than dicot (is a angiosperm).

Premise with knowledge: The scientist observed its appearance.
Ask-for: Cause
Correct hypothesis with knowledge: Nobody has seen edaphosauruses (is a synapsid) before.
Wrong hypothesis with knowledge: A scientist has picked up a paramecium (is a ciliate) and curious about it.

In the first example, the same knowledge is added for "dicot" in the premise and the wrong hypothesis, as well as for "monocot" in the wrong hypothesis. Adding knowledge increases the similarity between the wrong hypothesis and the premise, likely misleading the model to conclude a stronger causal relation between them.

In the second example, the added knowledge for "edaphosauruses" and "paramecium" is scientific terminology, which may still be difficult for the model to understand. It is possible that the model find it more difficult to reason about the sentences when the additional scientific terminology is added.

These examples illustrate that the idea of using domain knowledge to help with causal reasoning is intuitive but not very easy to realize. The injected knowledge may not be relevant in a way that helps the model understand the original sentences, and instead confuse the model.