

Retrieval Augmented Generation based Large Language Models for Causality Mining

Thushara Manjari Naduvilakandy, Hyeju Jang, Mohammad Al Hasan*

Department of Computer Science

Indiana University Indianapolis

tnaduvil@iu.edu, hyejuj@iu.edu, alhasan@iu.edu

Abstract

Causality detection and mining are important tasks in information retrieval due to their enormous use in information extraction, and knowledge graph construction. To solve these tasks, in existing literature there exist several solutions—both unsupervised and supervised. However, the unsupervised methods suffer from poor performance and they often require significant human intervention for causal rule selection, leading to poor generalization across different domains. On the other hand, supervised methods suffer from the lack of large training datasets. Recently, large language models (LLMs) with effective prompt engineering are found to be effective to overcome the issue of unavailability of large training dataset. Yet, in existing literature, there does not exist comprehensive works on causality detection and mining using LLM prompting. In this paper, we present several retrieval-augmented generation (RAG) based dynamic prompting schemes to enhance LLM performance in causality detection and extraction tasks. Extensive experiments over three datasets and five LLMs validate the superiority of our proposed RAG-based dynamic prompting over other static prompting schemes.

1 Introduction

The expression of causality, along with causal reasoning, is a defining characteristic of human cognition (Goddu and Gopnik, 2024), which makes natural language text a trove of causal knowledge. Extracting this knowledge from text is a significant task due to its widespread applications in natural language processing (NLP), including question answering (Khoo et al., 1998; Girju, 2003), event prediction (Radinsky et al., 2012; Silverstein et al., 2000) and medical text mining (Hashimoto et al.,

2014; Riaz and Girju, 2010). The importance of causality extraction task has intensified even further with the proliferation of AI Chatbots and conversational AI agents, which rely on synthesized causal knowledge. Therefore, sophisticated and scalable models are essential to accurately extract causal knowledge from various textual sources, ranging from scientific documents to news articles.

Causality in natural language text refers to the relationship between two entities in a sentence, where the occurrence of one entity (the cause) leads to the occurrence of the other (the effect). There are two prominent research tasks related to causality: causality detection (Blanco et al., 2008; Hidey and McKeown, 2016; Kayesh et al., 2019) and causality extraction (Dasgupta et al., 2018; Li et al., 2021; Kabir, 2023). The first task detects whether a given sentence is causal or not, which is typically framed as a binary classification task. The second task labels the cause and the effect phrases within a given sentence, given that it is causal. This task is often formulated as a sequence labeling task, where each token in the sentence is labeled as part of a cause phrase, effect phrase, or neither.

In traditional machine learning, many unsupervised (Kaplan and Berry-Rogge, 1991; Garcia et al., 1997; Girju et al., 2002) and supervised (Li et al., 2021; Dasgupta et al., 2018; Kabir, 2023) methods are proposed for causality mining. Among these, the unsupervised methods require significant human intervention, and the supervised methods suffer from the lack of large training datasets. With the emergence of pre-trained large language models (LLMs) such as GPT (Brown, 2020; Achiam et al., 2023), Llama (Touvron et al., 2023), Gemini (Anil et al., 2023), and Mixtral (Jiang et al., 2023), many supervised NLP tasks are now being solved using LLMs, primarily due to their ability to produce superior results only with minimal supervision. This is exciting news for causality mining, as supervised models for causality mining, trained on

* Corresponding author.

Code: https://github.com/thusharamanjari/RAG_LLM_CausalityMining

datasets from one domain, often perform poorly on datasets from different domains. Thus, LLMs with few-shot learning capabilities could be an attractive approach for causality mining.

Performance improvements of LLMs in specific tasks like causality mining can be achieved through fine-tuning, prompt engineering, and Retrieval-Augmented Generation (RAG) approaches. Fine-tuning methods train weights of a few terminal layers of the LLMs, prompt engineering selects effective in-context examples for few-shot training of LLMs, and RAG improves LLM outputs by generating prompt dynamically by leveraging relevant knowledge from external sources. For causality mining using LLMs, Jin et al. (2023) proposed to improve LLM performance in causal inference by fine-tuning, but their results show that fine-tuned models fail to generalize, yielding poorer results on out-of-distribution data. Besides, fine-tuning large language models requires significant resources and diverse training data as they are not good for incorporating dynamic knowledge based on varying scenarios. These limitations of fine-tuning open the frontier of research in prompt engineering and RAG to improve LLM performance.

In this paper, we propose two novel RAG (Retrieval-Augmented Generation) approaches namely **Pattern RAG** and **kNN+Pattern RAG**, for solving the causality mining task using LLMs. Both the approaches leverage a fewshot example repository—an external resource which we build (offline) for supporting our proposed RAG approaches. The fewshot example repository contains a collection of causal sentences with the cause and the effect phrases correctly tagged. Each of these sentences in this repository are also indexed by a causal connective. During causal inference using an LLM, given an input sentence both the RAG approaches judiciously retrieve relevant examples dynamically from the fewshot example repository, and augment the causality detection LLM prompts with these examples for improving LLM performance. We conduct extensive experiments on three datasets and five different LLMs to validate the performance of our proposed RAG approaches against traditional supervised ML approaches and other LLM prompting approaches, namely Zeroshot, Random fewshot and kNN RAG. Experimental results show that our approaches outperform the baseline methods for both causality detection and causality extraction tasks.

Input Sentence	Causal Connectives
However, as illustrated by these and other cases reported to date, the onset of <cause> troglitazone </cause> -induced <effect> liver injury </effect> is insidious and temporally variable.	<i>induced</i>
When a <effect> tsunami </effect> is generated by a strong offshore <cause> earthquake </cause>, its first waves would reach the outer coast minutes after the ground stops shaking.	<i>is generated by</i>
Highly viscous <cause> lavas </cause> lead to a violent <effect> eruption </effect>.	<i>lead to</i>

Table 1: Causal sentences from the fewshot example repository with LLM generated causal connectives.

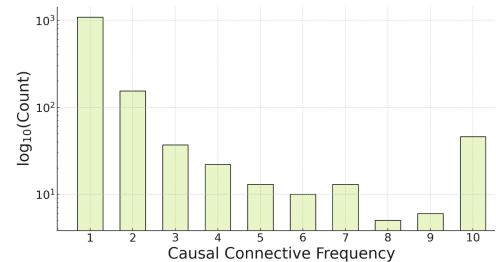


Figure 1: Logarithmic of the number of causal connectives in each frequency category ranging from 1 to 10.

2 Related Works

Causality mining research in the past few decades can be broadly categorized into three main approaches: rule-based (non-statistical), traditional machine learning-based, and deep learning-based. Rule-based approaches (Kaplan and Berry-Rogghe, 1991; Joskowicz et al., 1989; Kontos and Sidiropoulou, 1991; Garcia et al., 1997; Girju et al., 2002; Khoo et al., 1998; Sadek, 2013) mainly use different linguistic patterns and causality connectives (e.g., “caused”, “lead to” and “triggered”) to detect causality. These methods require significant human intervention for identifying the rules. Often, rules extracted from one domain may not work for another domain as effectively. Traditional machine learning approaches (Girju et al., 2002; Bethard and Martin, 2008; Pal et al., 2010; Sorgente et al., 2013; Pakray and Gelbukh, 2014) are more advantageous than rule-based approaches as they require less human effort. These methods automate the pattern extraction process by utilizing NLP tools like WordNet, Google N-grams, POS tagging, and apply them in models like decision trees, SVM, and Naive Bayes for causality extraction.

The advent of deep learning models and the attention mechanism help overcome the feature spar-

Freq.	Five Causal Connective sample
1	generates, probable cause of, that caused the, as the consequence of, after the treatment by
2	led to, can lead to, may be induced by, effect, root causes of
3	were caused by, derived from, experienced, was generated from, after initiation of, adverse effect of
4	caused the, after the use of, association of, causing, induces
5	can induce, that resulted in, leading causes of, leading to, was the cause of
6	leads to, because of, ensued from, produces, can cause
7	most common cause of, association between, as a complication of, side effect of, instigated by
8	result in, created by, as a result of, radiated, emits
9	resulting from, results in, had caused, radiating from, had caused
10	triggered by, induced by, arose from, is one of the main causes of, -associated

Table 2: Causal connective examples for each frequency category

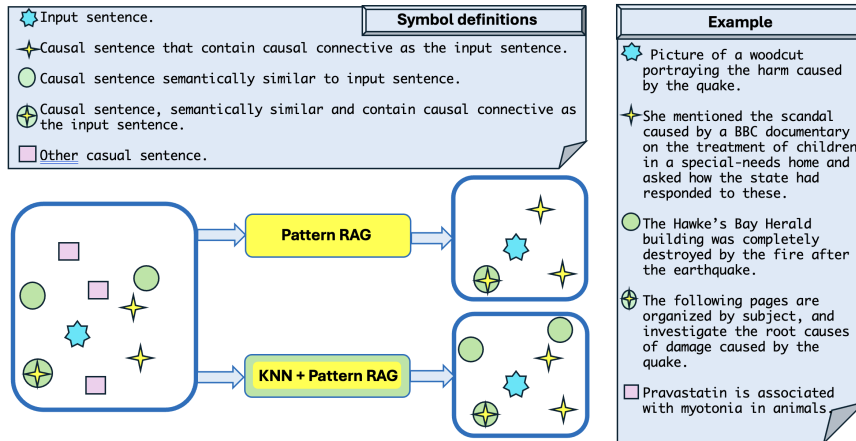


Figure 2: Dynamic fewshot selection mechanism of Pattern RAG and k NN+Pattern RAG. An example sentence from SemEval dataset and related sentences that are selected by different RAG schemes are also shown.

sity problem and made the training more effective (De Silva et al., 2017; Ponti and Korhonen, 2017; Dasgupta et al., 2018; Wu and He, 2019; Zhao et al., 2019; Ali et al., 2021; Tian et al., 2022; Lan et al., 2023; Chen and Li, 2021). Prior studies use various deep learning architectures, including Self-attentive BiLSTM-CRF with Transferred Embeddings (SCITE) (Li et al., 2021), linguistically informed Bi-LSTM model (Dasgupta et al., 2018), bidirectional LSTM with a CRF layer (Bi-LSTM-CRF) (Huang et al., 2015) and dependency-aware transformer based model (DEPBERT) (Kabir, 2023). However, achieving better performance with deep learning models requires high computational resources and a large training corpus, preferably from the same domain, which poses a significant obstacle. The advance of LLMs and prompt engineering helps to overcome the limitations of deep learning models. Zhang et al. (2024) utilize the idea of RAG-based LLMs to deduce causal relationships from a large corpus of scientific literature for the causal graph discovery task.

Liu et al. (2021) introduce the idea of selecting dynamic few-shot examples for a test instance

based on similarity measures. They experiment these k NN based in-context example selection method for sentiment classification, table-to-text generation, and question answering tasks. Similar approaches are also used for relation extraction tasks (Nori et al., 2023; Liu et al., 2023; Efeoglu and Paschke, 2024). We make minor modification to Liu et al. (2021) to adapt it our task and use it as one of the baseline.

k NN-based method chooses semantically similar examples, but they are not always helpful for cause effect phrase extraction task. Specifically, when the phrases are not identified by the semantic meaning of entire sentence, but by causal connectives, such a method performs poorly. Wang et al. (2023) utilizes a retriever trained on a labeled dataset, which involves a costly training phase. In contrast, our pattern-based method does not require such a retriever, offering a more efficient and lightweight alternative. Zhang et al. (2022) uses reinforcement learning to iteratively train a policy, which demands significant time and computational resources. Li et al. (2023) proposes a Unified Demonstration Retriever (UDR) that retrieves examples across

diverse tasks using a multi-task listwise ranking framework, reducing storage and deployment costs compared to task-specific retrievers. These works have explored the in context example selection methods but haven’t explored the causality mining tasks. Since cause-effect phrase extraction is a sequence labeling task, incorporating these baselines requires to adapt these methods substantially, which in-itself can be a new research direction.

3 Methodology

Inspired by recent advancements in Retrieval-Augmented Generation (RAG) demonstrating the efficacy of dynamically selected few-shot examples for enhancing LLM performance (Liu et al., 2021), we propose a novel approach to causality mining that utilizes an external knowledge base to augment LLM prompts. We construct a concise knowledge base of causal examples, and develop two RAG-based methods—**Pattern RAG** and **k NN+Pattern RAG**—to dynamically select relevant instances for a given task. These approaches retrieve in-context examples from our fewshot example repository by leveraging causal connectives and sentence embedding similarity.

3.1 Fewshot Example Repository Creation

The fewshot example repository is created off-line, which is to be used as an external source to support retrieval-augmented prompt generation with a large language model. The desiderata of this repository, which stores example causality sentences are as follows: (1) it should contain examples similar to any test instance provided by a user; (2) the examples should be domain-neutral; and (3) they should be concise.

To fulfill these requirements, we first combine all the causal sentences from the training datasets of SemEval, ADE and Li et al. (4,082 sentences in total). Each sentence is then fed into GPT-3.5-turbo with an ICL prompt to identify the causal connectives. Some sample sentences with their identified causal connectives (from GPT-3.5-turbo) are shown in Table 1.

The sentences are then indexed by the causal connectives identified by GPT, and stored in a data structure, called *Fewshot Example DB*. Since our experiment does not require many examples, and also to make the repository memory efficient, we keep up to 10 random examples per unique causal connective. Note that some example sentences are

shared by multiple causal connectives, as GPT does not always extract the exact casual connectives; for instance, it may extract “caused by” in one instance and “caused by the” in another, resulting in two different connectives. At the end, our *Fewshot Example DB* contains 2,365 instances indexed by 1,394 unique causal connectives, with 80 connectives having at least 5 examples, while the rest occur 1 to 5 times. Figure 1 shows the causal connectives count in each frequency category of *Fewshot example DB*. We manually verify the quality of causal connectives by randomly selecting five causal connective from each frequency category, ranging from 1 to 10, in our repository. This process is repeated over several iterations, and one of the results is shown in Table 2.

3.2 RAG Prompts for Causality Mining

LLMs offer powerful language understanding and generation capabilities, but they are not reliable information sources and they lack access to information beyond their training data. Retrieval-augmented generation enables LLMs to witness in-context examples which are highly relevant to the given task. These in-context examples are chosen by RAG approaches. In this work, for the causality extraction task, we propose two RAG ideas: Pattern RAG and k NN+Pattern RAG.

Pattern RAG: This is our first RAG scheme, where the fewshot examples are chosen by matching the causal connective of the input sentence with those of the sentences in the textitFewshot Example DB. For example, for input sentence “fever is caused by flu”, fewshot examples in the repository which has “cause by” connective are chosen for prompt augmentation. Since causal connective detection is not always exact, we select those sentences for which the causal connective is more than 90 percent similar to the causal connective of the input. If we find more than 10 matched examples, we randomly choose 10 of them and filter out the rest. This Pattern RAG scheme gets similar fewshot examples based on the presence of similar causal connectives in the examples. Our hypothesis is that, with these fewshot examples, LLM will be able to align the causal connective between the input sentence and the fewshot example, and be able to identify the cause and effect phrases more effectively. In Figure 2, the input sentence has the “caused by” pattern, the retrieved fewshot examples also have “caused by” in them (see the sentence tagged with a yellow star).

LLM	Method	SemEval				ADE				Li et al. dataset			
		Acc.	F1	P	R	Acc.	F1	P	R	Acc.	F1	P	R
GPT-3.5-turbo	Zeroshot	0.68	0.73	0.60	0.93	0.76	0.81	0.72	0.93	0.65	0.57	0.40	0.95
	Random fewshot	0.85	0.84	0.83	0.85	0.73	0.67	0.97	0.51	0.89	0.72	0.86	0.63
	<i>k</i> NN RAG	0.84	0.81	0.88	0.75	0.85	0.84	0.92	0.78	0.91	0.84	0.77	0.93
	Pattern RAG	0.86	0.83	0.90	0.78	0.86	0.85	0.96	0.77	0.93	0.86	0.85	0.87
	<i>k</i> NN+Pattern RAG	0.86	0.84	0.89	0.80	0.88	0.88	0.92	0.85	0.92	0.85	0.76	0.96
GPT-4o	Zeroshot	0.85	0.83	0.83	0.84	0.81	0.81	0.89	0.75	0.89	0.79	0.72	0.88
	Random fewshot	0.87	0.84	0.90	0.79	0.85	0.84	0.95	0.76	0.91	0.82	0.86	0.78
	<i>k</i> NN RAG	0.90	0.88	0.92	0.84	0.85	0.84	0.95	0.76	0.96	0.92	0.87	0.97
	Pattern RAG	0.88	0.86	0.89	0.84	0.89	0.89	0.95	0.84	0.94	0.88	0.85	0.92
	<i>k</i> NN+Pattern RAG	0.91	0.90	0.92	0.87	0.90	0.90	0.96	0.84	0.96	0.93	0.89	0.98
Llama3-8b	Zeroshot	0.79	0.79	0.70	0.91	0.83	0.85	0.78	0.93	0.73	0.62	0.47	0.94
	Random fewshot	0.78	0.74	0.78	0.71	0.87	0.88	0.87	0.90	0.84	0.72	0.63	0.83
	<i>k</i> NN RAG	0.72	0.75	0.62	0.96	0.74	0.80	0.69	0.95	0.57	0.53	0.36	0.98
	Pattern RAG	0.83	0.83	0.79	0.86	0.88	0.89	0.85	0.93	0.83	0.73	0.59	0.94
	<i>k</i> NN+Pattern RAG	0.82	0.82	0.77	0.87	0.81	0.83	0.80	0.86	0.74	0.64	0.48	0.95
Gemma2-9b-it	Zeroshot	0.81	0.77	0.82	0.73	0.67	0.59	0.88	0.45	0.89	0.77	0.79	0.75
	Random fewshot	0.79	0.75	0.84	0.67	0.71	0.66	0.91	0.52	0.88	0.73	0.79	0.68
	<i>k</i> NN RAG	0.83	0.80	0.90	0.71	0.60	0.43	0.86	0.28	0.92	0.85	0.85	0.85
	Pattern RAG	0.80	0.76	0.85	0.69	0.64	0.51	0.89	0.36	0.91	0.80	0.85	0.76
	<i>k</i> NN+Pattern RAG	0.83	0.81	0.84	0.78	0.65	0.56	0.85	0.42	0.91	0.84	0.78	0.90
Mixtral-8x7b	Zeroshot	0.79	0.78	0.72	0.86	0.78	0.80	0.78	0.82	0.76	0.64	0.50	0.88
	Random fewshot	0.75	0.69	0.77	0.63	0.72	0.70	0.84	0.60	0.81	0.64	0.62	0.65
	<i>k</i> NN RAG	0.81	0.80	0.79	0.81	0.78	0.78	0.84	0.73	0.84	0.75	0.61	0.96
	Pattern RAG	0.83	0.81	0.78	0.85	0.80	0.81	0.81	0.81	0.82	0.72	0.58	0.93
	<i>k</i> NN+Pattern RAG	0.80	0.80	0.71	0.92	0.81	0.84	0.77	0.92	0.75	0.66	0.49	0.99

Table 3: Causality detection results.

***k*NN+Pattern RAG:** In this RAG scheme we first identify causal sentences that are semantically similar to the input. We use “text-embedding-ada-002” model to obtain vector representations of the input sentence (test instances) and all example instances in our Fewshot Example DB. The *k*-nearest neighbor search algorithm is applied to find the 10 most similar examples to the given test instance from the DB. For *k*NN+Pattern RAG, we concatenate the 10 examples identified by *k*NN RAG and those retrieved by Pattern RAG, resulting in a total of 20 examples for fewshot prompt augmentation. *k*NN RAG retrieves 10 examples that are semantically similar to the input sentence, while Pattern RAG retrieves all the examples that show a causality relation with the same causal connective present in the input sentence. Our hypothesis is that this combination will help the LLM by leveraging both sentential semantics and causal pattern syntax. Figure 2 provides a pictorial depiction of both our RAG approaches.

4 Experiments

We test our proposed RAG approaches with five different LLMs on three datasets, comparing our results for two tasks against competing methods. The first task is causality detection, a binary classification task. The input to this task consists of sentences labeled as 1 (causal sentence) and 0 (non-causal sentence). The LLM outputs a response of 1 if the sentence has a causality relation, otherwise it outputs 0. For evaluation, We use standard classification evaluation metrics: accuracy, F1 score, precision and recall.

The second task is causality extraction, a sequence labeling task where the LLM identifies cause and effect phrases from the input sentence. The labeled dataset in our experiments typically marks single words as cause and effect. However, in real-world scenarios, more detailed phrase representations of cause and effect are often more informative than single words. To accommodate this, our designed prompts guide the LLM to extract

LLM	Method	SemEval	ADE	Li et al. dataset		
		Accuracy	Accuracy	F1	Precision	Recall
DEPBERT	Supervised	0.65	0.61	0.12	0.12	0.13
GPT-3.5-turbo	Zeroshot	0.85	0.75	0.54	0.57	0.52
	Random fewshot	0.83	0.79	0.47	0.56	0.40
	k NN RAG	0.91	0.83	0.46	0.60	0.38
	Pattern RAG	0.90	0.80	0.52	0.61	0.46
	k NN+Pattern RAG	0.91	0.82	0.51	0.62	0.44
GPT-4o	Zeroshot	0.82	0.66	0.76	0.79	0.73
	Random fewshot	0.82	0.77	0.75	0.77	0.73
	k NN RAG	0.87	0.79	0.77	0.83	0.72
	Pattern RAG	0.89	0.83	0.80	0.83	0.78
	k NN+Pattern RAG	0.89	0.84	0.75	0.78	0.73
Llama3-8b	Zeroshot	0.77	0.75	0.65	0.84	0.53
	Random fewshot	0.76	0.78	0.65	0.85	0.52
	k NN RAG	0.83	0.76	0.64	0.82	0.52
	Pattern RAG	0.83	0.77	0.64	0.83	0.53
	k NN+Pattern RAG	0.83	0.75	0.66	0.83	0.55
Gemma2-9b-it	Zeroshot	0.76	0.76	0.64	0.76	0.55
	Random fewshot	0.77	0.81	0.68	0.83	0.58
	k NN RAG	0.83	0.83	0.64	0.81	0.53
	Pattern RAG	0.83	0.83	0.72	0.82	0.61
	k NN+Pattern RAG	0.87	0.84	0.74	0.90	0.63
Mixtral-8x7b	Zeroshot	0.73	0.71	0.66	0.81	0.56
	Random fewshot	0.74	0.79	0.66	0.84	0.55
	k NN RAG	0.81	0.80	0.62	0.80	0.51
	Pattern RAG	0.81	0.83	0.67	0.86	0.55
	k NN+Pattern RAG	0.80	0.81	0.66	0.80	0.56

Table 4: Causality extraction results.

phrases instead of single words when identifying cause and effect within a sentence. For example, in the input sentence, “*The truck carried homemade weapons, and the blast was caused by the mishandling of weapons,*” the ground truth dataset labels “*mishandling*” as the cause and “*blast*” as the effect. However, with our prompts, the LLM generates “*mishandling of weapons*” as the cause and “*blast*” as the effect. Additionally, we explicitly instruct the LLM in the prompt to ensure there is no overlap between the predicted cause and effect phrases.

To evaluate the performance of causality extraction task, we use different metrics based on whether a sentence has only one cause-effect phrase or multiple cause-effect phrases in the ground truth data. For the SemEval and ADE datasets, each sentence has only a single set of cause-effect phrase; so for these datasets we only check whether the predicted cause and effect phrases match with the ground truth cause and effect phrases, respectively. If the

ground truth word is presented in the predicted phrases (for both cause and effect), we consider the causality extraction task on that sentence a success. Accuracy is simply the fraction of sentences for which causality extraction is a success. Note that, for these two datasets, we add a constraint in the prompt so that LLM extracts only one cause-effect pair from each sentence.

The Li et al. dataset have sentences which have multiple cause-effect phrases, so for this dataset accuracy is hard to produce. So we use precision, recall, and F1 metrics as below. From all the sentences in Li et al., we first create triplets (S, C, E) , each denoting a distinct cause (C) and effect (E) phrase pair in the sentence S . Let the set of these triplets be called the triplet test set (T_t) . Say, T_p is the collection of triplets formed from the LLM for all the sentences. Then, precision of the LLM model for the Li et al. dataset can be calculated as $P = |T_t \cap T_p|/|T_p|$, and recall can be calculated

as $R = |T_t \cap T_p|/|T_t|$. From P and R , we then compute the $F1$ value. While computing intersection between T_p and T_t , a predicted triplet in T_p is considered a match with a ground truth triplet in T_t , when their sentences are identical, and the ground truth cause and effect phrases are present in the predicted cause and effect phrase.

4.1 Baseline Methods

As baseline methods, we use two static prompt strategies: Zeroshot and Random fewshot for both the detection and extraction tasks. And a dynamic in context example selection baseline, k NN RAG. For the causality extraction task, we use the three deep learning based methods Bi-LSTM (Dasgupta), Bi-LSTM-CRF (SCITE) and DEPBERT (Kabir, 2023) which does not use LLM.

Zeroshot: In this baseline approach for both causality detection and extraction, we provide an ICL prompt (shown in Appendix) that combines the concepts of causal connectives with a detailed task explanation for the model. Experiments conducted under Zeroshot setting on different LLMs demonstrate the performance of LLMs on causality mining tasks when no prior examples are available.

Random fewshot: For this baseline approach, we randomly select 10 examples from *Fewshot Example DB* and concatenate them with the Zeroshot setting prompt for both tasks. For causality detection, an input example includes only the sentence with cause and effect phrases whereas for causality extraction, we additionally tag the causal connective of the example. Note that, the prompts and example formats for Random fewshot, k NN RAG, Pattern RAG, and k NN+Pattern RAG are the same, differing only in the choice of example selections. **k NN-RAG:** In existing RAG works [35], examples that are semantically similar to the input are chosen to augment the LLM prompt. We adapt this approach as a baseline for causality mining and call it as k NN RAG. We apply the k -nearest neighbor search algorithm on the vector representations to find the 10 most similar examples to the given test instance from the DB. We then concatenate these examples with our prompts for causality detection and extraction. In Figure 2, we see that for an input sentence related to destruction caused by an earthquake, the k NN RAG selects the examples with similar meanings (shown in green circles).

Dasgupta (Bi-LSTM): Dasgupta et al. (2018) proposed one of the earliest deep neural network-based method for causality extraction. Their approach

combines embeddings from Word2Vec with a linguistic feature vector, which are then input to a bi-directional LSTM (bi-LSTM) model.

SCITE (Bi-LSTM-CRF): Li et al. (2021) proposed a causality extraction model that leverages a BiLSTM CRF back bone, enhanced with Flair embeddings and multihead self-attention.

DEPBERT (Transformer): (Kabir, 2023) DEPBERT is a state-of-the-art transformer-based supervised model that leverages the dependency tree of a sentence on top of a BERT model to extract cause and effect phrases.

4.2 Datasets

For our experiments, we use three well-known causality datasets: SemEval (Hendrickx et al., 2019), ADE (Gurulingappa et al., 2012), and Li et al. dataset (Li et al., 2021). Their detailed discussion is provided in Appendix. Statistics of the dataset is provided in Table 9 (in Appendix). Table 8 and 7 shows few example inputs for causality detection and extraction tasks, respectively.

Methods	Accuracy
Dasgupta (Bi-LSTM)	0.78
SCITE (Bi-LSTM-CRF)	0.78
DEPBERT (Transformer)	0.86
Pattern RAG	0.88
kNN+Pattern RAG	0.90

Table 5: Accuracy comparison of our RAG methods (GPT-3.5-turbo Pattern RAG and k NN+Pattern RAG) with deep learning methods on SemEval dataset.

5 Results

Table 5 shows the comparison of three deep learning-based baseline methods trained on the entire SemEval train dataset with our RAG based methods, evaluated on the SemEval test set as used in the DEPBERT paper. Additionally, the best-performing deep learning method, DEPBERT, has been included in Table 4 to compare its performance against all the LLM results. This DEPBERT result is obtained by training on our few-shot DB and testing on the cleaned test datasets, as detailed in Table 9. Our experiment results for the causality detection and extraction tasks are shown in Table 3 and Table 4, respectively. From these tables, we can see that our RAG-based methods demonstrate significant performance improvements over the baseline methods across all three datasets.

For causality detection, the k NN+Pattern RAG method achieves the best performance when paired

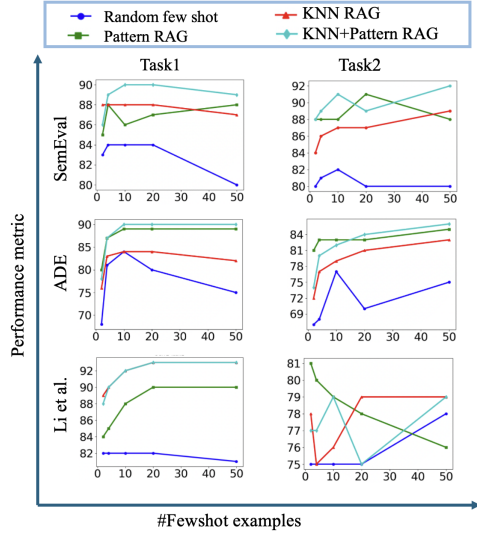


Figure 3: Plot of the number of examples vs. performance results of Causality detection (Task1) and Causality extraction (Task2) from GPT-4o model for different RAG based prompting

with the GPT-4o model. The corresponding best baseline method for GPT-4o across all three datasets is k NN RAG. The F1 improvements by the best of our RAG approach over the best baseline approach for SemEval, ADE, and Li et al. are +2%, +7%, and +1%, respectively.

For causality extraction, the SemEval dataset shows its best performance with the GPT-3.5-turbo model in the k NN RAG and k NN+Pattern RAG settings. The ADE and Li et al. datasets achieve their best performance with the GPT-4o model in k NN+Pattern RAG and Pattern RAG, respectively. Our best-performing RAG model shows no improvement in accuracy over the best baseline (k NN RAG) for the SemEval dataset, maintaining an accuracy of 0.91. However, significant improvements are observed for the ADE and Li et al. datasets. For ADE, the GPT-4o model using k NN+Pattern RAG improves accuracy by +6% (0.84 vs. 0.79). In the Li et al. dataset, the Pattern RAG method boosts F1 by +4% (0.80 vs. 0.77).

6 Performance vs Fewshot Count

In this experiment we validate whether the number of fewshot examples affects the performance of different promptings: Random fewshot, k NN RAG, Pattern RAG, and k NN+Pattern RAG. For obvious reason, Zeroshot is not used as it does not use any example. We show results for GPT-4o as this is the best performing LLM, Figure 3 shows the results. In this figure, we show six panels. In each panel, we show the change in F1 score with respect to the

number of fewshot examples for Random fewshot, k NN RAG, Pattern RAG and k NN+Pattern RAG. For causality detection, the Random fewshot selection approach shows lower F1 scores than those of all three RAG-based methods, regardless of the number of examples chosen. This indicates that the causality detection task benefits solely from the quality of the examples chosen across all three datasets, rather than the number of examples.

For causality extraction, we see that the RAG-based approaches show higher accuracy than the Random fewshot approach for different fewshot example counts in the SemEval and ADE datasets. In the Li et al. plot, we see that the Random fewshot outperforms Pattern RAG at a fewshot example count of 50. This is due to the fact that Pattern RAG dynamically selects examples based on the existence of causal connectives. Our *Fewshot example DB* has a limit of 10 examples for each pattern (causal connectives), so if an instance does not have multiple causal connectives, Pattern RAG cannot obtain more examples. Therefore, when the number of examples is increased to a high value like 50, Pattern RAG does not necessarily has 50 examples, while Random fewshot always have 50 examples.

We conclude from these results that simply increasing the number of examples does not improve LLM performance. The performance increase in the k NN+Pattern RAG in most of the LLM results is solely based on the quality of the examples chosen by concatenation, not on the increase in example count.

7 Causality extraction on multi-word and multi-cause-effect scenarios.

The dataset by Li et al. contains sentences labeled with multiple cause-effect pairs. Table 6 presents a comparative analysis of the transformer-based DEPBERT method and our best-performing Pattern RAG method using GPT-4o. Extracting multiple cause-effect pairs from a sentence poses significant challenges for traditional deep learning methods. However, our proposed LLM-based RAG method demonstrates a notable improvement in performance (F1 score). From the first example, we observe that DEPBERT struggles to predict all causes and effects correctly in a multi-labeled dataset. Our approach shows a good performance improvement but is hindered by the exact-match requirement of the evaluation metric (all words in the label should be present in the predicted phrase). For

Input Sentence	DEPBERT		Pattern RAG GPT 4o	
	Cause	Effect	Cause	Effect
Heat , wind and smoke cause flight delays.	× Heat	flight delays	Heat , wind, smoke	flight delays
Information about the foodborne illness caused by salmonella bacteria .	× by salmonella bacteria	× foodborne illness	salmonella bacteria	× foodborne illness
Eye discomfort from this staring effect is exacerbated by low humidity.	× low humidity	Eye discomfort, staring	× staring effect, low humidity	Eye discomfort

Table 6: Comparitive analysis of DEPBERT and the proposed approach on cause-effect extraction tasks. Cause and effect labels are indicated by blue and red color in the input sentence respectively. × indicates incorrect predictions, while denotes correct ones based on our evaluation metrics.

Dataset	Input Sentence	Cause label	Effect label
SemEval	Dogs develop a fever from stress and/or pain such as in a severe flea infestation.	stress	fever
ADE	Benzocaine-induced methemoglobinemia has been reported in man, dogs, and cats.	Benzocaine	methemoglobinemia
Li et al.	Paralysis or convulsions are caused by hormone deficiencies and imbalances .	– hormone deficiencies and imbalances – hormone deficiencies and imbalances	Paralysis convulsions

Table 7: Input examples for causality extraction.

instance, in example two, the actual effect phrase is "the foodborne illness," while our model predicts "foodborne illness," resulting in a misclassification due to the missing article. Example three highlights a scenario where our model predicts the cause phrase correctly, but the dataset lacks corresponding labels. This analysis shows that the causality extraction for a multi-word and multi-cause-effect scenarios, can be improved by a better evaluation metrics and labeling practices.

8 Conclusion

In this work, we proposed two retrieval-augmented generation (RAG) based dynamic prompting methods for LLMs to address the task of causality mining. These approaches leverage both semantic and pattern-based similarity between the input sentence and a collection of test sentences to construct dynamic prompts. Experimental results demonstrate that our proposed methods significantly enhance the performance of causality detection and causality extraction tasks.

9 Limitations

One major limitation of our approach is its focus on intra-sentential causality relations, without addressing inter-sentential causality relations. Extending our method to detect and extract inter-sentential causality could be a promising direction for future research, offering significant benefits to real-world

situations. For example, in the medical domain, understanding complex causal relationships across sentences is crucial. Besides, causality relations exist across all languages. Our experiments have been conducted only in English language. An in-depth understanding of the meaning and structure of different languages would help extend our RAG-based approaches to causality mining in different languages.

10 Acknowledgments

Dr. Hasan’s research is supported by National Science Foundation (NSF) grant number 2417275.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Wajid Ali, Wanli Zuo, Rahman Ali, Xianglin Zuo, and Gohar Rahman. 2021. Causality mining in natural languages using machine and deep learning techniques: A survey. *Applied Sciences*, 11(21):10064.
- Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 1.

- Steven Bethard and James H Martin. 2008. Learning semantic links from a corpus of parallel temporal and causal relations. In *Proceedings of ACL-08: HLT, Short Papers*, pages 177–180.
- Eduardo Blanco, Nuria Castell, and Dan I Moldovan. 2008. Causal relation extraction. In *Lrec*, volume 66, page 74.
- Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint ArXiv:2005.14165*.
- Chih-Yao Chen and Cheng-Te Li. 2021. Zs-bert: Towards zero-shot relation extraction with attribute representation learning. *arXiv preprint arXiv:2104.04697*.
- 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.
- Tharini N De Silva, Xiao Zhibo, Zhao Rui, and Mao Kezhi. 2017. Causal relation identification using convolutional neural networks and knowledge based features. *International Journal of Computer and Systems Engineering*, 11(6):696–701.
- Sefika Efeoglu and Adrian Paschke. 2024. Retrieval-augmented generation-based relation extraction. *arXiv preprint arXiv:2404.13397*.
- Daniela Garcia, EDF-DER, and IMA-TIEM. 1997. Coatis, an nlp system to locate expressions of actions connected by causality links. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 347–352. Springer.
- 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.
- Roxana Girju, Dan I Moldovan, et al. 2002. Text mining for causal relations. In *FLAIRS*, volume 2, pages 360–364.
- M.K. Goddu and A Gopnik. 2024. The development of human causal learning and reasoning. *Nature Review Psychology*, 3:319–339.
- Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892.
- Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. 2014. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 987–997.
- Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2019. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. *arXiv preprint arXiv:1911.10422*.
- Christopher Hidey and Kathleen McKeown. 2016. Identifying causal relations using parallel wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1424–1433.
- Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. 2023. Can large language models infer causation from correlation? *arXiv preprint arXiv:2306.05836*.
- Leo Joskowicz, T Ksiezyck, and Ralph Grishman. 1989. Deep domain models for discourse analysis. In *1989 The Annual AI Systems in Government Conference*, pages 195–196. IEEE Computer Society.
- Md Ahsanul Kabir. 2023. *Deep Learning Based Methods for Automatic Extraction of Syntactic Patterns and their Application for Knowledge Discovery*. Ph.D. thesis.
- Randy M Kaplan and Genevieve Berry-Rogghe. 1991. Knowledge-based acquisition of causal relationships in text. *Knowledge Acquisition*, 3(3):317–337.
- Humayun Kayesh, Md Saiful Islam, and Junhu Wang. 2019. On event causality detection in tweets. *arXiv preprint arXiv:1901.03526*.
- Christopher SG 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(4):177–186.
- John Kontos and Maria Sidiropoulou. 1991. On the acquisition of causal knowledge from scientific texts with attribute grammars. *International Journal of Applied Expert Systems*, 4(1):31–48.
- Yuquan Lan, Dongxu Li, Yunqi Zhang, Hui Zhao, and Gang Zhao. 2023. Modeling zero-shot relation classification as a multiple-choice problem. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

- Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. Unified demonstration retriever for in-context learning. *arXiv preprint arXiv:2305.04320*.
- Zhaoning Li, Qi Li, Xiaotian Zou, and Jiangtao Ren. 2021. Causality extraction based on self-attentive bilstm-crf with transferred embeddings. *Neurocomputing*, 423:207–219.
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.
- Junbao Liu, Xizhong Qin, Xiaoqin Ma, and Wensheng Ran. 2023. Freda: Few-shot relation extraction based on data augmentation. *Applied Sciences*, 13(14):8312.
- Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.
- Partha Pakray and Alexander Gelbukh. 2014. An open-domain cause-effect relation detection from paired nominals. In *Mexican International Conference on Artificial Intelligence*, pages 263–271. Springer.
- Santanu Pal, Partha Pakray, Dipankar Das, and Sivaji Bandyopadhyay. 2010. Ju: a supervised approach to identify semantic relations from paired nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 206–209.
- Edoardo Maria Ponti and Anna Korhonen. 2017. Event-related features in feedforward neural networks contribute to identifying causal relations in discourse. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 25–30.
- Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *Proceedings of the 21st international conference on World Wide Web*, pages 909–918.
- Mehwish Riaz and Roxana Girju. 2010. Another look at causality: Discovering scenario-specific contingency relationships with no supervision. In *2010 IEEE Fourth International Conference on Semantic Computing*, pages 361–368. IEEE.
- Jawad Sadek. 2013. Automatic detection of arabic causal relations. In *Natural Language Processing and Information Systems: 18th International Conference on Applications of Natural Language to Information Systems, NLDB 2013, Salford, UK, June 19-21, 2013. Proceedings 18*, pages 400–403. Springer.
- Craig Silverstein, Sergey Brin, Rajeev Motwani, and Jeff Ullman. 2000. Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery*, 4:163–192.
- Antonio Sorgente, Giuseppe Vettigli, and Francesco Mele. 2013. Automatic extraction of cause-effect relations in natural language text. *DART@ AI* IA*, 2013:37–48.
- Yuanhe Tian, Yan Song, and Fei Xia. 2022. Improving relation extraction through syntax-induced pre-training with dependency masking. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1875–1886.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Liang Wang, Nan Yang, and Furu Wei. 2023. Learning to retrieve in-context examples for large language models. *arXiv preprint arXiv:2307.07164*.
- Shanchuan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2361–2364.
- Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active example selection for in-context learning. *arXiv preprint arXiv:2211.04486*.
- Yuzhe Zhang, Yipeng Zhang, Yidong Gan, Lina Yao, and Chen Wang. 2024. Causal graph discovery with retrieval-augmented generation based large language models. *arXiv preprint arXiv:2402.15301*.
- Yi Zhao, Huaiyu Wan, Jianwei Gao, and Youfang Lin. 2019. Improving relation classification by entity pair graph. In *Asian Conference on Machine Learning*, pages 1156–1171. PMLR.

11 Appendix

The prompts for causal connective extraction, causality detection (zeroshot setting), and causality extraction (zeroshot setting) are shown in Figures 6, 5, and 4, respectively. For different RAG settings and for different causality tasks, the following texts and dynamically generated examples (as retrieved by various RAG methods) are appended to the associated prompt. “Below are 10 example sentences which show causality relation with their cause and effect phrase delimited between the XML tags <cause>, </cause>, <effect> and </effect>.” + EXAMPLES.”

Causality Extraction Prompt

You will be given a sentence that shows a causality relation. Your task is to identify all the cause phrase, effect phrase, and causality phrase from that sentence. Below is the hints and rules for the task.

Hints: In sentences showing a causality relation, you can identify a pair of phrases where one phrase causes the other (effect). There will also be a phrase between the cause and effect phrases that indicates the causality relation; this is the causality phrase. And in most sentences causality phrase can be seen between the cause and effect phrase.

Your output should be in JSON format with keys `cause`, `effect`, and `causality_phrase`: `{"cause1": "", "effect1": "", "causality_phrase1": "", "cause2": "", "effect2": "", "causality_phrase2": "", "cause3": "", "effect3": "", "causality_phrase3": "", ...}`. Make sure your JSON output is closed with `}` at the end.

Rule 1: If you are not able to find the cause phrase, effect phrase, and causality phrase from the given sentence you can output empty string for the keys in the Json format and do not guess the answer.

Rule 2: Ensure that there are no overlapping phrases between the cause phrase, effect phrase, and causality phrase. For instance, consider the input sentence "Fever is caused by virus." Your identified output without overlap should be: `{"cause": "virus", "effect": "fever", "causality_phrase": "caused by"}`. Your output should not contain overlap, as shown below: `{"cause": "by virus", "effect": "fever is caused", "causality_phrase": "caused by"}`. This means that each phrase (cause, effect, and causality phrase) should be distinct and not overlap with each other.

Rule 3: There could be multiple cause-effect phrases in the sentence. You should extract all of them. For instance, consider the input sentence "Mutations in MCFD2 causing F5F8D are null mutations but also missense mutations, which cause destabilization of its flexible structure and the loss of its interaction with ERGIC-53 .". Your identified output will be: `{"cause1": "Mutations in MCFD2", "effect1": "F5F8D", "causality_phrase1": "causing", "cause2": "missense mutations", "effect2": "destabilization of its flexible structure", "causality_phrase2": "cause", "cause3": "missense mutations", "effect3": "the loss of its interaction with ERGIC-53", "causality_phrase3": "cause"}`

Figure 4: Causality Extraction Prompt

A Dataset Description

SemEval: For our causality detection task, we select 249 causal sentences from the test set that are absent in our *Fewshot Example DB* and a random sample of 300 non-causal sentences from the test data (9). We label the causal sentences with a 1 and the non-causal sentences with a 0. Table 8 shows a sample of input from the SemEval data for the causality detection task. For the causality extraction task, we use the same 249 causal sentences from the detection task and their provided cause and effect labels as test keys. Table 7 shows a sample of input sentences along with their provided cause and effect labels.

ADE: This dataset contains 6,821 causal sentences involving drugs and their effects. After filtering sentences having multiple cause-effect phrases, we retain 4,271 instances. We randomly partition these into 2,000 training instances (later used for the *Fewshot Example DB* creation) and 2,271 test instances. For our causality detection task, we use these 2,271 test instances (causal sentences absent in our *Fewshot Example DB*) and combine them with a random sample of 2,000 non-causal sentences from the training dataset of the Li et al. dataset. An ADE input sample is shown in Table 8. For the causality extraction task, we use the 2,271 causal sentences as input and use the drug

Dataset	Input Sentence	label
SemEval	Dogs develop a fever from stress and/or pain such as in a severe flea infestation.	1
	The researchers placed the compound in a tube, which then was put inside a magnet.	0
ADE	Benzocaine-induced methemoglobinemia has been reported in man, dogs, and cats.	1
	The man placed the cartridge into the printer.	0
Li et al.	Paralysis or convulsions are caused by hormone deficiencies and imbalances ..	1
	This theme has been covered in science fiction like Star Trek.	0

Table 8: Task1 Causality detection input examples

and effect labels as ground truth.

Li et al. : For the causality detection task, we use the entire test data (Table 9) as our input (all causal sentences are absent in our *Fewshot Example DB*). For the causality extraction task, we use the 191 causal sentences as input and use the provided cause-effect labels as our test keys. The Li et al. dataset has multiple cause-effect pairs labeled in a given sentence, resulting in a total of 296 unique triplets of (cause, effect, sentence).

Causality Detection Prompt

You will be given a sentence, and your task is to identify whether the sentence shows a causality relation or not. If the sentence shows a causality relation, you can output 1; if it doesn't, output 0. Your output should be in Json format: { output : }.

Hint: If the sentence shows causality relation, you will be able to identify a pair of phrase in the sentence where one phrase will be the cause of the other phrase (effect).

Figure 5: Causality Detection Prompt

Causal Connective Extraction Prompt

You will be provided a sentence which have a causality relation. The cause phrase is delimited between the XML tags <cause> and </cause>. The effect phrase is delimited between the XML tags <effect> and </effect>. Go through the sentence and understand why the phrase between the tags <cause> and </cause> is a cause phrase and the phrase between the tags <effect> and </effect> is the effect phrase in the given sentence. There will be a phrase between cause and effect phrase in the sentence which will help you to understand that the phrase between <cause> and </cause> is a cause phrase and phrase between <effect> and </effect> is the effect phrase in the given sentence. Once you identify that phrase you can output that phrase as a Json format : {output: }.

14 examples sentence and their identified output phrase is given below.

example1: In this article, we describe a Japanese patient with <effect> severe interstitial pneumonia </effect> probably caused by <cause> sorafenib treatment </cause> for metastatic renal cell carcinoma.
output: caused by

example2: A <effect> case of SIADH </effect> induced by <cause> mizoribin administration </cause>.
output: induced by

... ..

example14: <effect> Failures in physical containment </effect> may occur due to <cause> incorrect design </cause>.
output: due to"

Figure 6: Causal Connective Extraction Prompt

Dataset	Split	#total unique instances	#causal relations	#non-causal relations
<i>Fewshot Ex-ample DB</i>	-	2,365	2,365	-
SemEval	Train	8,000	1,003	6,997
	Test	2,717	328	2,389
	Task1 input	549	249	300
	Task2 input	249	249	-
ADE	Train	2,000	2,000	-
	Test	2,271	2,271	-
	Task1 input	4,271	2,271	2,000
	Task2 input	2,271	2,271	-
Li et al.	Train	4,450	1,079	3,371
	Test	786	191	595
	Task1 input	786	191	595
	Task2 input	191	191	-

Table 9: Dataset statistics. (Task1: Causality detection, Task2: Causality extraction)