CALM-Bench: A Multi-task Benchmark for Evaluating Causality Aware Language Models

Dhairya Dalal † and **Mihael Arcan**‡ and **Paul Buitelaar**† ‡
†SFI Centre for Research and Training in Artificial Intelligence
† ‡Insight SFI Research Centre for Data Analytics
Data Science Institute, University of Galway
d.dalal1@nuigalway.ie,
{mihael.arcan,paul.buitelaar}@universityofgalway.ie

Abstract

Causal reasoning is a critical component of human cognition and is required across a range of question-answering (QA) tasks (such as abductive reasoning, commonsense QA, and procedural reasoning). Research on causal QA has been underdefined, task-specific, and limited in complexity. Recent advances in foundation language models (such as BERT, ERNIE, and T5) have shown the efficacy of pre-trained models across diverse QA tasks. However, there is limited research exploring the causal reasoning capabilities of those language models and no standard evaluation benchmark. To unify causal QA research, we propose *CALM-Bench*, a multi-task benchmark for evaluating causality-aware language models (CALM). We present a standardized definition of causal QA tasks and show empirically that causal reasoning can be generalized and transferred across different QA tasks. Additionally, we share a strong multi-task baseline model which outperforms single-task fine-tuned models on the *CALM-Bench* tasks.

1 Introduction

Causal reasoning is a crucial aspect of human cognition and is critical to the development of our mental models of reality (Neeleman et al., 2012; Johnson-Laird and Khemlani, 2017; Griffiths, 2017). Theories of causation have been studied extensively across philosophy (Beebee et al., 2009), physics (Dowe, 2009), cognitive science (Waldmann, 2017), and probability and statistics (Pearl, 2009), amongst many other fields. Explorations of causality in the language domain tend to be semantic, linguistic, or logical in nature as access to direct observational data or event probabilities is not assumed nor is required. Descriptions of causality can be linguistically valid but factually incorrect (e.g. butter is the leading cause of factory deaths). Therefore, causal reasoning in language should ideally be logically consistent and grounded in commonsense knowledge. The *counterfactual theory of causation* (Lewis, 1973) provides a useful definition of causation for language applications. It posits that causation is relational (there is a cause and effect), temporal (the cause must precede the effect), and counterfactual (if the causing event had not occurred, the effect would not have occurred). Various natural language processing (NLP) applications require identifying causal relations and reasoning over those relations.

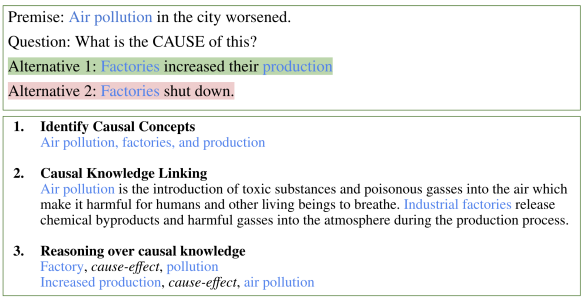

Figure 1: An CQA example from the COPA (Gordon et al., 2012). CQA requires identifying causal concepts, linking those concepts to causal relations, and reasoning over those relations.

These NLP applications can be split into two general categories: causal relation identification (CRI) and causal question-answering (CQA). CRI tasks aim to identify and extract cause/effect spans from descriptions of causal events. CRI requires linguistic knowledge - relying on lexical triggers (i.e. causative verbs and causal connectives) and grammatical structures (Neeleman et al., 2012; Girju, 2003). Historically, the majority of NLP research on causality has focused on CRI.

In contrast to CRI, CQA tasks require both background causal knowledge and reasoning. Consider the question *Air pollution in the city worsened. What is the cause of this?* (Figure 1). To answer this question, commonsense knowledge about factories, pollution, and the ability to infer both causal and counterfactual relations is required. General

Findings of the Association for Computational Linguistics: EACL 2023, pages 296–311
May 2-6, 2023

Task	Example	Size	Question Type	Format	Knowledge
aNLI (Bhagavatula et al., 2020)	**Context:** Jessie wants to save the planet. This summer has been the hottest in all history. **Question: Which hypothesis is the most plausible for the provided observations?** A: Jessie decides to buy a new truck. **B: Jessie decides to sell her truck and use public transportation instead.**	174,226 Train: 169,654 Val: 1,532 Test: 3,040	cause prediction	multiple-choice	social, world
COPA (Gordon et al., 2012)	**Question: Air pollution in the city worsened. What is the cause of this?** **A: Factories increased their production.** B: Factories shut down.	1,000 Train: 800* Val: 200 Test: 500	cause prediction effect prediction	multiple-choice	world
CosmosQA (Huang et al., 2019)	**Context:** Two things happened today in Beijing. First off, incoming journalists were amazed to find China had successfully lifted the brown haze in city. Skies were crystal blue and the air felt noticeably lighter. **Question: Why did the sky appear clearer?** A: None of the above choices. B: The citizens learned to ignore the gloomy skies. **C: The citizens made an effort to cut down on pollution.** D: A large storm had recently passed.	35,210 Train: 25,262 Val: 2,985 Test: 6,963	cause prediction effect prediction	multiple-choice	social, world
E-Care (Du et al., 2022)	**Question: The city is determined to control air pollution. What is the effect?** **A: They have to reduce the number of automobiles.** B: Environmental pollution has been increased.	17,051 Train: 14,929 Val: 2,122 Test: blind	cause prediction effect prediction	multiple-choice	social, world, science
ROPES (Lin et al., 2019)	**Context:** There are two planets, Glarnak and Bornak, that share the same atmospheric composition. The planets have nearly identical ecosystems and topography. The main difference between the two planets is the level of global warming on each planet. Glarnak is experiencing a strong impact from global warming. Bornak, though, is experiencing practically no effects of global warming. **Question: Which planet has more pollutants in the atmosphere? Glarnak**	14,322 Train: 10,924 Val: 1,688 Test: 1,710	cause prediction cause comparison effect prediction effect comparison	reading comprehension	science, world
WIQA (Tandon et al., 2019)	**Context:** 1. A seed is in soil. 2: The seed germinates. 3: The plant grows roots. 4: The plant grows out of the ground. 5: The plant gets bigger. 6: The plant flowers. 7: The flower produces fruit. 8: The fruit releases seeds. 9: The plant dies. **Question: Suppose less pollution in the environment happens, how will it affect the population of plants? A: More** B: Less C: No Effect	39,705 Train: 29,808 Val: 6,894 Test: 3,003	effect prediction	multiple-choice	science, world

Table 1: CALM-Bench is a mutli-task causal question answering benchmark consisting of six diverse QA tasks requiring both causal reasoning and knowledge.

work on CQA is often under-defined and limited based on the task definition. For example, previous work defined CQA as answering variations of *What is the cause/effect of X?* style questions where the model had to select the most plausible cause or effect from a set of candidate options. While this task requires causal knowledge, it could be recast as an information retrieval problem with no further requirement of causal reasoning. A stronger definition of CQA would allow for more principled explorations of causal reasoning (e.g. reasoning over causal chains, abductive inference, counterfactual reasoning, etc) and aid in the development of stronger NLP models.

Recent advances in foundation language models have demonstrated the effectiveness of pre-trained models across a wide range of NLP and general language understanding tasks. The term *foundation model* (Bommasani et al., 2021) describes any monolithic neural model (e.g. BERT (Devlin et al., 2019)) that captures general knowledge through pre-training and is able to transfer that knowledge to a wide range of downstream tasks. Foundation language models exhibit general reasoning capabilities (Clark et al., 2021), factual knowledge recall (Petroni et al., 2019), and superior performance on a wide range of QA tasks (Khashabi et al., 2020; He et al., 2021; Lourie et al., 2021a). Knowledge in foundation language models is usually injected through denoising objectives (e.g. masked token prediction) (Sun et al., 2020). However, interpreting and extracting that knowledge is difficult (requiring specialized probing tasks) and these models can be susceptible to exploiting superficial (Kavumba et al., 2019). CQA tasks could provide a unique opportunity to develop both explainable models (through producing causal explanation chains) and expand the reasoning capabilities of those models in QA settings. To date, no comprehensive study has explored the causal reasoning capabilities of foundation language models.

We aim to unify research around CQA research by providing a definition for CQA rooted in the cognitive understanding of causal learning and propose *CALM-Bench*, a multi-task causal question-answering benchmark for evaluating causality-aware language models (CALM). *CALM-Bench* (Table 1) consists of six different QA tasks (aNLI (Bhagavatula et al., 2020), COPA (Gordon et al., 2012), CosmosQA (Huang et al., 2019), E-Care (Du et al., 2022), WIQA (Tandon et al., 2019), and ROPES (Lin et al., 2019)) that require both causal knowledge and causal reasoning. We show empirically that causal reasoning can be generalized across the different tasks in *CALM-Bench*. We present a multi-task learning (MTL) setup that outperforms all single-task fine-tuned baselines and demonstrates strong results on the COPA task in a zero-shot setting. Relevant details about the code and model weights can be found on GitHub [1].

2 Causal question-answering

We define CQA broadly as any QA task which requires both *causal reasoning* and *causal knowl-*

[1] https://github.com/dhairyadalal/CALM-Bench

edge provided a real or hypothetical description of events. Cognitive theories of causal learning provide a framework for understanding and evaluating the process of causal question-answering in NLP applications. The inferential theory of causal learning posits that causal learning is a slow and effortful cognitive process that involves drawing causal conclusions over propositional premises (Boddez et al., 2017).

Propositions represent our causal knowledge and contain both qualified relational information (e.g. increase of greenhouse gasses in the atmosphere causes global warming) and propositional beliefs (I believe that greenhouse gasses cause global warming). Propositions are compositional (given the propositions: factories cause air pollution and pollution leads to global warming, we can infer that factories cause global warming) and directional (i.e. we would not infer that global warming causes factories). A key aspect of causal learning is the ability to generalize specific causal knowledge to new situations which is known as causal mechanism knowledge. (Johnson and Ahn, 2017; Ahn et al., 1995).

Causal mechanism knowledge is the mental representation of a system of physical or abstract parts/processes and the expectation of causal interactions between those components that can be generalized to new situations. For example, an arson investigator relies on their mechanism knowledge of fire catalysts and forensic experience to ascertain human involvement. Causal mechanism knowledge can be succinctly represented as propositional statements. Causal bridging inferences describe the relationship between causal knowledge and reasoning. Singer et al. (1992) found that individuals invoke causal statements to bridge two events and then validate those statements against prior commonsense and causal knowledge. For example, given the events *Anna added butter to the hot pan.* and *The butter melted.*, we implicitly invoke the bridging statement *heat caused the butter to melt* based on our prior knowledge.

Solving CQA tasks can be decomposed into three general steps: *causal concept identification*, *causal knowledge linking*, and *causal reasoning*. Consider Figure 1, the causal concepts of air pollution and factories are identified and then linked to background knowledge in order to produce causal knowledge. Causal knowledge can be expressed as relational triples (e.g. factory, cause-effect, pollution) which are effectively propositional statements. The final step requires reasoning over that knowledge through both inferential and counterfactual reasoning. We infer that the increase in factory production results in worsening air pollution based on causal knowledge that factory production causes pollution. The counterfactual, if factories shut down then air pollution would not increase, allows us to eliminate the second option. Arriving at the correct answer in this example is difficult without any background causal knowledge and reasoning over that knowledge.

An important aspect of causal learning is the ability to generalize causal mechanism knowledge to novel situations and task settings. We can see in Table 1 that while thematically all the examples are about the causal relationship between global warming and air pollution, each question requires different types of reasoning over the same knowledge. With the aNLI example, global warming is not mentioned explicitly but must be inferred from social commonsense knowledge (i.e. through the bridging inferences that *saving the planet* and *the hottest summer* are related to global warming) and then use abductive reasoning to select the most plausible hypothesis. The COPA example requires counterfactual reasoning to eliminate the option that factories shutting down would not contribute to air pollution and inferential reasoning to infer that increased factory production results in more air pollution. The WIQA example requires both understanding the life cycle of a plant as a procedural chain and predicting the magnitude impact of environmental pollution as a downstream effect on the plant population. Finally, the ROPES example involves generalizing mechanism knowledge to a fictional setting in order to identify which planet is more likely to have pollutants in the air.

CALM-bench consists of diverse QA tasks requiring social, world, and science knowledge. Our empirical experiments aim to validate the assumption that causal reasoning is transferable across these QA tasks in *CALM-Bench* and produce strong baselines for future research in this space.

3 Related Work

3.1 Causal question-answering

COPA was one of the first QA benchmark tasks which required both background commonsense knowledge and causal reasoning. It is also included as part of the SuperGlue (Wang et al., 2019) bench-

mark. COPA can be considered solved by modern massive foundation models which achieve near human performance (99% accuracy). However, these models are very large (the top three models having more than 10 billion+ parameters), are trained on multi-terabyte scale corpora, and require significant computing resources. Sharp et al. (2016) constructed the first CQA dataset from the Yahoo! Answers corpus using the templates *What causes ...* and *What is the result of ...* to identify causal questions. Sharp et al. (2016) and Xie and Mu (2019) investigated different strategies for training distributed causal embeddings for re-ranking answer options for those causal questions. Hassanzadeh et al. (2019) and Kayesh et al. (2020) explored binary causal questions (i.e. could X cause y) answering using a mixture of co-occurrence statistics and cosine similarity threshold derived from fixed BERT embeddings. The proposed solutions were specific to the task format (i.e. learning threshold values for predicting the yes option). causalqa introduced CausalQA, a corpus of 1.1 million causality-related questions and answers extracted from various datasets primarily related to open-domain web queries (e.g. GooAQ (Khashabi et al., 2021), MS-Marco (Nguyen et al., 2016)). Causal questions were identified using templates spanning *What*, *How*, and *Why* style questions whose intent is to enquire about causes and effects.

Both the CausalQA and the Yahoo Answers! causal questions focus on causal knowledge retrieval or basic reading comprehension without further requirement of causal reasoning. Causal knowledge retrieval can be generalized to information retrieval where the goal is to ensure the retrieved passage contains causal explanations related to the query. Here linguistic cues (Khoo et al., 1998; Girju et al., 2007; Neeleman et al., 2012) or semantic similarity (Dalal et al., 2021b) can be used to identify relevant passages. Likewise, answering *What*, *How*, and *Why* style questions in the context of reading comprehension (e.g. SQuAD (Rajpurkar et al., 2016)) focus more on the lexical overlap between the question and supporting text and linguistic cues associated with the question typologies. CALM-Bench aims to address this gap by focusing QA tasks that require both causal knowledge and causal reasoning.

Most recently, CQA research has investigated augmenting foundation language models with external knowledge for CQA. Dalal et al. (2021a) proposes augmentation with external causal knowledge graph embeddings derived from CauseNet (Heindorf et al., 2020) for QA on the COPA and WIQA tasks and Hosseini et al. (2022) explores injecting the commonsense knowledge from the ATOMIC (Sap et al., 2019) commonsense knowledge base using the BERT masked language modeling pretraining objective for the COPA task. Recent interest in question-answering has led to the development of many large-scale and complex QA tasks. *CALM-bench* consists of curated tasks that require causal reasoning and are described in Section 4.

3.2 Commonsense Reasoning

Commonsense reasoning is closely related to CQA and can be considered a broader superset of CQA depending on the task. Several of the *CALM-Bench* tasks (aNLI, COPA, CosmosQA, and E-CARE) require causal reasoning over commonsense knowledge, and the aNLI, COPA, and CosmosQA tasks were first introduced as commonsense QA tasks. Recent work on commonsense reasoning has focused on probing commonsense knowledge found in foundation language models (Zhou et al., 2020), strategies for effective knowledge augmentation (Fan et al., 2020), and the generation of commonsense knowledge (Bosselut et al., 2019). Lourie et al. (2021b) introduced the first multi-task commonsense QA benchmark (RAINBOW) and a universal model (UNICORN) for general commonsense QA. UNICORN is a T5-11b model (Raffel et al., 2020) trained on the RAINBOW multi-set tasks and fine-tuned in a multi-task setting. Our approach and motivation for multi-task CQA benchmark were greatly inspired by (Lourie et al., 2021b). *CALM-bench* shares two of its tasks (aNLI and CosmosQA) with the RAINBOW benchmark and we consider multi-task learning in our experiments.

3.3 Causal Relation Identification

CRI is often the first step for aggregating causal knowledge when building automated CQA systems (Hassanzadeh et al., 2020). Extracted causal relations are often useful for generating causal knowledge graphs (Heindorf et al., 2020) and developing causal knowledge representations (Sharp et al., 2016; Dalal et al., 2021a) which can be used to improve model performance in CQA tasks. CRI tasks have been studied extensively in the computational linguistics and NLP domain (Yang et al., 2022; Drury et al., 2022). Early methods relied on lexical triggers and linguistic cues (Khoo et al.,

1998; Girju et al., 2007; Neeleman et al., 2012). More recent approaches have explored using neural methods with word embedding features (Dasgupta et al., 2018), self-supervision (Zuo et al., 2021), and external knowledge (Liu et al., 2020). Several efforts have been undertaken to unify CRI research. Tan et al. (2022) introduced the UniCausal benchmark which consolidates six annotated CRI corpora across the tasks of causal sequence classification, cause-effect span detection, and causal pair classification. (Hosseini et al., 2021) introduced the CREST schema and toolkit which converts thirteen commonly used CRI datasets into a unified format.

4 CALM-Bench Tasks

CALM-Bench (Table 1) consists of five multiple-choice tasks (aNLI, COPA, Cosmos QA, E-Care, and WIQA) and a reading comprehension task (ROPES). These tasks require diverse causal knowledge which can be broadly summarized as social (sociological norms of human behavior), world (general commonsense knowledge), and science (specific scientific knowledge of natural processes such as the precipitation cycle or plant life cycle). Questions either require predicting the cause or effect (i.e. cause and effect prediction) provided a description of events or comparing entities (i.e. cause and effect comparison) in a causal system.

Abductive Natural Language Inference (aNLI) (Bhagavatula et al., 2020) is an abductive reasoning task over narratives of social situations. Provided a sequential pair of social observations, the model must predict which of the two provided hypotheses best explains the observations.

Choice of Plausible Alternatives (COPA) (Gordon et al., 2012) is a commonsense causal reasoning task. Provided a premise, the goal is to select the most likely cause or effect from a pair of options. (Kavumba et al., 2019) introduced 500 additional training examples in Balanced-COPA to mitigate the corpus-level artifacts that were likely to be exploited by language models during fine-tuning.

COSMOS QA (Huang et al., 2019) is a multiple-choice QA task requiring social commonsense knowledge. Provided a narrative about people in everyday situations, the goal is to identify the most plausible cause or effect about agents in the story.

E-Care (Du et al., 2022) consists of two causal reasoning tasks. The first task, similar to COPA, requires identifying the most likely cause or effect of the provided premise. The second task requires generating a causal explanation of the correct answer option. We only consider the first task as part of CALM-Bench.

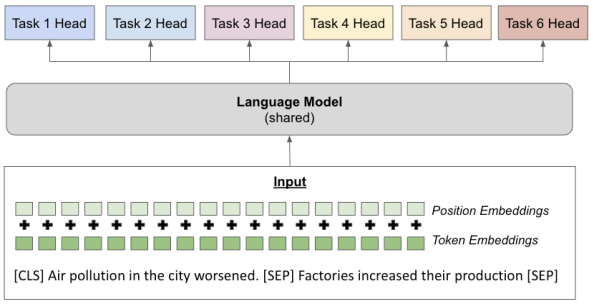

Figure 2: Our MTL model adapts the hard-parameter sharing architecture (Baxter, 2004) where the language model is shared across all the task heads. During training, the task losses are averaged and backpropagated to produce causality-aware contextual embeddings which are effective across all the *CALM-Bench* tasks (Table 4).

Reasoning over Paragraph Effects (ROPES) (Lin et al., 2019) is a reading comprehension task. Provided a knowledge passage, the model is required to reason over the causal and qualitative relations in the passage and apply them to answering questions about a hypothetical situation. 70% of background passages contain causal relations and 26% contain both causal and qualitative relations.

What If question-answering (WIQA) (Tandon et al., 2019) is a multiple-choice QA task requiring reasoning over procedural descriptions of natural processes. WIQA requires predicting the downstream magnitude (more, less, no effect) effect of a perturbation to an individual step in the procedural chain.

5 Methodology

5.1 Language Models

Our experiments consider two different foundation language models, BERT (Devlin et al., 2019) and ERNIE 2.0 (Sun et al., 2020). BERT and derivative models (e.g. RoBERTa (Liu et al., 2019b), DeBERTa (He et al., 2021), etc) contain unspecified distributional knowledge which is learned through the random masked language modeling pretraining objective. In a contrast, ERNIE 2.0 injects external knowledge through a variety of pretraining objectives including masked knowledge prediction, discourse relation prediction, and the IR relevance task. ERNIE 2.0's underlying transformer encoder has the same architecture and parameters as the

BERT Transfer Results						
Trained On ⇓ Evaluated On ⇒	***aNLI***	***COPA***	***CosmosQA***	***E-Care***	***WIQA***	***ROPES***
Single Task FT Baseline	*0.61*	*0.64*	*0.57*	*0.76*	*0.65*	*0.58*
aNLI	-	+0.11	+0.04	0	-0.01	+0.01
COPA	+0.02	-	+0.04	-0.04	0	-0.06
CosmosQA	+0.01	+0.05	-	0	-0.01	+0.02
E-Care	+0.02	+0.13	-0.02	-	-0.02	-0.05
WIQA	0	0	+0.03	-0.02	-	-0.05
ROPES	+0.02	+0.07	+0.03	-0.04	-0.02	-

Table 2: This table contains the transfer learning results for the BERT model. Results are read across the rows where the first column in each row contains the base task selected for transfer learning and the remainder of the columns are the evaluation results across the target tasks. We provide the single-task finetuned baseline in the second row and the pp difference between for each experiment. All results presented are accuracy scores with exception of ROPES which is exact match.

ERNIE 2.0 Transfer Results						
Trained on ⇓ Evaluated On ⇒	***aNLI***	***COPA***	***CosmosQA***	***E-Care***	***WIQA***	***ROPES***
Single Task FT Baseline	*0.64*	*0.71*	*0.63*	*0.76*	*0.64*	*0.53*
aNLI	-	+0.07	0	+0.02	+0.01	+0.08
COPA	0	-	-0.01	+0.01	+0.02	-0.03
CosmosQA	+0.02	+0.01	-	0	+0.02	-0.12
E-Care	0	+0.08	+0.02	-	+0.02	+.11
WIQA	0	+0.01	0	-0.01	-	+0.03
ROPES	+0.02	-0.06	-0.01	+0.01	+0.02	-

Table 3: This table contains the transfer learning results for the ERNIE 2.0 model. In contrast the BERT model, we observe general consistent positive improvement across nearly all tasks. This suggests that language models with grounded knowledge tend to both do better on CQA tasks and are able to transfer causal reasoning across tasks more effectively.

BERT model and is trained on similar data. ERNIE 2.0 is trained on additional Reddit and Discovery data but the primary difference is in its knowledge-focused pretraining objectives.

We hypothesize that ERNIE 2.0 will outperform BERT across the CQA task as grounded knowledge is a requisite for causal reasoning in our definition. The BERT and ERNIE 2.0 implementations come from the Huggingface Transformers library (Wolf et al., 2020). We use the pretrained base models for both (`bert-base-uncased` [2] and `nghuyong/ernie-2.0-base-en` respectively [3]).

5.2 Language Model Training

Single-task fine-tuning and multi-task fine-tuning are used to train our models on the CQA tasks. Sequential fine-tuning (Pratt, 1992) was also investigated but found to be inconsistent and not as effective as the other methods (Appendix A.6.2). Following the task head paradigm introduced in Devlin et al. (2019), we develop separate classification heads for each task(see Appendix A.1 for

[2]https://huggingface.co/bert-base-uncased

[3]https://huggingface.co/nghuyong/ernie-2.0-base-en

	aNLI	*COPA*	*CosmosQA*	*E-Care*	*ROPES*	WIQA	Score
Fine-tuned Baseline							
Bert-base	0.61	0.64	0.57	0.76	0.58	0.65	0.64
ERNIE-base	0.64	0.71	0.63	0.76	0.53	0.64	0.65
MTL Baseline							
Bert-base MTL	0.62	0.75	0.58	0.72	**0.61**	0.72	0.67
ERNIE-base MTL	**0.65**	**0.80**	**0.65**	**0.78**	0.58	**0.77**	**0.71**

Table 4: We present the baselines results for CALM-bench. All the task are evaluated using the accuracy metric with the exception of ROPES which displays exact match. Results are presented for the test sets for COPA and WIQA and on validations sets for aNLI, CosmosQA, E-Care, and ROPES. We find that MTL models outperform the single-task finetuned models consistently with ERNIE-base MTL model having the best results.

more details). The pooled *CLS* embedding from the last layer in the language model is fed into the classification head to map the language model's contextualized output into the task's classification space. In the single-task setting, each task is trained independently. The cross-entropy loss is calculated per training batch and back-propagated through all the layers in the language model.

For the multi-task learning (MTL) model, we adapt a hard-parameter sharing model (Baxter, 2004) and train it using the multi-task fine-tuning strategy (Liu et al., 2019a). Our MTL model (Figure 2) consists of a shared base language model and separate task heads for tasks in CALM-Bench. For each train step, a train batch is sampled for each task and the task-specific losses are calculated. The task losses are averaged before backpropagation. The MTL model is trained for 8,000 steps on the aNLI, CosmosQA, E-Care, and WIQA tasks. The ROPEs task is not included in training as its format is significantly different from the multiple-choice tasks and resulted in lower performance in our early experiments. COPA was also omitted from the MTL training given its small size (800 training examples) and instead saved for zero-shot evaluation. At evaluation time, we fine-tune the MTL model on each target task for one additional epoch and then evaluate the model on the target evaluation set.

A hyperparameter search is run to identify the optimal random seed and the learning rate for each task (see Appendix A.5.1). Four of the tasks (aNLI, CosmosQA, E-CARE, and ROPES) have private test sets and a public leaderboard. For these tasks, we treat the validation set as the test set during evaluation and generate a new validation split from the training data to be used for training validation. The general intuition is that fine-tuned language models should have the best task-specific performance. If causal reasoning is transferrable, we should see improvements over the single-task fine-tuned models in both the transfer learning and multi-task learning experiments.

6 Empirical Findings

6.1 Single-task Fine-tuned Baselines

The baseline results for the single-task fine-tuned language models for all tasks can be found in Table 4. We find the ERNIE model on average outperforms the BERT model across most of the CQA tasks with an average improvement of 5.3pp on the aNLI, COPA, and CosmosQA tasks. However, ERNIE does underperform the BERT model on both the ROPES and WIQA tasks and shows no improvement on the E-Care task. These results are used as the baseline for the transfer learning and MTL experiments in Table 4.

6.2 Transferability of Causal Reasoning

We conduct sixty experiments to see if causal reasoning can be generalized and transferred across the QA tasks in CALM-Bench. For each experiment, we select a base task (e.g. aNLI) and a different target evaluation task (e.g. COPA). The language model is first fine-tuned on the base task and then fine-tuned on the target task. That model is then evaluated on the target task. Each transfer learning experiment is independent and the final results are summarized in Table 2 and 3.

Across both BERT and ERNIE 2.0 models, we observe that task-specific causal knowledge and reasoning are transferable. However, the pattern of transferrence differs across both models.

For the BERT model, the E-Care, WIQA, and ROPES tasks generally see degradation in accuracy

and exact match. However, there is improvement across aNLI, COPA, and CosmosQA tasks with COPA receiving an average of 7pp gain. We hypothesize this may due to two factors. As noted earlier, there is no grounded knowledge in BERT. BERT has to learn both task-specific knowledge and reasoning processes associated with each task. Tasks with similar knowledge requirements (aNLI, COPA, and CosmosQA) benefit from each other and the shared task format (multiple-choice). In contrast, the ROPES and WIQA tasks have different task heads and knowledge requirements. BERT is likely suffering from catastrophic forgetting when fine-tuning on the target task.

In contrast, we find consistent general improvement across all the tasks with the ERNIE 2.0 model. ERNIE 2.0 contains grounded knowledge which allows for better transfer learning across the tasks. This was observed with WIQA and ROPES seeing average improvements of 1.8pp and 1.4pp in contrast to the average losses of -0.08pp and -1.5pp with the BERT model.

To summarize, we provide empirical evidence that causal reasoning and knowledge can be transferred across different CQA tasks. We further find validate our assumptions that CQA requires both reasoning capabilities and grounded knowledge as the knowledge-rich ERNIE demonstrates more consistent improvement across the *CALM-Bench* tasks.

6.3 Multi-Task Learning Results

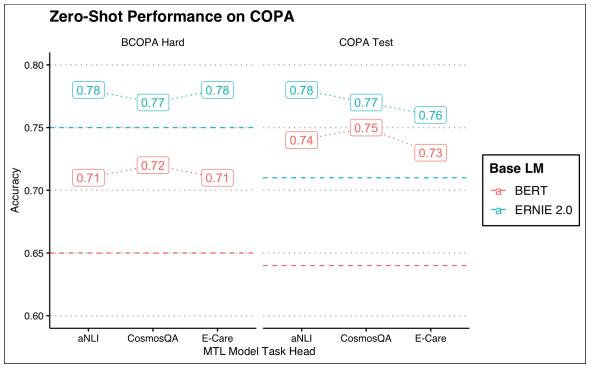

Figure 3: Zero-shot results on the COPA task. We present results for both the primary COPA Test set and the BCOPA Hard (Kavumba et al., 2019) subset. The dashed lines are the single-task fine-tuned baselines. Despite each task head having not seen the COPA examples during the MTL training, they outperform the single-task fine-tuned baselines.

In Table 4, we provide the baseline results for both the single-task fine-tuned and multi-task models on the CALM-Benchmark. The score column equally averages all metrics to provide a single value for comparing the different approaches. The MTL baselines outperform all the single-task fine-tuned baselines with the ERNIE MTL model providing the best results. These results further corroborate our claim that causal knowledge and reasoning are generalizable across diverse QA tasks. However, we do observe that task format matters. The inclusion of ROPES (a reading comprehension task) during multi-task training resulted in generally lower performance. As a result, our final MTL model was only trained on the subset of multiple-choice tasks (aNLI, CosmosQA, E-Care, and WIQA). Future work may consider alternative ways to weight task-specific losses or different model architectures (e.g. T5 (Raffel et al., 2020)) which can map all tasks to the same text-to-text format.

In the context of multiple-choice CQA, we find consistent and positive improvement across all tasks in both the single-task transfer learning and multi-task learning scenarios. We run an additional zero-shot experiment where is task head in the MTL model is used to evaluate the COPA test and BCOPA hard test examples. Figure 3 shows that both the BERT and ERNIE single-task fine-tuned baselines are outperformed by an average of +10pp and +6.6pp on the test set and see an average of +6.3pp and +1.3pp improvement on the BCOPA hard subset. For comparison, Hosseini et al. (2022) fine-tune a BERT large (345 million parameters) model on 780,000 knowledge triples from the ATOMIC commonsense knowledge base. Their BERT-Large-ATOMIC model achieves 88% accuarcy on the COPA test set and 73% accuracy on the BCOPA hard subset. Our smaller ERNIE 2.0 MTL model achieves 80% fine-tuned accuracy on the COPA test set with fewer parameters (110 million) and less training data. Further, our MTL model outperforms the BERT-Large-ATOMIC model on the BCOPA hard subset with the zero-shot MTL heads averaging around 77% accuracy and fine-tuned model achieving 79% accuracy.

7 Conclusion

In this paper, we provide a unified definition of causal question-answering in the context of natural language applications. Drawing from the cognitive science literature, we posit that CQA tasks

require both causal reasoning and causal knowledge. Based on this definition, we introduce the *CALM-bench*, the first multi-task CQA benchmark to evaluate the general causal reasoning capabilities of foundation language models. We provide empirical evidence which validates the intuition that causal reasoning and knowledge are transferable across the CQA tasks. Knowledge-enriched language models like ERNIE are likely to outperform distributional models (i.e. BERT) across all tasks in both the single-task fine-tuning and multi-task fine-tuning settings. Finally, we provide a set of strong baselines for future work exploring causal question-answering and the causal reasoning capabilities of language models.

While our experiments show causal knowledge is transferable, these models are still opaque. CQA provides a unique opportunity for model explainability through causal explanation structures and reasoning chains. The E-Care and WIQA task have annotated explanations that provide a useful starting point. Causal knowledge sources like CauseNet (Heindorf et al., 2020), ConceptNet (Speer et al., 2017), and Wikidata [4] can also be used to generate causal explanations. We believe the next evolution of foundation language models will have stronger causal reasoning capabilities and implicit structured causal knowledge. CALM-bench provides a starting point for further research on causal question-answering.

Limitations

Our research assumes the English language due to the lack of multi-lingual QA datasets. Future work may consider developing CQA tasks in other languages.

Additionally, we used the base models for BERT and ERNIE 2.0 in our experiments for all experiments. The public leaderboards for most of the tasks in *CALM-Bench* feature larger models with the billion parameter plus models occupying the top spots. Future work can explore scaling our experimental setup to the large and extra-large versions of our language models used as well as considering more modern architectures such DeBERTa (He et al., 2021) and ERNIE 3.0 (Sun et al., 2021). A challenge for multi-task training with large models is that the batch size for each task must be significantly reduced to ensure the model fits in GPU memory. Smaller batch sizes lead to unstable training and convergence. Tricks like gradient accumulation and modern optimization libraries (e.g. DeepSpeed [5] and Fairscale [6]) can be explored.

Finally, our multi-task model is not truly universal in the sense that a new task head is required for each additional CQA task. While there is transferability across the multiple-choice formats, the model does struggle to generalize causal reasoning across different formats like reading comprehension. Our encoder-only approach is unable to handle generation tasks. As a result, the E-CARE and aNLI explanation tasks are excluded. Lourie et al. (2021b) found success using encoder-decoder models where all tasks are converted to a text-to-text format. While (Lourie et al., 2021b) only considered multiple-choice tasks, future work could explore including reading comprehension and explanation generation tasks using models like UnifiedQA (Khashabi et al., 2020) and T5.

Acknowledgements

This work was supported by the Science Foundation Ireland under grants SFI/18/CRT/6223 (Centre for Research Training in Artificial Intelligence), SFI/12/RC/2289_P2 (Insight), and co-funded by the European Regional Development Fund.

References

Wookyoung Ahn, Charles W. Kalish, Douglas L. Medin, and Susan A. Gelman. 1995. The role of covariation versus mechanism information in causal attribution. *Cognition*, 54(3):299–352.

Jonathan Baxter. 2004. A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning*, 28:7–39.

Helen Beebee, Christopher Hitchcock, and Peter Menzies. 2009. Introduction. In *The Oxford Handbook of Causation*. Oxford University Press.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Yannick Boddez, Jan De Houwer, and Tom Beckers. 2017. The inferential reasoning theory of causal learning: Toward a multi-process prepositional account. Oxford library of psychology., pages 53–64. New York, NY, US.

[4] https://www.wikidata.org/

[5] https://github.com/microsoft/DeepSpeed

[6] https://github.com/facebookresearch/fairscale

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2021. Transformers as soft reasoners over language. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*.

Dhairya Dalal, Mihael Arcan, and Paul Buitelaar. 2021a. Enhancing multiple-choice question answering with causal knowledge. In *Proceedings of Deep Learning Inside Out (DeeLIO): The Second Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*. Association for Computational Linguistics.

Dhairya Dalal, Sharmi Dev Gupta, and Bentolhoda Binaei. 2021b. A semantic search pipeline for causality-driven adhoc information retrieval.

Tirthankar Dasgupta, Rupsa Saha, Lipika Dey, and Abir Naskar. 2018. Automatic extraction of causal relations from text using linguistically informed deep neural networks. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 306–316. Association for Computational Linguistics.

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

Phil Dowe. 2009. Causal Process Theories. In *The Oxford Handbook of Causation*. Oxford University Press.

Brett Drury, Hugo Gonçalo Oliveira, and Alneu de Andrade Lopes. 2022. A survey of the extraction and applications of causal relations. *Natural Language Engineering*, 28(3):361–400.

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)*. Association for Computational Linguistics.

William Falcon and The PyTorch Lightning team. 2019. PyTorch Lightning.

Zhihao Fan, Yeyun Gong, Zhongyu Wei, Siyuan Wang, Yameng Huang, Jian Jiao, Xuanjing Huang, Nan Duan, and Ruofei Zhang. 2020. An enhanced knowledge injection model for commonsense generation. *CoRR*, abs/2012.00366.

Roxana Girju. 2003. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering*, pages 76–83. Association for Computational Linguistics.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. SemEval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 13–18, Prague, Czech Republic. Association for Computational Linguistics.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. Association for Computational Linguistics.

Thomas L Griffiths. 2017. Formalizing prior knowledge in causal induction. *The oxford handbook of causal reasoning*, pages 115–126.

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.

Oktie Hassanzadeh, Debarun Bhattacharjya, Mark Feblowitz, Kavitha Srinivas, Michael Perrone, Shirin Sohrabi, and Michael Katz. 2020. Causal knowledge extraction through large-scale text mining. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13610–13611.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Wei Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *2021 International Conference on Learning Representations*.

Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020. Causenet: Towards a causality graph extracted from the web. In *CIKM*. ACM.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Pedram Hosseini, David A Broniatowski, and Mona Diab. 2021. Predicting directionality in causal relations in text. *arXiv preprint arXiv:2103.13606*.

Pedram Hosseini, David A. Broniatowski, and Mona Diab. 2022. Knowledge-augmented language models for cause-effect relation classification. In *Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022)*, pages 43–48. Association for Computational Linguistics.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: machine reading comprehension with contextual commonsense reasoning. *CoRR*, abs/1909.00277.

Samuel G. B. Johnson and Woo-kyoung Ahn. 2017. 127Causal Mechanisms. In *The Oxford Handbook of Causal Reasoning*. Oxford University Press.

Philip Nicholas Johnson-Laird and Sangeet Khemlani. 2017. Mental models and causation. *Oxford handbook of causal reasoning*, pages 1–42.

Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. 2019. When choosing plausible alternatives, clever hans can be clever. *arXiv preprint arXiv:1911.00225*.

Humayun Kayesh, Md. Saiful Islam, Junhu Wang, Shikha Anirban, A.S.M. Kayes, and Paul Watters. 2020. Answering binary causal questions: A transfer learning based approach. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–907. Association for Computational Linguistics.

Daniel Khashabi, Amos Ng, Tushar Khot, Ashish Sabharwal, Hannaneh Hajishirzi, and Chris Callison-Burch. 2021. GooAQ: Open question answering with diverse answer types. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 421–433, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christopher S. G. Khoo, Jaklin Kornfilt, Robert N. Oddy, and Sung-Hyon Myaeng. 1998. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing*, 13:177–186.

David Kellog Lewis. 1973. *Counterfactuals*. Blackwell.

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning over paragraph effects in situations. *CoRR*, abs/1908.05852.

Jian Liu, Yubo Chen, and Jun Zhao. 2020. Knowledge enhanced event causality identification with mention masking generalizations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3608–3614. International Joint Conferences on Artificial Intelligence Organization.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *ACL*.

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

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021a. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. In *AAAI*.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021b. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. *arXiv preprint arXiv:2103.13009*.

Ad Neeleman, Hans Van de Koot, et al. 2012. The linguistic expression of causation. *The theta system: Argument structure at the interface*, 20.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset.

Judea Pearl. 2009. *Causality: Models, Reasoning and Inference*, 2nd edition. Cambridge University Press.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases?

L. Y. Pratt. 1992. Discriminability-based transfer between neural networks. In *Advances in Neural Information Processing Systems*, volume 5. Morgan-Kaufmann.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Maarten Sap, Ronan LeBras, 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.

Rebecca Sharp, Mihai Surdeanu, Peter Jansen, Peter Clark, and Michael Hammond. 2016. Creating causal embeddings for question answering with minimal supervision. In *ACL 2016 Proceedings of Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Murray Singer, Michael Halldorson, Jeffrey C Lear, and Peter Andrusiak. 1992. Validation of causal bridging inferences in discourse understanding. *Journal of Memory and Language*, 31(4):507–524.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8968–8975.

Fiona Anting Tan, Xinyu Zuo, and See-Kiong Ng. 2022. Unicausal: Unified benchmark and model for causal text mining. *arXiv preprint arXiv:2208.09163*.

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. WIQA: A dataset for "what if..." reasoning over procedural text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6076–6085. Association for Computational Linguistics.

Michael R. Waldmann. 2017. *The Oxford Handbook of Causal Reasoning*. Oxford University Press.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhipeng Xie and Feiteng Mu. 2019. Distributed representation of words in cause and effect spaces. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7330–7337.

Jie Yang, Soyeon Caren Han, and Josiah Poon. 2022. A survey on extraction of causal relations from natural language text. *Knowl. Inf. Syst.*, 64(5):1161–1186.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. Curran Associates Inc., Red Hook, NY, USA.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. *CoRR*, abs/1808.05326.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *AAAI*.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021. Improving event causality identification via self-supervised representation learning on external causal statement. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2162–2172, Online. Association for Computational Linguistics.

A Appendix

A.1 Training Details

A.1.1 Training Environment

All models were trained on a single Nvidia A100 GPU and the `a2-highgpu-1g` Google Cloud Compute (GCP) instance. The GCP instance has 12 virtual CPUs and 85 GB of memory.

Model training was implemented using the Pytorch Lightning library (Falcon and The PyTorch

Lightning team, 2019). To ensure reproducibility we use the Pytorch Lightning `seed_everything` function which sets the random seed for the pytorch, numpy and the python.random libraries and the seeds used for data sampling.

The AdamW optimizer (Loshchilov and Hutter, 2017) and FP16 precision were used during training. Task specific learning rates were selected through a hyperpameter search (see Appendix A.5.1). For single-task fine-tuning experiments, the model was trained for 5 epochs and the model with the best validation accuracy was selected for evaluation. For the MTL experiment we train the model for 10,000 steps and checkpoint the model every 1,000 steps. The checkpoint (8,000 steps) with best average validation accuracy/exact match was selected for evaluation on the test set.

A.2 Multiple-Choice Tasks

In this section we detail the input format and the classification heads for the multiple-choice tasks in *CALM-Bench*. The aNLI, COPA, CosmosQA, and E-Care tasks all converted to the SWAG data format (Zellers et al., 2018) and we adapt the Huggingface `BERTforMultipleChoice` task head as the classification head.

The WIQA task is treated as simple multi-class classification problem. Provided a procedural description, question, and the answer options (more, less, and no effect) the input format is as follows: `[CLS] procedural description [SEP] question [SEP] more [SEP] less [SEP] no effect[SEP]`. The classification head is a single layer feed forward network which maps the pooled CLS token embedding of the language model's last layer into the label space.

A.3 Reading Comprehension Task

We treat the ROPES task as a SQuAD (Rajpurkar et al., 2016) style reading comprehension task and adapt the XLNET reading comprehension task head (Yang et al., 2019). Provided a question, hypothetical situation, and background passage we format the input as follows: `[CLS] question [SEP] hypothetical situation [SEP] background [SEP]`

The objective of the task head is to identify the answer span in the provided input text. The pooled CLS embedding of the last layer in the language model is fed to a feed forward network which independently predicts the start and end positions of the answer span in the input text. Beam search is run to identify the most probable start and end position, after which the answer text is extracted. Unlike SQuAD, the answer span is not always present in the situation description or background passage, but it is guaranteed to specified in the question text. As a result, we do not mask the question token positions during for the task head.

A.4 Sequence Classification Tasks

The causal sequence identification and counterfactual sequence identification tasks (Appendix A.6.1 are treated as binary classification tasks. The pooled CLS embedding of the last layer in the language model is fed to a feed forward network which maps it to a binary classification space.

A.5 Relation Extraction Tasks

We treat causal and counterfactual relation extraction tasks (Appendix A.6.1) as token classification tasks and adopt a custom BIO tagging format (Ramshaw and Marcus, 1995). Causal and counterfactual entities are tagged with the <cause>, <effect>, <antecedent>, and <consequent> begin and inside tags (e.g. <B-cause> and <I-cause>). All other tokens are labelled with the outside tag (<O>). The token embeddings of the last layer in the model are fed into a single layer feed forward network which predicts for each token the most probable tag.

A.5.1 Hyperparameter Details

We run a hyperparameter search for the random seed and learning rate for each task in *CALM-Bench*. We search over the following learning rates: [0, 1, 42, 1988, 2022, 3023] and randomly selected seeds: [1e-5, 3e-5, 2e-5, 5e-5]. The search is conducted in a two-stage process where we first identify the best learning rate and then identify the best random seed. During the search trial, the model is trained for 100 steps with the provided hyperparameter and then evaluated on validation set. The best hyperparameters are summarized in Table 5 and Table 6.

A.6 Additional Experiments

A.6.1 Transfer Learning Across CRI and CALM-Bench

For analyzing the relationship between CRI and CQA in the transfer learning context, we consider the following CRI tasks:

- **Causal sequence identification**: a binary classification task to evaluate if the sentence

Model	Huggingface Alias	Parameters	Task	Seed	Learning Rate	Batch Size
BERT	bert-base-uncased	110 million	aNLI	3023	2e-5	24
BERT	bert-base-uncased	110 million	COPA	1	1e-5	24
BERT	bert-base-uncased	110 million	CosmosQA	3023	2e-5	24
BERT	bert-base-uncased	110 million	E-CARE	42	2e-5	24
BERT	bert-base-uncased	110 million	ROPES	0	5e-5	24
BERT	bert-base-uncased	110 million	WIQA	1988	2e-5	24
ERNIE 2.0	nghuyong/ernie-2.0-base-en	110 million	aNLI	0	2e-5	24
ERNIE 2.0	nghuyong/ernie-2.0-base-en	110 million	COPA	42	2e-5	24
ERNIE 2.0	nghuyong/ernie-2.0-base-en	110 million	CosmosQA	0	3e-5	24
ERNIE 2.0	nghuyong/ernie-2.0-base-en	110 million	E-CARE	2022	3e-5	24
ERNIE 2.0	nghuyong/ernie-2.0-base-en	110 million	ROPES	42	3e-5	24
ERNIE 2.0	nghuyong/ernie-2.0-base-en	110 million	WIQA	1988	2e-5	24

Table 5: This table summarizes the best single-task fine-tuning hyperparameters task in *CALM-Bench*.

contains causal relata (i.e. cause and effect entities and a causal relation)

- **Causal relation tagging**: a sequence tagging task that requires identifying cause and effect spans provided a sequence of a token representing a sentence.
- **Counterfactual sequence identification**: a binary classification task to evaluate if the sentence contains counterfactual relata (i.e. antecedent and consequent entities and a counterfactual relation)
- **Counterfactual relation tagging**: a sequence tagging task requires the identification of consequent and antecedent spans from a sequence of tokens representing a sentence

SemEval 2007 Task 4 (Girju et al., 2007) and 2010 Task 8 (Hendrickx et al., 2010) tasks require classifying the relation given a pairs of entities in a sentence. We combine the SemEval 2007 Task 4 and 2010 Task 8 datasets to generate examples for causal relation identification and tagging. CREST (Hosseini et al., 2022) is used to convert all examples from the 2007 and 2010 tasks into a standardized sequence tagging format. For counterfactual tasks, we use the SemEval 2020 Task 5a and 5b datasets.

Table 8 and Table 9 summarize the results for these additional experiments. We find similar patterns to our CQA transfer learning experiments. With BERT, transfer between CRI and CQA tasks is not consistent. However, the ERNIE 2.0 model shows consistent improvement from CQA tasks to the Causal Id and Causal Relation identification tasks. Across both models there seems to be no transfer learning improvements on the counterfactual relation identification tasks.

A.6.2 Sequential fine-tuning Results

Table 7 summarizes the results of the sequential fine-tuning experiment with the BERT model. We start with a pretrained BERT model and then sequentially train it on the following multiple-choice tasks: WIQA, aNLI, CosmosQA, and E-Care. The model initially sees improvements over the single-task fine-tuned baseline results. However, as additional tasks are added, performance starts to degrade across several tasks. Due to the unstable results of sequential fine-tuning, we choose instead to pursue multi-task learning.

Model	Huggingface Alias	Parameters	Task	Seed	Learning Rate	Batch Size
BERT	bert-base-uncased	110 million	Causal Sequence Identification	42	5e-5	24
BERT	bert-base-uncased	110 million	Causal Relation Identification	1	5e-5	24
BERT	bert-base-uncased	110 million	Counterfactual Sequence Identification	1	1e-5	24
BERT	bert-base-uncased	110 million	Counterfactual Relation Identification	3023	5e-5	24
ERNIE 2.0	nghuyong/ernie-2.0-base-en	110 million	Causal Sequence Identification	0	2e-5	24
ERNIE 2.0	nghuyong/ernie-2.0-base-en	110 million	Causal Relation Identification	0	5e-5	24
ERNIE 2.0	nghuyong/ernie-2.0-base-en	110 million	Counterfactual Sequence Identification	42	3e-5	24
ERNIE 2.0	nghuyong/ernie-2.0-base-en	110 million	Counterfactual Relation Identification	2022	5e-5	24

Table 6: This table summarizes the best hyperparameter used for all CRI transfer learning experiments.

	aNLI	**COPA**	**CosmosQA**	**E-Care**	**ROPES**	**WIQA**
BERT single task-fine baseline	*0.61*	*0.64*	*0.57*	*0.76*	*0.51*	*0.65*
+ WIQA and aNLI	0.61	0.74	0.60	0.75	0.34	0.77
+ CosmosQa	0.61	0.72	0.57	0.75	0.30	0.75
+ E-Care	0.60	0.72	0.59	0.76	0.45	0.70

Table 7: Results from the sequential fine-tuning experiment. As additional tasks are added the model's performance starts to degrade across all tasks.

	Causal QA Tasks						**Relation Identification Tasks**			
	aNLI	**COPA**	**CosmosQA**	**E-Care**	**ROPES**	**WIQA**	**Causal Id.**	**Causal Rel.**	**CF Id.**	**CF Rel.**
Baseline	*.61*	*.64*	*.57*	*.76*	*.58*	*.65*	*.96*	*.68*	*.96*	*.62*
aNLI	N/A	+.11	+.04	0	+.01	-.01	0	+.01	0	-.02
COPA	+.02	N/A	+.04	-.04	-.06	0	+.01	0	+.01	-.02
CosmosQA	+.01	+.05	N/A	0	+.02	-.01	0	-.01	+.01	-.02
E-Care	+.02	+.13	-.02	N/A	-.05	-.02	-.02	+.02	+.01	0
ROPES	+.02	+.07	+.03	-.04	N/A	-.02	-.04	0	0	-.03
WIQA	0	0	+.03	-.02	-.05	N/A	+.01	+.02	+.01	0
Causal Id.	0	0	+.01	-.02	-.11	-.01	N/A	+.02	0	-.03
Causal Rel.	+.01	-.17	+.02	-.05	0	0	+.01	N/A	0	-.02
CF Id.	+.01	+.04	+.02	0	-.05	+.01	+.01	+.01	N/A	-.01
CF Rel.	0	+.01	+.03	-.04	+.02	+.01	+.01	0	0	N/A

Table 8: This heatmap table summarizes the transfer learning results of BERT model on the CALM-bench and CRI tasks.

	Causal QA Tasks						Relation Identification Tasks			
	aNLI	**COPA**	**CosmosQA**	**E-Care**	**ROPES**	**WIQA**	**Causal Id.**	**Causal Rel.**	**CF Id.**	**CF Rel.**
Baseline	*.64*	*.71*	*.63*	*.76*	*.53*	*.64*	*.94*	*.66*	*.96*	*.64*
aNLI	N/A	+.07	0	+.02	+.08	+.01	+.01	+.03	0	-.02
COPA	0	N/A	-.01	+.01	-.03	+.02	+.03	+.02	0	0
CosmosQA	+.02	+.01	N/A	-.01	-.12	+.01	+.02	+.03	+.01	-.03
E-Care	0	+.08	+.02	N/A	+.11	+.02	+.02	+.05	0	-.02
ROPES	+.02	-.06	-.01	+.01	N/A	+.02	+.02	+.01	0	+.04
WIQA	0	+.01	0	-.01	+.03	N/A	+.03	+.02	+.01	-.01
Causal Id.	+.01	0	+.02	-.01	-.11	-.03	N/A	+.03	0	0
Causal Rel.	+.01	-.08	-.02	-.01	+.04	+.01	+.03	N/A	0	-.06
CF Id.	0	+.03	+.02	-.01	+.11	+.02	+.03	+.01	N/A	-.02
CF Rel.	+.02	+.01	0	-.01	-.03	+.02	+.03	+.03	0	N/A

Table 9: This heatmap table summarizes the transfer learning results of ERNIE 2.0 model on the CALM-bench and CRI tasks.