Leipzig University
Institute of Computer Science
Degree Programme Computer Science, B.Sc.

Probing Large Language Models for Causal Knowledge

Bachelor's Thesis

Ruben Kohlmeyer

1. Referee: Jun.-Prof. Dr. Martin Potthast
2. Referee: Ferdinand Schlatt

Submission date: August 31, 2023

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Leipzig, August 31, 2023

. . . .
Ruben Kohlmeyer

Acknowledgements

First of all, I would like to express my sincere thanks to my professor Jun.-Prof. Dr. Martin Potthast for providing me with the interesting topic of my bachelor's thesis.

This thesis would not have been possible without my supervisor Ferdinand Schlatt. I want to give my great thanks to him for his continuous support throughout the whole thesis, for the intensive and regular discussions of the content, and for his helpful and constructive feedback.

Many thanks also go out to the Webis Group for allowing me to run my experiments on their hardware and for making their large causal knowledge base available to me.

I want to thank Lia Zaitchenko, Lea Pfeiffer, and my mother for proofreading this work and providing helpful suggestions for its improvement.

Last but not least I want to thank my dear friends and family for their support and encouragement during these last months.

Abstract

In this thesis we explore to what degree large language models (LLMs) can extract, learn and reproduce causal knowledge found on the internet. For this, we look at two LLMs trained on large amounts of English natural language texts and compare their output to the CauseNet, a large-scale graph of claimed causal relations. We test the output of the LLMs for a large number of very simple causal questions such as "smoking causes x. what could x be?" using evaluation, mask-filling, and text-generation. From these outputs, we devise metrics for gauging the confidence of each causal relation in the LLM.

We show that off-the-shelf LLMs possess a rudimentary ability to answer open-ended questions about causal inference. LLMs were more likely to generate and repeat causal claims that were frequently found in their training data, showing a correlation between the confidence the LLM has in a causal claim and that claims support in the CauseNet. Our adapted text-generation-based approach shows promising results in the task of aiding the construction of causal graphs, outperforming an evaluation-based approach from a previous paper by Long et al.. Furthermore, we find that using different techniques of prompt engineering as well as increasing the size of the model increases the performance of this task.

Contents

Chapter 1

Introduction

With recent advancements such as ChatGPT, Large Language Models (LLMs) have once again entered the public eye. LLMs are able to memorize knowledge contained in their training data and are able to regurgitate that knowledge through text generation [Hwang et al., 2021]. Parts of the training data may contain causal knowledge. This knowledge about causal relations is a fundamental part of how humans interact with their surroundings. Assigning cause and effect to events and phenomena is vital to building knowledge and understanding of the world.

In this thesis we want to find out to what degree LLMs can extract, learn and reproduce causal knowledge found on the internet. To answer this question, we look at GPT-2 [Radford et al., 2019] and BERT [Devlin et al., 2019], which are both LLMs trained on large amounts of English natural language texts.

To gauge how well the LLM does at this task, we compare its output to the CauseNet [Heindorf et al., 2020] (see figure 1.1), a large-scale graph of claimed causal relations. The relations in the CauseNet were extracted from natural language texts in the 2012 ClueWeb12 web crawl, as well as semi-structured sources like info-boxes from Wikipedia. The LLMs we look at are similarly trained on text found on the internet. This makes the CauseNet a good baseline for comparison. Each relation in the CauseNet graph links a cause to an effect. Additionally for each relation, the CauseNet provides the source count and the number of linguistic patterns (Support) that found this relation.

LLMs have been shown to possess the ability to extract and reproduce causal knowledge from their training data at a small scale [Long et al. [2023], Hobbhahn et al. [2022]]. To investigate this, the previous research used a relatively small sample of handcrafted phrases and relations or known benchmark sets for causal tasks, such as SemEval at 1 730 word pairs [Hassanzadeh et al.,

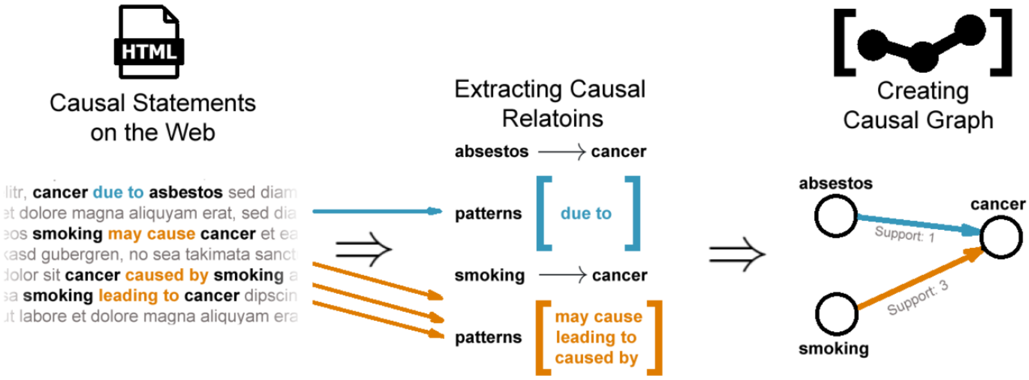

HTML
Causal Statements on the Web
litr, cancer due to asbestos sed diam
et dolore magna aliquyam erat, sed dia
eos smoking may cause cancer et ea
kasd gubergren, no sea takimata sanct
dolor sit cancer caused by smoking a
sa smoking leading to cancer dipscin
ut labore et dolore magna aliquyam era
⇒
Extracting Causal Relatoins
absestos ⟶ cancer
patterns
due to
smoking ⟶ cancer
patterns
may cause
leading to
caused by
⇒
Creating Causal Graph
absestos
cancer
Support: 1
Support: 3
smoking

Figure 1.1: A diagram illustrating the compilation of causal relations into the CauseNet.

2019]. The methods used in this thesis probe for causal knowledge at a much larger scale, relying on the large number of automatically extracted relations of the CauseNet, instead of a small number of manually selected ones.

Our research utilizes the CauseNet metrics of support and source count to further analyze the output of the LLMs and check for a potential correlation between relations favored by the LLMs and relations that score highly in the CauseNet.

To probe for causal relations we employ three approaches, each with their advantages and limitations.

(1) Perplexity, where the LLM is given a causal relation in the form of a short sentence and returns a score that represents how unfamiliar or perplexed the LLM is with that sentence, giving an insight into which sentences it would be likely to generate itself. However, this score is also influenced by the length, phrasing, or starting word of the sentence, instead of just the content of the causal relation. (2) Mask-filling, where the LLM is given an incomplete causal relation in the form of a short sentence with some parts masked out. The LLM returns words to fill in these masks, as well as a score of confidence for each answer given. Since the different words fill in the same mask, this score is not influenced by the varying length of the phrasing of the sentence in the way that perplexity is. The output of mask-filling always has the same form, making it easier to parse and analyze, but limiting the LLM in its answers. (3) Text generation, where the LLM is tasked to output its own causal relations. This allows the LLM the greatest degree of freedom in the answers given. While this freedom allows us to probe for a larger variety and number of outputs, it also introduces potential errors and makes the outputs harder to parse.

We also use text generation to see if it could be used to help construct an example causal graph, using a much smaller causal graph as ground truth rather than the CauseNet.

For all of this probing, we use pre-trained versions of GPT-2 and BERT transformer models. We compare how well different sizes of these models perform at these tasks and we employ different strategies of phrasing the sentences and prompts (such as n-shot prompting or switching linking words) to see if it can improve the LLM's performance.

Under our conditions, GPT-2 can achieve a precision of up to 13% and up to 6% recall at the task of generating relations found in CauseNet and a precision of up to 64% and up to 61% recall at the task of recreating a much smaller causal graph. The numbers around the CauseNet are quite low when compared to the results of papers that work with smaller sample sizes. They however broadly still showcase similar abilities of the LLM to learn and reproduce causal knowledge. In our tasks of prompting, mask-filling, and perplexity evaluation (when adjusted for starting word and input length) the LLM's confidence clearly correlates with that of the CauseNet. We show that, similar to the relations in the CauseNet, LLMs are most confident in short, very general causal relations, found frequently on the internet. These relations are more likely to be learned and generated by the LLM. Our results show that this ability to evaluate and reproduce causal relations scales with model size and is heavily influenced by different phrasings of the prompts, matching the results of Hobbhahn et al. [2022] in this regard.

Chapter 2

Related Works

In the following, we will go over the research directly related to this thesis followed by a brief introduction to LLMs and their ability to learn and reproduce knowledge. We will explain why causal knowledge in particular is a topic of interest, how causal tasks have been categorized, and how causal knowledge can be represented in causal graphs. We will go through different causal graphs, how they were created, and how they are composed. We will explore other works that aim to answer similar questions about the causal capabilities of LLMs, their methods, and their results. And finally we will give an overview of the CauseNet, its construction, and its use in this thesis.

2.1 LLMs and Knowledge

Recent advancements in AI have shown that LLMs possess impressive proficiency in natural language tasks. They are able to answer complex questions about a variety of topics, correctly identify and follow different tasks and generate large amounts of text e.g. for emails, essays, stories, or sections of program code [Tang et al., 2023]. They frequently show the ability to recall and apply knowledge from all across different domains. The ability to capture and memorize knowledge provides an opportunity to unlock various applications of LLMs in fields involving healthcare, research, and risk assessment [Long et al., 2023].

LLMs such as GPT2 [Radford et al., 2019] and BERT [Devlin et al., 2019] are trained on large amounts of English language texts in an unsupervised fashion. This means that they are trained on raw text with no human labeling, which allowed the researchers to use lots of publicly available data. In this process of training, LLMs are able to internalize relationships between concepts and knowledge about those concepts from data in the corpus. This knowledge then influences the answer an LLM might give when prompted with a question or asked to complete a certain task. GPT2 is a model trained to predict the next word at the end of a sequence, while BERT was trained to fill in missing words anywhere in a sequence. These models can have variants with different numbers of parameters and be specifically pre-trained to fulfill various different tasks. We will go into more detail regarding which specific models were used for this thesis in the next chapter.

2.2 Causal Tasks

Some of the most common questions are causal in nature. They are fundamental to humans' understanding of the world. Causality plays a role in determining cause and effect in a situation, inferring causation from correlating data, predicting the potential effects of an action, or determining the best action to take while considering many complex factors. Zhang et al. [2023] classify causal tasks into three types.

- **Type 1** are basic causal inference tasks that can be solved using domain knowledge. This includes finding potential causes of a given effect, potential effects resulting from a given cause, or determining what in a situation is the cause and what is the effect.

  Example: "I washed my car. My car got dirty. Which sentence is the cause of the other?"

- **Type 2** are causal discovery tasks that aim to discover a causal link from data. This includes estimating causation based on correlation or reasoning if an action had an effect on the data. This task cannot simply be solved by memorizing domain knowledge but requires a different skill set in high-precision causal reasoning.

  Example: "Here is a list of data on carbon emissions A:[...] and global temperatures B:[...] over the years. Determine if there is a link between A and B."

- **Type 3** are complex causal question tasks that require an understanding of the potential consequences of actions, the relation between many different inter-dependent factors, and the best path forward to reach a desired outcome. Depending on the setting this may require high precision in mathematical or logical reasoning.

  Example: "A patient returns for the third time with lumbago. The epidural steroid injections helped him before, but not for long. I injected 12mn betamethasone the last two times. What is the dose that I should use this time?"

Zhang et al. [2023] find that current LLMs are very promising at answering type 1 causal questions owing to their large collection of knowledge. However, they struggle with tasks of type 2 and type 3, since they require an understanding of underlying causal mechanisms to infer causality from data or make recommendations for very specific situations.

In Kıcıman et al. [2023], the researchers further distinguished causality along two axes. The first axis categorizes approaches by their primary emphasis on data analysis (covariance-based causality) or logical reasoning (logic-based causality). The second axis categorizes causality into reasoning about the broad relationship between variables (type causality) and reasoning about the cause and effect of specific events (actual causality).

Kıcıman et al. [2023] further differentiate between 4 main categories of causal tasks, those being effect inference, causal discovery, attribution, and judgment. These can broadly be assigned to the 3 types of causal tasks.

Effect inference (determining a causal relation) and *attribution* (determining the potential cause or causes) have approaches that rely on both covariance-based and logic-based reasoning, but are typically seen used for inferring type causality. Effect inference and attribution broadly match the categorization of type 1 task described above.

Causal discovery tasks (similar to those described in type 2 tasks above) were categorized as mostly falling into covariance-based type causality.

Judgment tasks extend attribution to questions about reward, blame, morality, and intent. A judgment task can take into account any or all covariance-based, logic-based, type, and actual causality. Having to take into account complex factors and the goal of determining the desired outcome in the context of morality or law puts the judgment tasks into type 3 as described above.

LLMs' ability to answer questions of causal discovery (type 2) has been a subject of recent research [Jin et al. [2023], Zhang et al. [2023], Tu et al. [2023]], which has shown that unspecialized LLMs perform poorly in this task, while LLMs specifically trained in this task were able to reach considerable accuracy. However, while tasks like this and many other causal tasks may also require extensive domain knowledge to answer, the experiments of this thesis will be focusing exclusively on tasks of type 1 (specifically type causality), as they are the most helpful when trying to answer the main questions of this thesis.

2.3 Probing for Causal Knowledge

In this section, we want to look at the different types of causal tasks that have been studied in regard to LLM capabilities in related research. For this we look at the kinds of methods used, the types and sizes of benchmarks used, and results relating to causality obtained from these experiments, as well as other useful observations that were gathered.

Hobbhahn et al. [2022] evaluated the performance of GPT3 at two similar causal tasks. Task one involved using real-world knowledge to identify cause and effect in a sentence and task two involved identifying cause and effect in a constructed scenario completely divorced from the real world. Both of these are tasks of basic causal inference (type 1). The knowledge tested in task one is taken from a part of BigBench, a benchmark for LLMs maintained by Google. The prompts tested in task two were manually constructed specifically to remove any real-world knowledge from them. The exact number of evaluated prompts was not stated in the paper, but we estimate it to be around 100 different prompts per setup. The researchers found, that GPT3 performed with an accuracy of up to 98% under ideal conditions, but was very susceptible to being influenced by the phrasing of a prompt, not just its content. Their experiments used different sizes of GPT3 to test how its performance at these tasks scaled with the size of the model, finding that larger models outperformed smaller models. Furthermore, the researchers employed different prompting strategies and found that preceding a prompt with correct examples (n-shot setup) and following a "question: ...? answer: ..." pattern improved their results. The aforementioned varying model sizes as well as their prompting techniques will also be employed in the text generation setup of this thesis.

Long et al. [2023] looked at LLMs' performance in the task of constructing causal graphs, specifically in the medical field. For this, they used the LLM to identify whether or not the relationship between two concepts was a cause-effect relationship. Similar to Hobbhahn et al. [2022] this is a type 1 causal task, but instead of using text generation, they relied on an approach in which GPT3 would rank two statements implying the presence or absence of a causal relationship. In this thesis we plan to also use a slightly altered version of this approach, instead utilizing text generation to rank these two statements. For their experiments they worked with a very small sample of just 18 manually crafted causal relations to evaluate the LLM's performance, yielding an average accuracy of 66.7%. We will also use this same sample as one of our benchmark sets, which will allow us to compare our approach to the approach taken by Long et al. [2023].

Hassanzadeh et al. [2019] tested the ability of LLMs to answer simple binary causal inference questions. To evaluate this ability the researchers worked with example pairs of cause and effect taken from 4 benchmarks handcrafted by human experts (SemEval with 1 730 word pairs, NATO-SFA with 118 word pairs, Risk Models with 804 word pairs, and CE Pairs with 320 word pairs). The LLM was tasked with identifying if a causal relationship existed between a given pair of two concepts. Thereby all of these tasks are type 1 tasks that rely on causal knowledge to answer. As established by Zhang et al. [2023] LLMs perform well at these type 1 tasks. Matching this the researchers found that their BERT implementation reached an accuracy above 50% in these tasks. To evaluate the preferred answer of the LLM they introduced a method called NLM-BERT, wherein BERT is used to encode a sentence “x may cause y” and compute the cosine similarity to its top-k most similar sentences to obtain a score. Since the experiments of this thesis already involve prompting BERT to generate its own relations, we can instead simply use the score it provides.

Li et al. [2021] trained a BERT-based model to increase its accuracy of different causal tasks. To train and evaluate their model they used a variety of benchmarks that can be categorized into two different types. The first task was causal classification, for which the researchers used the same benchmarks as Hassanzadeh et al. [2019]. Additionally, the researchers worked with COPA, another causal inference task using 500 examples, as well, as CausalQA and CosmoQA, two causal question-answering tasks, which fall under type 3 tasks. Similar to Hassanzadeh et al. [2019] NLM-BERT was used to evaluate the answer of the LLM. Matching the previous research, unspecialized versions of BERT performed above 50% at the type 1 tasks, but worse at the type 3 tasks. However, through specifically training their own model for causal tasks they were able to significantly increase accuracy for all tasks compared to the baseline.

Kıcıman et al. [2023] conducted four causal experiments using various state-of-the-art LLMs, working with the Tübingen Cause-Effect-Pair dataset containing 108 more common cause-effect pairs from across different domains and the Neuropathic Pain dataset containing 475 cause-effect pairs that rely on very specific domain knowledge to answer. In addition, they used the Arctic Sea Ice knowledge graph consisting of 48 edges to evaluate the LLMs’ abilities to reconstruct a small causal graph using causal inference. These 3 tasks were all type 1 causal tasks. The largest models had no problem solving either of these tasks with impressive accuracy. The researchers evaluated the LLMs’ abilities to complete tasks of counterfactual reasoning (type 3) using the Big-Bench benchmark containing 275 causal examples. Here too the results show that the larger models (especially GPT4) displayed substantial ability. Finally, they evaluated the LLMs’ performances at different related tasks, such as identifying necessary and sufficient factors and inferring normality from data.

Tu et al. [2023] conducted a small study on ChatGPT’s causal capabilities. Despite setting out to investigate tasks of causal discovery (type 2) the researchers admit that the tasks they used relied on causal knowledge and meta information instead of observational data, bringing them closer to causal inference tasks (type 1). For their experiments, they selected 100 examples from the Neuropathic Pain dataset and evaluated the outputs of ChatGPT manually. In addition to the binary answer given by ChatGPT, they took into account the additional reasoning the LLM would provide in its output. The results show that ChatGPT achieved a high precision but low recall in this task.

Jin et al. [2023] conducted a large-scale experiment (400 000 samples) to investigate the capabilities of LLMs at different causal discovery tasks (type 2). They found that off-the-shelf LLMs performed poorly at discovering causation from correlation, but that they could be specially trained to achieve high accuracy at this task.

Li et al. [2020] compared the performance of different LLMs at two type 1 causal tasks using 100 samples each from two datasets. The first task was a cause-effect classification task using the COPA dataset used by Li et al. [2021]. The second task was a causal inference task where either cause or effect had to be filled in by the LLM. This setup provided the LLM with an incomplete sentence (like "babies cry because ...") and prompted the LLM to finish the sentence through text generation. The discussed setup differs from binary- or multiple-choice tasks more commonly used in related research, which tests the LLM's ability to correctly identify causal relations without relying on a choice between predetermined options. Our thesis employs a comparable approach for generating effects from causes using LLMs. In their paper, Li et al. [2020] found that conventional transformer models like GPT2 produced low accuracy for the task of generating their own causal relations and performed considerably better at identifying cause and effect from a predetermined list. Unsurprisingly LLMs specifically trained for similar causal tasks like CausalBERT were able to outperform the conventional models.

Analyzing the existing research in this field we learned that almost all of the studies done previously worked with a comparably small sample size for their benchmarking (less than 10 000 samples) when probing for type 1 causal capabilities. Jin et al. [2023] used a large-scale benchmark of 400 000 samples, however only evaluated LLMs' performance at causal discovery tasks (type 2). Papers like Li et al. [2020] relied on a large number of causal relations to train their LLM, but used only a small subset of those relations to benchmark the LLM's performance. This thesis aims to build on the different approaches seen in the discussed papers (like mask-filling and effect-inference) and apply them to a much larger sample. In order to accomplish this we cannot rely on manually evaluating the output, but instead use automated approaches. The automated approaches for evaluating outputs in related research papers mostly rely on the restricted and formulaic nature of multiple-choice questions, which finds only limited application in our thesis. The specifics of our automated approach will be discussed in the next chapter.

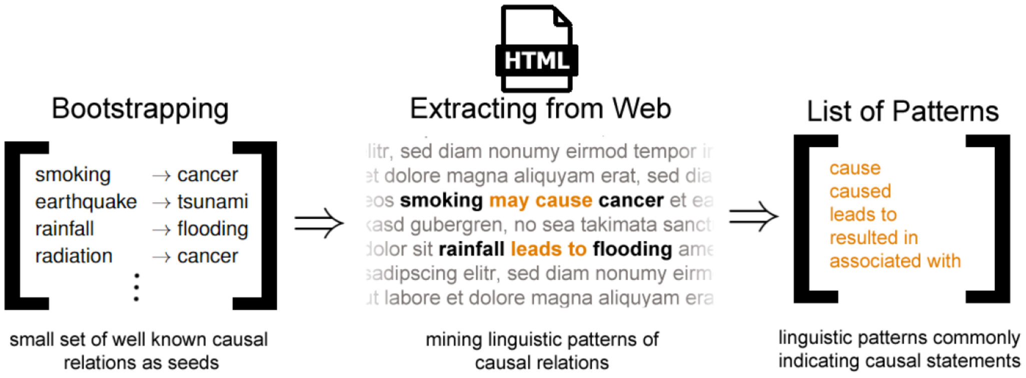

Bootstrapping
smoking → cancer
earthquake → tsunami
rainfall → flooding
radiation → cancer
small set of well known causal relations as seeds
HTML
Extracting from Web
litr, sed diam nonumy eirmod tempor
et dolore magna aliquyam erat, sed dia
eos smoking may cause cancer et ea
asd gubergren, no sea takimata sanct
dolor sit rainfall leads to flooding ame
sadipscing elitr, sed diam nonumy eirm
ut labore et dolore magna aliquyam era
mining linguistic patterns of causal relations
List of Patterns
cause
caused
leads to
resulted in
associated with
linguistic patterns commonly indicating causal statements

Figure 2.1: A diagram showing the bootstrapping process used to create the CauseNet.

2.4 CauseNet

The CauseNet [Heindorf et al., 2020] is a large-scale knowledge base of claimed causal relations. It is comprised of over 11 million causal relations, making it one of the largest knowledge bases for causal knowledge. Before we explore how the CauseNet was utilized as a ground truth in the next chapter, we will first go into how the CauseNet was constructed, how it is composed, how it can be used, as well, as give examples of causal claims contained within it.

The CauseNet was compiled from a large amount of natural language texts with the purpose of cataloging causal beliefs commonly expressed on the internet. It draws its data from two sources: the ClueWeb12, a web crawl made up of over 730 million English web pages, and Wikipedia, taking into account the natural language text on the site as well as information contained in lists and info-boxes.

The CauseNet was automatically generated in three steps (see figure 2.1) with minimal supervision. First, a small set of well-known causal relations were chosen as seeds, which included relations like "smoking → cancer" and "earthquake → tsunami". Using these seeds, Wikipedia was searched for sentences that contained both cause and effect from any seed, which were then mined for linguistic patterns expressing the causal relation. Gathering these linguistic patterns gives insights into how causal relations are expressed in natural language text. In the next step, the extracted patterns were used to find more causal relations in the corpus. In the final step, some of the collected relations were selected based on the number of different patterns that found them (support) and added back into the pool of seeds. This process was iterated twice, yielding 53 different linguistic patterns that were then used to search the entire web crawl.

For each causal relation, the CauseNet keeps track of cause, and effect, the total number of times this relation was found in the data, and the number of different linguistic patterns that found this relation. The latter number is referred to as the support of a relation.

For example if the relation “smoking → lung cancer” was represented in the data with the sentences

“smoking **causes** lung cancer” and
“smoking **leads to** lung cancer”

the relation would be given a support of 2.

The support metric provides a way to express in how many different ways a given relation was expressed in the corpus, which can be used to discern the CauseNet’s “confidence” in the relation. A relation expressed in many different ways is more likely to accurately represent beliefs about causal claims found on the internet. Causal relations assigned high support in the CauseNet tend to be common and short general statements between broad concepts. These are relations very frequently expressed in a variety of ways. In contrast, relations with low support in the CauseNet tend to be very narrow, domain-specific, and wordy.

For our ground truth, we only consider the “High-Precision-CauseNet”, a sub-graph of the CauseNet, containing only relations with support of 2 or higher, which aims to increase precision by reducing the number of only weakly supported causal relations and still leaves around 198 000 claimed causal relations.

For the automated parsing of LLM outputs we make use of the full “High-Recall-CauseNet”, which contains over 11 million claimed causal relations.

In this thesis, we will be working with the CauseNet as our ground truth. Since the LLMs examined in this thesis were trained on a similar corpus to the CauseNet (large-scale natural language text extracted from the internet), this allows us to infer connections between the causal claims found in the training data and the causal relations learned and reproduced by the LLMs. The large size of the CauseNet enables us to prompt at a larger scale (over 10 000 samples) than related research (see section 2.3). As discussed in a later chapter all our prompts were sampled from a subset of causal relations in the CauseNet. The large scale of our prompting set allows us to use an additional metric to evaluate the LLMs’ confidence in a relation by counting how often it repeats that relation across different prompts. This approach will be discussed in the next chapter.

We will also make extensive use of the support metric of the CauseNet. We expect that similar to the most supported relations in the CauseNet, LLMs also hold the strongest confidence in commonly found, general causal statements. Using this, in our thesis, we set out to see if such a correlation can be observed between causal relations commonly found on the internet (high CauseNet support) and the causal relations favored by the LLM.

While working with the CauseNet it is important to keep in mind a few caveats. (1) In order to evaluate how well an LLM learns causal knowledge from its training data one has to compare its output to the causal knowledge contained in its training data. We made the assumption that the training data is similar enough to the data from which the CauseNet was constructed for us to be able to make this connection. However, there are differences between the two datasets that must be taken note of. The CauseNet was constructed from a web crawl taken from 2012 as well as Wikipedia, while GPT2 was trained on a number of English language books and websites from 2019 or later and BERT was trained on the Brown corpus and Wikipedia. While there is likely considerable overlap between the different training sets and the CauseNet, it is hard to say how much this might influence the accuracy of our results. (2) We need to keep in mind that the CauseNet is a collection of *claimed* causal relations. The results of this thesis, therefore, cannot show how accurately an LLM performs at producing *correct* causal relations, only how well it can learn the causal knowledge encoded in its training data.

Chapter 3

Methods

In this thesis, we use three different methods of probing LLMs for causal knowledge: (1) Perplexity, (2) mask-filling, and (3) text generation (see figure 3.1). For each of those methods, we will go into detail on how the prompts given to the LLM were selected, what metrics of confidence we used, and how the output of the LLM was analyzed.

To compare the capability to generate and evaluate causal relations at different sizes of LLM, we chose four different variants of GPT2 [Radford et al., 2019] and three different variants of BERT [Devlin et al., 2019]. For the tasks of perplexity evaluation and text generation, we chose GPT2 (124M parameters), GPT2-Medium (355M Parameters), GPT2-Large (774M parameters), and GPT2-xl (1.5B parameters) as pre-trained transformer models from Huggingface [Wolf et al., 2020]. For the task of mask-filling we chose BERT-Base-Uncased (110M parameters), BERT-Large-Uncased (340M parameters), and BERT-Large-Uncased-Whole-Word-Masking (340M parameters) also as pre-trained transformer models from Huggingface.

As a ground truth, we worked with the CauseNet, meaning that we examined how many of the causal relations generated by the LLM were also present in the CauseNet, as well as comparing our different confidence metrics to support and source numbers in the CauseNet. Even the smaller "High Precision CauseNet" contains around 198 000 claimed relations, which is orders of magnitude larger than benchmark data sets such as NATO-SFA or SemEVAL used by previous research like Li et al. [2021] and Hassanzadeh et al. [2019]. This large ground truth allows us to probe the LLMs at a larger scale than in previous papers. We work with the CauseNet as the ground truth, as well as a source of sample relations for creating prompts.

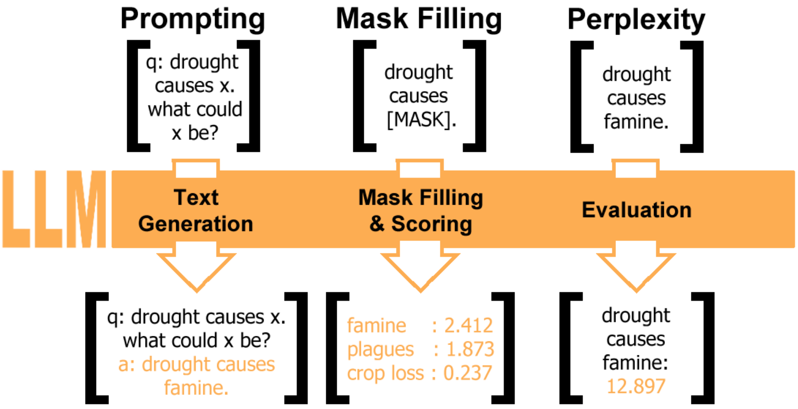


Figure 3.1: Examples of input and output of the three probing methods used in this thesis

3.1 Perplexity

Perplexity is a metric that measures the uncertainty of generating a word in a sequence, based on the previous words. It is calculated using the average inverse log-likelihood, which results in texts that are more likely to be generated by the LLM having a lower perplexity [Tang et al., 2023].

By evaluating the perplexity for a large and diverse sample of relations it is possible to gain insight into which relations the LLM favors. To create this sample we pseudo-randomly selected over 12 000 relations from the CauseNet, making sure it would include relations ranging in support from 2 to 39.

In order to evaluate the perplexity of a relation ([cause], [effect]) we constructed a simple natural language sentence that encoded it. To construct these sentences we employed two different approaches. In the first approach all sentences took the form "[cause] causes [effect].", only using "causes" as the linking word. In the second approach instead of "causes" for each relation we used the most common linking word with which the relation was found in the CauseNet. Examples of relations and the resulting sentences can be found in figure 3.2.

The perplexity of a word is influenced by the words that precede it. This needs to be kept in mind when comparing the perplexities of different sentences. In short sentences, the starting word has a large influence on the perplexity. An unlikely starting word could increase the calculated perplexity of an otherwise coherent sentence. To account for this we separated our results by starting words and different sentence lengths.

Such problems with the perplexity method have to be accounted for or even worked around, making this method not ideal when trying to probe for causal

heart failure → death
smoking → lung cancer

(a) 2 relations selected from the CauseNet sample

First approach:
"heart failure causes death"
"smoking causes lung cancer"
Second approach:
"heart failure results in death"
"lung cancer caused by smoking"

(b) Sentences constructed from the relations in (a). Linked with "causes" or with their most common linking words in the CauseNet.

Figure 3.2: Examples of sentences used in perplexity evaluation

knowledge. In contrast to mask-filling and text generation, the perplexity method also does not use the LLM's ability to generate its own relations.

3.2 Mask Filling

For a mask-filling task, an LLM is given a sentence with some of the words masked out such as "[MASK] causes death.". The LLM is then tasked to predict which words should replace those masks. It returns its top-k predictions along with the probability score of each one.

Similar to 3.1 we can use a sample of relations from the CauseNet to construct natural language sentences, this time with parts of them masked out. Each sentence was constructed with 10 different linking words and the first concept in the sentence was masked out. Since half of the linking words reversed the order of cause and effect in the sentence the masked out concept was not always the cause.

Compared to text generation (see 3.3), the method of mask-filling is more restrictive in terms of which outputs the LLM can produce. On one side this can be beneficial, making the outputs easier to parse automatically. On the other hand, it also excludes some causal relations from being generated by the LLM. In the prompt, either the cause or the effect is already filled in and cannot be changed by the LLM. This means that the LLM cannot generate relations in which both cause and effect are arbitrary. Since the LLM can only ever replace a mask with a single word, there are some causal relations that can never be generated through this method. The LLM is further limited by its own vocabulary since it can only fill the mask with words that appear in its vocabulary. Considering all these restrictions, only around 64 000 (32%) of the around 198 000 CauseNet relations could ever possibly be generated using this method. However, an upside of using mask-filling as compared to text generation is that in addition to generating its own causal relations, the LLM also returns a probability score, which can serve as a measure of confidence for that relation.

For the relations generated in this way by the LLM, we used two metrics to represent which relations the LLM favors. The first metric is the probability score provided by the LLM (if a relation was generated multiple times with different scores, we use the average). The second metric is the number of times the same relation was generated by different prompts. Finally, we calculated the precision and recall of the mask-filling setup and compared the output to the CauseNet, looking for a potential correlation between the LLM's metrics of confidence and the CauseNet's support and source number.

3.3 Text Generation

For the task of text generation, an LLM is given a prompt and continues it by generating text. The generated text is in large parts based on the form and content of the prompt. We use this fact to instruct the LLM to generate causal relations while being careful not to skew its results.

Similar to perplexity (3.1) and mask-filling (3.2) we worked with a sample from the CauseNet to construct natural language sentences from. But other than in perplexity and mask-filling the prompts for text generation have to be more complex to ensure that the LLM can reliably generate causal relations. Compared to mask-filling the LLM has a greater degree of freedom for its response in text generation. Where in mask-filling it is limited to filling in one word at a time, in text generation the length of the responses can be arbitrary. As such the opportunity for prompting more- and more diverse causal relations arises. However this also introduces a lot of noise in the form of results that do not generate causal relations at all. To reduce this noise we employ n-shot prompting as our prompting strategy. Instead of simply asking the LLM to generate causal relations, we first provide an example of a question and an answer in the correct form. The LLM can then learn from this correct example to generate text that follows similar patterns [Wei et al., 2023].

From the sample of CauseNet relations, we took the causes and asked the LLM to generate text filling in possible effects. In addition to this, the prompt would include either one or three example pairs of correct question and answer for 1-shot and 3-shot prompting respectively. An example of a 1-shot prompt is illustrated in figure 3.3. The relations chosen for these examples were also semi-randomly taken from the CauseNet while making sure to not skew their distribution towards high or low support. The question-and-answer examples were constructed in a very formulaic way. We chose this form to make it easier to extract the causal relations from the text generated by the LLM. The LLM

high fever	→ death
high fever	→ convulsions
high fever	→ hair loss
cancer	→ hospitalization

(a) 4 relations selected from the CauseNet sample

“q: high fever causes x. what could x be?
a:
1. high fever causes death.
2. high fever causes convulsions.
3. high fever causes hair loss.
q: cancer causes x. what could x be?”

(b) A prompt constructed from the relations in (a)

Figure 3.3: An example of a 1-shot prompt used in text generation

was, for example, encouraged to express its causal relations in short, simple sentences, each separated by a line break.

To construct the set of prompts each unique cause in the sample was combined with 5 different example questions and answers each. We did this to reduce the probability of our results being too heavily influenced by specific examples. For each pair of example answer and question, we then asked the LLM to generate answers of 40 tokens or fewer.

For 3-shot prompting, we picked the example questions and answers in the same way but added 3 pairs of question and answer before each prompt.

We asked the LLM to generate text for over 23 000 prompts each for the 1-shot and the 3-shot approach.

For the relations generated in this way by the LLM, we counted the number of times the same relation was generated by different prompts. We used this metric to represent which relations the LLM favors. We then calculated precision and recall as compared the our ground truth, as well, as the correlation between the CauseNet’s metrics (support and source number) and the LLM’s metric.

In addition to our comparison to the CauseNet, a very large causal graph, we also examined if the method of text generation could also be used to reconstruct a small causal graph. For this, we chose the set of 4 small causal graphs used in the paper by Long et al. [2023]. For this, we constructed a set of 1-shot prompts based on the relations on this much smaller causal graph. These prompts are similar in structure to the prompts constructed from the CauseNet. However, we also prompted the LLM to generate statements implying the absence of a causal relation such as “smoking does not cause diabetes”. In the paper by Long et al. [2023] had the LLM score two statements, one implying the presence and one implying the absence of a causal relation, in order to determine if a causal relation between two concepts should be present. We used a slightly different method based on the number of times each relation was generated to decide if there should be an edge or no edge between any pair of two concepts. For this, we used the number of times the LLM generated a sentence implying the presence or absence of the edge as our result. For each possible relation in the graph, we consider the LLM to be accurate if it generated more sentences agreeing than disagreeing with the graph.

3.4 Text Parsing

Extracting causal relations from the text generated by LLMs (as done in 3.3) is not trivial. The generated text can encode the causal relation for example in the form of a comma-separated list, a natural language sentence, or even not encode any causal relations at all. At a smaller scale, it is possible to manually go through all of the generated text and extract the causal relations encoded within it. However, the scale at which we were prompting made this approach unfeasible for us. Instead, we chose to automate this process.

For the purposes of parsing the output text, we assumed that each answer to the question was separated by a line break and contained a linguistic pattern that directly indicated cause and effect. If no causal relation could be found to be indicated in a given line of generated text, the entire line was treated as an effect. This effect in combination with the cause from the prompt would then be added to the list of extracted relations.

This parsing algorithm performed better at parsing some types of texts over others, preferring short and clear sentences that stated both cause and effect explicitly. We encouraged the LLM to generate answers of this form, by providing examples of just that in the n-shot prompts. To analyze how well the LLM could adhere to this pattern, we categorized different ways that the generated text veered from the pattern. Examples of these ways that we categorized can be found in figure 4.4.

If the n-shot pattern was not followed, each line of the full text response generated was simply assumed to be a claimed effect corresponding to the cause in the prompt. However, this would also be applied to responses consisting of full sentences or nonsensical outputs that do not make a claim to any causal relation at all. To discern if an output could be plausibly interpreted as a claim to a causal relation, we utilized the High-Recall-CauseNet. If a pattern-less result from the LLM never occurred as cause or effect (or as part of cause or effect) in the High-Recall-CauseNet, it was regarded as nonsense (making no claim to any causal relation).

Taking a manual sample of 205 pattern-less relations that were tagged in this way yielded an accuracy of 71% for this method of differentiating nonsense from causal claims.

Chapter 4

Results

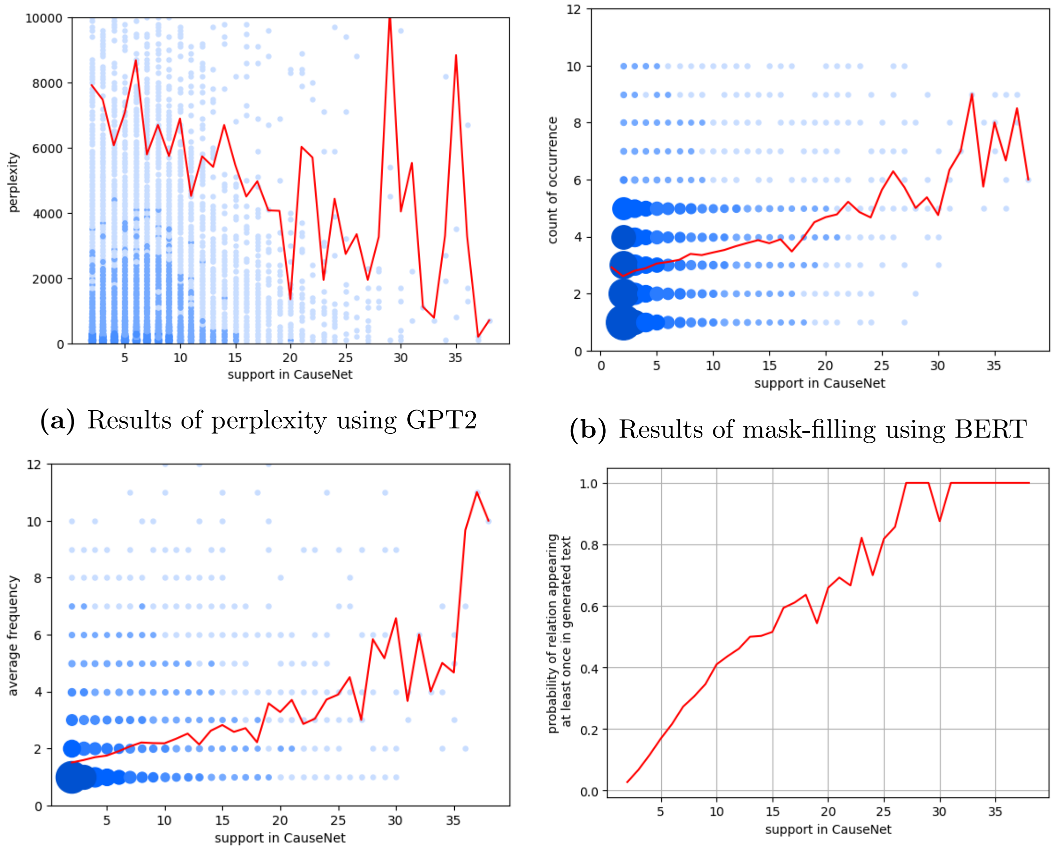
In the previous chapter, we detailed the different methods of probing for causal relations in LLMs. In this chapter, we want to answer how well LLMs can extract, learn and generate causal relations found on the internet. To do this we analyze the results of our experiments, give examples, discuss their overall composition, and compare them to the ground truth. Furthermore, we will show how different methods of creating prompts and parsing generated text can influence the LLMs' performance.

4.1 Correlation with CauseNet

In our setups, we used LLMs to assign the metrics of perplexity, score (in mask-filling), and count (in mask-filling and text-generation) to assign numbers to causal relations. We used these metrics to find out which relations the LLM was confident in. A high count metric of a relation means that the relation was generated multiple times from different prompts, indicating high confidence. Likewise, a high score metric directly shows that the LLM was confident in its answer. The inverse is true for perplexity since it reflects the degree to which an LLM is unfamiliar with an output. Here a low perplexity indicates higher confidence.

We then used these metrics to compare them with the metrics of support and source number in the CauseNet. For mask-filling and text-generation a positive correlation between the metrics would indicate that the LLM prefers similar causal relations to the CauseNet. Since perplexity is inverted, a negative correlation between it and the CauseNet metrics would indicate a similar overlap in preferred relations.

In figure 4.1 (c) we see that in the text generation setup the majority of the generated relations cluster in the lower left corner. These relations have low support in the CauseNet and are only generated very few times by the



(a) Results of perplexity using GPT2

(b) Results of mask-filling using BERT

(c) Results of text-generation using GPT2

(d) Recall of text-generation split by support

Figure 4.1: The results of text generation, mask-filling, and perplexity. Each relation is plotted according to its count of occurrence (or perplexity) and its support in the CauseNet. The average of all relations with the same support is shown in red. (d) shows the probability that a relation of specific support is generated at least once by the LLM. For (b) and (c) the values in the graph are discretized since the count of occurrence can only be a whole number.

LLM. This corresponds to the overall makeup of the CauseNet. Similar to the distribution here, the majority of the CauseNet is made up of relations with low support that were only found a couple of times in the corpus. We also see that relations with higher support in the CauseNet were on average more likely to be repeated by the LLM. Furthermore, we found that a relation with a support over 30 had a near 100% chance to be generated at least once, a relation with support of 15 had a 50% chance and a relation of support 5 had only a 20% chance to be generated at least once. From this, we can see that the LLM was more likely to generate and repeat generations of high CauseNet-support.

Similar observations are also true for the results of mask-filling (sub-figure (b)). The majority of generated relations cluster towards both low support and low count. However, unlike in text generation, we also see a significant clustering in relations up to a count of 5, above which the occurrences quickly decline. We suspect that this is due to the construction of the prompts used in mask-filling. Each prompt would be asked 5 times with varying linking words, leading to the LLM being more likely to repeat the same relation up to 5 times.

In the perplexity setup (subfigure (a)), we see a less pronounced but still present clustering of relations with low support and low perplexity. At face value, this contradicts our expectations, since we would expect lower support relations to have a higher perplexity. We also found that lower support relations, while largely clustered around lower perplexities, also tended to be spread into much higher perplexities. Relations of low support were also more likely to produce outliers with extremely high perplexity. Since the average of a set is very susceptible to outliers, we suspect that this artificially increased the average perplexity for lower support relations.

To quantify the correlation between our different metrics and the CauseNet, we calculated the correlation coefficient. For this, we used Spearman's rank coefficient, the results of which are shown in table 4.1 and discussed in the following paragraphs.

As discussed in Chapter 3, perplexity alone is not a great measure of the confidence an LLM has in a causal relation, since it is heavily influenced by factors such as the length of the sentence and what starting word is used. To account for this we used different approaches. For the first approach ((a) in table 4.1) we compared the perplexity of all relations produced by our setups. The first setup (naive) used only "causes" as a linking word, while the other setup used the most common linking word for each relation. We found that the results from approach (a) have no statistically significant correlation with either support or source number in the CauseNet. To account for the influence that sentence length has over perplexity, we calculated the correlation for only relations with the same number of words (b). This yielded small and negative, but statistically significant coefficients. When separated by sentence length like this we found that on average, shorter sentences had a higher perplexity. We suspect this to be because of the greater influence that an unexpected starting word has in a short sentence. To account for the difference in starting words we looked at only relations with the same starting word ((c) and (d)). In the example in table 4.1 we looked at the perplexities only of relations that started with "smoking causes ...". We found that accounting for starting word and sentence length yields a statistically significant negative correlation between perplexity and support for almost all starting words that we checked. However, we also found that this did not hold true for "conditions" as the starting word,

approach	model	metrics		method	coeff.
perplexity	gpt2-xl	perpl.-support	(a)	naive	*0.0221
	gpt2-xl	perpl.-support		linking word	*0.1413
	gpt2-xl	perpl.-sourceNo.		linking word	*0.1120
	gpt2-xl	perpl.-support	(b)	length 3	-0.0508
	gpt2-xl	perpl.-support		length 5	-0.2689
	gpt2-xl	perpl.-support		length 7+	-0.3653
	gpt2-xl	perpl.-support	(c)	“smoking”	*-0.1222
	gpt2-xl	perpl.-support	(d)	“smoking” & len. 3	-0.3046
mask-fill.	BERT-base	count-support		one-sided	0.1845
	BERT-base	score-support		one-sided	0.1907
	BERT-large	score-support		one-sided	0.2207
text-gen.	gpt2-base	count-support		1-shot	0.0687
	gpt2-xl	count-support		1-shot	0.1841
	gpt2-xl	count-support		3-shot	0.2160
	gpt2-xl	count-sourceNo.		3-shot	0.1895

Table 4.1: Spearman rank coefficients between different LLM metrics with the metrics of the CauseNet. We compared coefficients for different model sizes and methods. The coefficients for some perplexity correlations were not statistically significant, these are marked (*) in the table.

which produced no correlation even when controlling for sentence length.

Looking at the results of mask-filling we see a consistently small but significant correlation between both score and the CauseNet as well as count and the CauseNet. Of the two metrics produced by the mask-filling setup, the score showed a slightly higher correlation with support than the count did. We also found that for both score and count the correlation with the CauseNet scaled with the size of the model. This positive correlation shows that BERT tends to favor relations with higher support and will both assign them higher scores and repeat them more frequently.

The results of the text generation show a small, but significant correlation with the CauseNet, similar to the results of mask-filling. This correlation increases with the size of the model as well as with the number of n-shot examples. Comparing the two CauseNet metrics to the text-generation count shows us that count shows a greater correlation with the support metric. We suspect that this is the case since the count metric in text generation is similar to the support in the CauseNet. Support indicates how many different linguistic patterns found a given causal relation, while count indicates how many different prompts generated a given causal relation.

Looking at the overall results of text generation and mask-filling we find that the correlations shown prove a measurable link between the causal knowledge found in the training data of an LLM and the causal knowledge it can reproduce. LLMs are more likely to repeat common, general causal relations that are also favored in the CauseNet. This shows that support in the CauseNet is a helpful metric for predicting the degree to which an LLM can memorize and reproduce causal claims.

However, this link between the CauseNet and the LLM does not show up when looking at perplexity, making perplexity alone an unreliable metric for evaluating an LLM's confidence in a causal relation. However, by accounting for influence outside of the causal relations itself (sentence length, starting word) we found that under certain conditions perplexity can be used as an indicator of confidence for causal relations.

4.2 Precision and Recall

The methods of mask-filling and text generation both prompted the LLM to create its own causal relations. We will look at the precision and recall of these results for different sizes of LLM and prompting strategies. Since we allow and even encourage the LLM to generate the same relation multiple times we distinguish between macro- and micro-precision. For the purposes of macro-precision, we consider all instances of the same relation being generated to form one class and count how many of these classes are also present in the ground truth. For micro-precision, we count each relation separately.

4.2.1 Mask-Filling

As discussed in Methods, this method as well as the LLM used places some restraints on the possible outputs. The LLM can only ever replace the mask with one word, so a prompt of the form "[MASK] causes cancer" could produce the answer "smoking", but not the answer "cigarette smoking". Furthermore, in order for the LLM to use it, the word that replaces the mask has to be in its vocabulary. In our one-sided approach we also always provided either an effect or a cause from the sample. This meant, however, that relations where both cause and effect were not in the sample could not be generated by the LLM since one side would always be unchangeable. Of the around 198 000 relations in the CauseNet only 63 682 (32%) could ever possibly be generated by the LLM with this method. This restriction needs to be taken into account when looking at recall, since we calculate it in regards to only relations generateable by this method, instead of in regards to all relations in the CauseNet.

As seen in table 4.2 we found that less than 10% of the relations generated by BERT also appeared in the CauseNet. Looking closer at the relations

model	micro-precision	macro-precision	recall
base	9.12%	6.18%	39.49%
large	**9.44%**	7.90%	40.97%
large-wwm	9.35%	**7.91%**	**41.24%**

Table 4.2: Results of the Mask-Filling method, comparing the models BERT-base-uncased, BERT-large-uncased and BERT-large-uncased-whole-word-masking all using the one-sided approach described in Methods.

reveals that this is in part because of words like “it”, “this” or “he”, single letters or abbreviations like “mr” or “j” or first names like “thomas” frequently getting filled in for cause or effect. While “thomas causes cancer.” or “smoking causes it.” are linguistically sound sentences, they do not contain valid causal relations. We further found that BERT achieved a higher micro- than macro-precision. This observation holds true for all the sizes of BERT that we tested. This shows that BERT was more likely to repeat relations that also appear in the CauseNet than those not in the CauseNet, which matches the correlation shown in section 4.1

In these mask-filling experiments, we were able to reach a recall of around 40%, which is impressive considering the size of our ground truth. Judging from the precision we can gather that BERT has some ability to generate causal relations. These results are quite low compared to the results of other causal inference tasks (such as cause and effect identification using COPA [Li et al., 2021]) evaluated in other papers that worked with the same sizes of BERT. This suggests that, while showing promising results in some causal tasks, BERT still struggles to generate its own causal relations at a larger scale. However, this ability could be increased by using a larger version of the model.

4.2.2 Text Generation

In these experiments, we used 23 050 different prompts and were able to generate up to 137 560 unique causal relations. We go into more detail about the distribution and form of these relations in section 4.3.

In our text-generation experiments, we find that the ability of an LLM to generate causal relations correlates with the size of the LLM and the number of shots used in the prompt. As shown in table 4.3 we see that in terms of precision and recall 3-shot outperforms 1-shot and larger variants of GPT2 outperform smaller variants in every instance. These findings match the results of Hobbhahn et al. [2022] and show that they still hold true when prompting at a larger scale.

model	method	micro-precision	macro-precision	recall
GPT2	1-shot	3.83%	3.72%	1.29%
	3-shot	7.63%	5.59%	3.60%
GPT2-medium	1-shot	6.25%	5.22%	2.15%
	3-shot	11.05%	8.04%	4.86%
GPT2-large	1-shot	7.75%	6.28%	3.02%
	3-shot	12.55%	8.53%	5.25%
GPT2-xl	1-shot	9.44%	7.41%	3.65%
	3-shot	**13.93%**	**9.38%**	**6.14%**

Table 4.3: Results of the Text-Generation method using GPT2, comparing micro-precision, macro-precision, and recall by model size and method.

Looking at micro- and macro-precision we find that the micro-precision is always higher, similar to the results of the mask-filling. This shows us that GPT2 was more likely to repeat relations also present in the CauseNet, matching the correlation shown in section 4.1.

These experiments showed that GPT2 has the ability to generate plausible causal relations. The GPT2-xl variant achieved the highest precision with close to 14% in the 3-shot setup, as well as a recall of just over 6%. At first glance, these results appear quite low when compared to other experiments done in similar papers like Long et al. [2023] or Hobbhahn et al. [2022]. In these papers, GPT3 reached accuracies of over 90% in tasks of causal classification. However, in those papers, the size of the models and the nature of the tasks differed greatly from the ones used in this thesis. However comparing our results to experiments that are similar in task and model size, we find that our results more closely align with the previous research. In a paper by Li et al. [2020] GPT2 achieved a precision of 8% at the task of generating possible effects corresponding to causes. This number was reached by manually evaluating its small sample instead of large-scale automatic parsing in our case.

4.2.3 Constructing Small Causal Graphs

In this setup, we evaluated the ability of GPT2 to reconstruct a smaller causal graph. For this we used the collection of 4 small causal graphs from a paper by Long et al. [2023] containing 18 causal relations in total as seen in figure 4.2. In our first approach, we employed the exact same method used in text generation to construct prompts and generate and parse outputs, using the smaller graphs in place of the CauseNet as ground truth. This unsurprisingly yielded a much lower precision of only 0.41%, in part because the fewer relations are contained in the ground truth, the lower the likelihood of GPT2 generating them. The

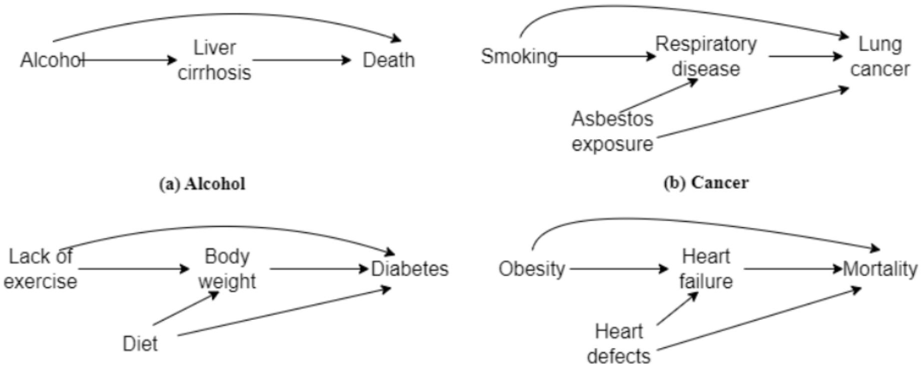


Figure 4.2: The 4 small causal graphs used in Long et al. [2023].

opposite was true for recall. GPT2 was able to generate 17 out of the 18 causal relations at least once, yielding a recall of 94.44%. This is likewise unsurprising due to the small size of the new graph. This first approach yielded an accuracy of only 0.41%, showing that the same exact text generation method was ineffective at reconstructing smaller causal graphs.

In our second approach we altered the text-generation method slightly by implementing the idea of comparing positive to negative statements as used by Long et al. [2023]. Instead of taking all generated relations to be part of a new causal graph that is then compared to the ground truth we use the generated relations to determine if there is a causal link between two concepts from the ground truth or not. This way of using generated relations to help answer binary questions yielded a much better result. The second approach reached an accuracy of 77.19%, outperforming the base case in the paper by Long et al. [2023] (66.70%). This is very impressive considering we used a smaller model and error-prone automatic parsing.

Our results show that a slightly adjusted method of text generation can be used to achieve considerably improved performance in the task of reconstructing small causal graphs. Comparing this result to the results of the text-generation method, we see that using the output of GPT2 to answer binary causal questions yields better results than simply using all causal relations generated by GPT2 to construct a causal graph. This ability to answer binary causal questions matches previous research. We showed that our metric of count, when used to determine the answer to these binary questions, can achieve comparable performance.

4.3 Text Parsing

As we have shown in this thesis we probed LLMs for causal relations at a large scale. Experiments using perplexity or mask filling are easily scaled up, owing to the predictable form of the output that they give. The same is not true for text generation, however, necessitating another step to the pipeline in which the outputs of the text generation would be parsed into causal relations. In other studies, this was done by human experts for a small sample of the output for the purposes of benchmarking performance. This human expert approach was unfeasible at our scale, so we instead chose to automate the parsing of the outputs.

However, the automation of the parsing step runs the risk of introducing additional error. To decrease this error and improve our parsing algorithm we analyzed the types of outputs most commonly produced by the LLM. We differentiate between outputs that followed the pattern of the n-shot prompt (further divided into the categories of "Baseline", "Copied Effects" and "Ignored Cause") and outputs that did not follow the n-shot pattern ("Free-form Outputs"). Outputs that follow a predictable pattern such as "x causes y" are naturally easier to parse correctly. In the following, we included a description of each of those categories.

Breaking down the results of the text generation by these categories (table 4.4) shows that outputs following the n-shot pattern performed more than two times better than non-pattern outputs at generating relations of the CauseNet. We found that the ratio of free-form to patterned outputs appears not to be correlated to model size, however, larger versions of GPT2 show a higher precision when compared to the CauseNet. This implies that larger models have to rely less on the pattern to produce plausible causal claims. We also found that, just as expected, going from a 1-shot to a 3-shot prompt increased the number of relations that follow the pattern, making them easier to parse and more likely to appear in the CauseNet, increasing overall performance.

Baseline: Outputs are considered part of the baseline if they followed the structure of the n-shot example and are not part of any other category. Each baseline output consists of a short sentence containing a cause and an effect linked with a linking word implying a causal link. Following this pattern makes them very easy to parse automatically. Over half of the relations generated by the LLM are baseline relations and around 14.85% of those can also be found in the CauseNet.

Example:
Prompt: "death causes x. what could x be?"
Answers: "1. death causes sickness.

Type of Output	Total	% of Total	Total in CauseNet	% of CauseNet Overlap	precision
All	163 179	100.00%	19 892	100.00%	12.19%
Baseline	95 337	58.42%	13 344	67.08%	14.85%
Copied Effect	11 850	7.26%	4 169	20.96%	35.18%
Ignored Cause	23 088	14.15%	5 098	25.63%	22.08%
Free-form Output	35 498	21.75%	2 114	10.63%	5.96%

Table 4.4: A breakdown of the output of 3-shot text generation using GPT2-xl, broken into the different output patterns. Columns 2 and 3 show this breakdown in regards to all generated relations, columns 4 and 5 in regards to only relations that were generated that were also present in the CauseNet and column 6 shows the precision when compared to the CauseNet.

2. death causes a lack of assets.
3. ..."

Copied Effects: If a causal relation produced by the LLM contains a concept that also appears in the n-shot example, it is considered a copied effect. The effects getting copied tend towards more general concepts, though this is not always the case. It is likely that the inclusion of these relations in the n-shot example influences the LLM into repeating them. This is reflected in the fact that relations with copied effects are three times as likely to also appear in the CauseNet when compared to the baseline. This could be because the LLM tends to copy common and general concepts more. These are also more likely to be the effect of another relation as well, or it could simply be because the example effects are taken from the CauseNet, therefore increasing the likelihood that the whole relation also appears in it. However, this is not the case for all relations with copied effects since the same relation will be generated by a different prompt not containing the effect in its example.

Example:

N-shot example: "poor health causes x. what could x be?
a:
poor health causes poverty.
poor health causes lack of concentration."

Prompt: “death causes x. what could x be?”
Answers: “(1) death causes a lack of concentration.
(2) poor health causes poverty.”

In the example we see two answers to the question in the prompt, one repeating the effect from the n-shot example (1) and the other copying both cause and effect from the n-shot example (2). Relations that were taken as a whole from the n-shot example were not counted as relations claimed by the LLM to avoid our data being skewed by the LLM just copying our input.

Ignored Causes: Outputs are considered to have ignored the cause if it provided a causal relation in the correct form but that relation had a different cause than the one asked for in the question of the prompt. A lot of the time the LLM will replace the cause in the question with one that is broader or more generic. Despite not answering the exact question asked in the prompt, these outputs still contain a plausible causal claim. Relations with a copied cause have a higher likelihood of also occurring in the CauseNet than in the base case. This is likely because when introducing its own causes the LLM tends towards broader, more general concepts, which are likely to also appear in the CauseNet.

Example:
prompt: “bpa exposure causes x. what could x be?”
answer: “stress causes complications.”

Free-form Outputs: A considerable fraction of the generated relations do not semantically follow the structure demonstrated in the n-shot example. Despite the large number of free-form outputs, less than 6% of them also occurred in the CauseNet, making it the lowest-performing category of output in this regard. This is likely in part due to the LLM having difficulty following the question at all if it does not follow the pattern, but also in part due to inaccurate parsing of the results. Many of the outputs that diverge in this way fit into one of two categories: (1) Plausible claims to causal relations that are simply stated in a different natural language way than the n-shot example and (2) generated text that has nothing to do with or makes no statements about any causal relations (nonsense).

Example:
prompt: “death causes x. what could x be?”
answers: “(1) lack of time or energy
(2) when the answer to each of the above three questions is to be found in the answer to a fourth question exercise”

	precision	**recall**	**f1-score**
(a) all free-form are valid	12.19%	**6.14%**	8.17%
(b) free-form are nonsense	**15.94%**	4.97%	7.58%
(c) nonsense-filter	13.93%	**6.14%**	**8.52%**

Table 4.5: Results of different approaches of parsing Free-form Outputs from 3-shot text generation using the GPT2-xl model. Precision and f1-score here refer to micro-precision and micro-f1-score.

	Actual: valid	**Actual: nonsense**
Predicted: valid	**73**	27
Predicted: nonsense	32	**73**

Table 4.6: Confusion matrix of the nonsense filter algorithm with an overall accuracy of 71.21%.

In this example we see that the answer given in (1) does not follow the "x causes y." structure. Nonetheless, it could plausibly be an answer to a causal question. Many free-form answers given, come in the form of a single concept or a comma-separated list of concepts. Since the question is asking for the LLM to fill in the effect of a causal relation, simply stating a list of potential effects is a valid way to answer it.

The same cannot be said for the answer given in (2). In this case, a human reader would easily come to the conclusion that the generated text contains no answer to the question "what could x be?". However, automatically differentiating between answers of categories (1) and (2) is not trivial. For this, we devised an algorithm (nonsense filter) to automatically tag free-form answers as either (1) valid or (2) nonsense, which would not be counted towards the answers given by the LLM. Table 4.5 compares this algorithm to two naive tagging algorithms that would tag all free-form answers as either all valid (a) or all nonsense (b). We can see that the nonsense filter outperforms the naive algorithms in (a) and (b) in terms of f1-score.

As described in Chapter 3 the nonsense filter tags output as "valid" or "nonsense" based on their occurrence in the High-Recall-CauseNet. To evaluate how accurate this method was at tagging the output, we manually evaluated 205 free-form outputs and compared that to the prediction made by the nonsense filter (table 4.6). This yielded an accuracy of 71.21%. This was better than random chance, but still a far cry from having the output tagged manually. The large number of outputs that needed to be tagged however made automation necessary, even at the cost of potentially lowering overall accuracy.

Chapter 5

Conclusion

In this thesis, we set out to evaluate how well LLMs can internalize and reproduce causal knowledge. In the related works chapter, we discussed the different kinds of causal tasks tested on LLMs in related research. We found that the performance of LLMs at tasks of causal inference is well studied, but current research into this topic most commonly relied on small-scale sample sets for their evaluation. With the experiments of this thesis, we aimed to build on the different approaches seen in the related works and probe for causal knowledge at a much larger scale. For this we heavily relied on the CauseNet, a large-scale knowledge base containing over 198 000 claimed causal relations. These relations were extracted from a large English language corpus based on internet pages and are thus comparable to the training data of the LLMs we used. We used a subset of around 12 000 relations to prompt LLMs to create over 700 000 of their own causal claims.

First, we must acknowledge the limitations of our research. In this thesis, we worked under the assumption that the CauseNet contained similar causal relations to the large texts the LLMs were trained on. This is likely true to some degree, however, it might not be enough to accurately estimate how well an LLM learns causal knowledge from its training data. For the correlation to be as representative as possible of the LLM's ability to learn causal knowledge, we would need to use an LLM pre-trained using the Clueweb12 web crawl. However, creating such an LLM was outside the scope of this thesis. Additionally, probing for causal knowledge at a larger scale through text generation introduced the difficulty of automatically parsing a varied output. While we tried to filter out nonsense and incentivize adherence to a known output pattern, this process was not perfect leaving some degree of error in our measurements.

For our experiments, we employed three different approaches: perplexity, mask-filling, and text generation.

In our perplexity approach (section 3.1), we looked into using the calculated perplexity to evaluate an LLM's confidence in a causal relation. This, however, returned mixed results. We found that perplexity as a metric is heavily influenced by contextual factors of a causal sentence like word count and starting word. This means that perplexity on its own was found to be an unreliable indicator for gauging the causal knowledge of an LLM. Controlling for these factors allowed for clearer results, but also heavily restricted the kinds of relations that could be compared. We suspect that an approach that normalized perplexity in regards to word count and starting word could be more significant as a metric for the causal knowledge of an LLM and further research could be done to test this.

Our approaches using mask-filling and text generation yielded slightly more promising results. We found that through these approaches off-the-shelf versions of GPT2 and BERT were able to demonstrate a rudimentary ability to answer open-ended causal inference questions by learning and reciting knowledge contained in their training data. The LLMs were much more likely to repeat causal claims that showed up in the CauseNet. We also found a small but significant correlation between the LLMs' confidence in a causal claim and the support in the CauseNet (see section 4.1). This means that the LLMs were more likely to generate and repeat relations with high support in the CauseNet, showing that the CauseNet support is a useful indicator for the causal knowledge of LLMs.

We found that in all of our experiments, larger variants of the LLMs consistently outperformed smaller variants. In our experiments, GPT2-xl was able to reach a micro-precision of 13.9% and a recall of 6.1% in the task of text generation while BERT-large reached a precision of 9.4% and a recall of 40.9% in the mask-filling approach. In addition to the size of the LLM, we also found that the number of shots in the n-shot setup improved the LLM's performance.

Adapting the text-generation approach for the task of recreating causal graphs also returned promising results. The accuracy of 77% observed in our experiments outperformed the results of previous research on the same task with GPT3 [Long et al., 2023]. This is a surprising outcome considering we used GPT2 compared to GPT3, which is much larger. This shows that a large-scale text generation approach has promising potential in aiding the construction of causal graphs. Judging from the results of our experiments we predict that larger models like GPT3 would perform even better using this approach.

Bibliography

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Oktie Hassanzadeh, Debarun Bhattacharjya, Mark Feblowitz, Kavitha Srinivas, Michael Perrone, Shirin Sohrabi, and Michael Katz. 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, 7 2019. doi: 10.24963/ijcai.2019/695. URL `https://doi.org/10.24963/ijcai.2019/695`.

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

Marius Hobbhahn, Tom Lieberum, and David Seiler. Investigating causal understanding in LLMs. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*, 2022. URL `https://openreview.net/forum?id=st6jtGdW8Ke`.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs, 2021.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation?, 2023.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality, 2023.

Zhongyang Li, Xiao Ding, Ting Liu, J. Edward Hu, and Benjamin Van Durme. Guided generation of cause and effect. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, jul 2020. doi: 10.24963/ijcai.2020/502. URL `https://doi.org/10.24963%2Fijcai.2020%2F502`.

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

Stephanie Long, Tibor Schuster, Alexandre Piché, Department of Family Medicine, McGill University, Mila, Université de Montreal, and ServiceNow Research. Can large language models build causal graphs?, 2023.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. The science of detecting llm-generated texts, 2023.

Ruibo Tu, Chao Ma, and Cheng Zhang. Causal-discovery performance of chatgpt in the context of neuropathic pain diagnosis, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

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. Huggingface's transformers: State-of-the-art natural language processing, 2020.

Cheng Zhang, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, and James Vaughan. Understanding causality with large language models: Feasibility and opportunities, 2023.