Causal Retrieval with Semantic Consideration

Hyunseo Shin, Wonseok Hwang*
Department of Artificial Intelligence
University of Seoul
hseo98@uos.ac.kr, wonseok.hwang@uos.ac.kr

Abstract

Recent rapid progress in large language models (LLMs) leads to high performance conversational AI system. To extend the system to expert fields, such as biomedical or legal domains, it becomes standard to combine LLMs with information retrieval (IR) system and generate answer based on retrieved information (documents) for given queries. Thus, it is essential that the IR system should "understand" various intents included in the queries including but not limited to semantic similarity, causal relationship etc. However, existing IR systems primarily focuse on retrieving related information based only on semantic similarities between sentences or documents. Here, we develop CAWAI that can understand causal relation between queries and documents . By training dense retriever with dual constraints, causal loss and semantic loss, CAWAI shows strong generalization capability achieving up to +7.8% Hit@1 compared to DPR baseline in causal retrieval task where a target sentence is buried in 20m Wikipedia sentences.

1 Introduction

With recent advancement large language models (LLMs), it has become standard practice to enhance the performance of LLMs via retrieval-augmented generation (RAG). In RAG system, a document retriever plays a critical role, as it is essential to provide relevant documents for given queries to generate correct answers. Indeed, recent study analyzing the performance of RAG systems in legal domains shows that around 40–50% of hallucinations originate from the failure in document retrieval step [Magesh et al., 2024].

The performance of information retrieval (IR) systems can be roughly defined as providing "relevant" information (documents) for given queries. However, the notion of "relevant" often encompasses various aspects [van Opijnen and Santos, 2017]. For instance, if users want to find similar legal cases, "relevance" may indicate "semantic similarity". On the other hand, if users want to investigate the consequences of specific events, the retriever may need to find documents that include causal consequences of the "cause" mentioned in the queries. For instance, [Ye et al., 2024] categorizes queries based on structure and intent, emphasizing the importance of context for understanding user intent.

However, many existing IR systems primarily focus on semantic similarity between texts to retrieve information. This approach can limit the ability to find relevant information, such as causal relationships. The failure to capture these complex relationships often results in incomplete or misleading information, ultimately affecting decision-making processes for users who rely on accurate data.

Here, we propose CAWAI[2], a new method for training causal dense retriever. By training a dense retriever with dual constraints, causal loss and semantic loss, CAWAI can retrieve either the cause or the effect sentences that has causal relation with the input query. The evaluation result shows that,

*Corresponding author

[2]CAUSALITY AWARE DENSE RETRIEVER

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

CAWAI achieves significantly higher scores compared to existing baselines such as BM25, and DPR upto +7.8 Hit@1, +7.9 Hit@10, and +7.9 MRR@10 demonstrating its effectiveness in modeling and retrieving causal relationships.

In summary our contributions are as follow.

- We propose a CAWAI, a dense retriever specialized in causal retrieval tasks.
- CAWAI achieve significantly better performance compared to existing dense retriever baselines.

We will open source our model to encourage further research in developing causal retriever.

2 Related Work

2.1 Information Retrieval

Traditional keyword-based information retrieval methods like BM25 [Robertson and Zaragoza, 2009] rely heavily on lexical overlap between queries and documents, which limits their ability to capture deeper semantic relationships.

Karpukhin et al. [2020] address this limitation by proposing dense passage retrieval (DPR) that leverages powerful neural language model to convert queries and documents into dense vector representations allowing them to incorporate semantic information beyond simple keyword matching.

Recent studies on dense retrievers shows leveraging LLMs can improve retrieval accuracy. Lee et al. [2024] show the potential of distilling knowledge from LLMs to create compact yet versatile text embedding models. Luo et al. [2024] demonstrate that LLM-based dense retriever significantly outperforms traditional models through comprehensive experiments.

Erker et al. [2024] introduce a Triple-Encoders to compute distributed utterance. The method encodes each sentence independently, and creates contextualized embeddings by linearly combining representations from multiple subspaces. Smilarily, our method also uses three encoders, each specializing in capturing different aspects of causal relationships between sentences.

2.2 Causal Relationship Identification

Recent works in causal discovery with LLMs focus on identifying cause-effect relationships by leveraging causal graphs. Zečević et al. [2023] introduce a framework for causal discovery, where LLMs can return causal graphs through conditional independence statements. Similarly, Zhang et al. [2024] introduce a RAG-based approach for causal graph recovery, dynamically retrieving domain-specific text chunks and inferring relationships between factors using LLMs. While these methods offer insights for post-hoc causal analysis, they apply causal reasoning only after retrieval. In contrast, our work aims to incorporate causal cues at the retrieval stage itself, allowing the model to identify causal relationships from the beginning of the retrieval process.

3 Methods

In this section, we provide the details of CAWAI.

3.1 Model architecture

CAWAI utilizes three encoders: Cause Encoder, Effect Encoder, and Semantic Encoder. The architecture is illustrated in Figure 1.

Cause Encoder processes a text for cause event (e.g. Tom really has no energy to run.), denoted as e_1, generating an vector representation e_1'. Similarly, Effect Encoder processes a text for effect event (e.g. He takes a rest before running again.), e_2, corresponding to e_1, producing an encoded representation e_2'. Semantic Encoder, whose weights are fixed during training, takes both the original cause event e_1 and effect event e_2 as inputs (e_1 and e_2 are encoded separately) and outputs vector representations e_1'' and e_2''.

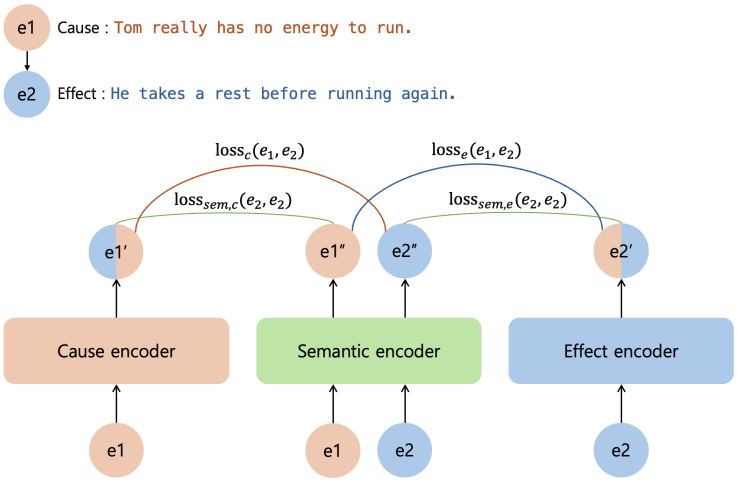


Figure 1: Architecture of CAWAI. CAWAI comprises three encoders: the Cause Encoder, the Semantic Encoder, and the Effect Encoder. Cause Encoder maps the texts corresponding to the causes of some effects into the effect vectors (e_1), Effect Encoder maps the texts corresponding to the effects of some causes into the cause vectors (e_2). Semantic Encoder ensures semantic preservation by minimizing the difference between its own semantic representations and those produced by the Cause and Effect encoders. The red line means loss$_c$, the blue line means loss$_e$ and the green line means loss$_{sem}$.

.

3.2 Training

In line with retrieval-based learning, Cause Encoder is trained to map an input cause event e_1 (text) to its corresponding effect event e_2'' (vector), thereby learning the cause-to-effect relationship. Conversely, Effect Encoder is trained to map the effect event e_2 (text) back to the cause event e_1'' (vector), learning the effect-to-cause relationship.

The weights of Semantic Encoder remain fixed during training, but a semantic preservation loss (loss$_{sem,c}$ in Figure 1), ensures that the encoded representation e_1' from Cause Encoder remains semantically close to the original cause event e_1, and similarly, the encoded representation e_2' from Effect Encoder remains close to the original effect event e_2. This process of semantic alignment facilitates the preservation of contextual nuances and intricate interactions between cause and effect events and helps in maintaining the semantic consistency of both events during training.

In-batch Negative Sampling We use in-batch negative sampling across all three encoders (Cause encoder, Effect encoder, and Semantic encoder)In Cause Encoder, for a given cause event $e_1(i)$) and its corresponding effect event $e_2(i)$), we define a set of negative effects $N(e_2(j)))$ that are sampled from $\{e_2(j)|j \neq i\}$ for all pairs in the batch, where i indicates a batch index and $e_2(i)$ represents the effect events for other cause-effect pairs in the batch. The resulting loss function can be written as

$$\text{loss}_c(e_1(i), e_2(i)) = -\log \frac{\exp(\text{sim}(e_1'(i), e_2''(i)))}{\sum_{j\neq i}^{N_{\text{batch}}} \exp(\text{sim}(e_1'(i), e_2''(j)))} \quad (1)$$

Here, $\text{sim}(e_1'(i), e_2''(i))$ denotes the similarity between the outputs of Cause Encoder ($e_1'(i)$) and the output of Semantic Encoder ($e_2''(i)$), where $e_2(j)''$ represents the encoded output of other effect events within the same batch.

Similarly, in Effect Encoder, for a given effect event e_2, we define negative causes $N(e_1)$ sampled from $\{e_1(i)|i \neq j\}$, forcing the model to map effect-to-cause more accurately by distinguishing from irrelevant cause events.

$$\text{loss}_e(e_1(i), e_2(i)) = -\log \frac{\exp(\text{sim}(e_2'(i), e_1''(i)))}{\sum_{j\neq i}^{N_{batch}} \exp(\text{sim}(e_2'(i), e_1''(j)))} \tag{2}$$

In addition to two losses above, we introduce semantic losses where the output of Cause Encoder (e_1') is contrasted with the output of Semantic Encoder (e_1''). Likewise, Effect Encoder's output e_2' is compared against its original effect sentence (e_2'') from Semantic Encoder ensuring the outputs stay semantically close to their respective inputs.

$$\text{loss}_{\text{sem},c}(e_1(i), e_1(i)) = -\log \frac{\exp(\text{sim}(e_1'(i), e_1''(i)))}{\sum_{j\neq i}^{N} \exp(\text{sim}(e_1'(i), e_1''(j)))} \tag{3}$$

We apply the same negative sampling technique for the effect events in Semantic encoder.

The total loss is then computed as:

$$\text{loss}_{total} = \beta(\text{loss}_{\text{sem},c} + \text{loss}_{\text{sem},e}) + \text{loss}_c + \text{loss}_e \tag{4}$$

This final loss ensures that the cause-to-effect and effect-to-cause mappings are learned effectively, while also preserving the semantic consistency of the original inputs.

The semantic loss terms imposes dual constraints to the vector representation, which helps regularize the model. For instance, e_1' should include the information about the corresponding effect text (e_2'') while preserving their own representation (cause text, e_1''). We also assume preserving semantic similarity may help as a keyword-based retrieval algorithm, such as BM25, can sometimes retrieve answer for given question.

4 Experiments

4.1 Datasets

We utilize the `e-CARE` [Du et al., 2022] and `BCOPA-CE` [Han and Wang, 2021] datasets. The `e-CARE` dataset is split into training, validation, and test sets in a ratio of 6:1:1. `BCOPA-CE` were used only for the training and validation of models The `BCOPA-CE` dataset consists of 500 triplets of `<cause, premise, effect>`, where the premise simultaneously acts as the effect of the cause sentence and the cause of the effect sentence. We transform the dataset into causal retrieval tasks. The resulting dataset comprises 1,000 cause-effect pairs, which are further separated into the training and validation set at ratio 9:1. To prevent data leakage, pairs generated from the same `<cause, premise, effect>` triplet are always included in the same dataset (either training or validation).

The resulting dataset consists of 13,692 training examples, 2,232 validation examples, and 2,136 test examples where each example consists of a pair of cause and effect texts. We use several datasets to evaluate performance of retrieval. We evaluate CAWAI with e-CARE test set while varying the domain and the size of the retrieval pool. The pool comprises the CoLA dataset(Wang et al. [2023]), which includes 1700 sentences, consisting of pairs of events in their temporal order from the RocStories corpus. The CoLA dataset closely resembles the domain in training dataset, although they differ in format. To simulate a real-world scenario, we augment the retrieval pool ranging from 2k sentences (wiki_S), 20k sentences (wiki_M), 200k sentences (wiki_L), 2m sentences (wiki_{XL}), to 20m sentences (wiki_{XXL}) from Wikipedia[3]. We also prepared another retrieval pool augmented similarly using the English corpus from RedPajama-Data-v2 [Computer, 2023][4].

4.2 Model

We used BM25(Robertson and Zaragoza [2009]), and DPR(Karpukhin et al. [2020]) as baseline models. We trained the BERT-base-uncased model as encoders for each model and selected the one with the highest accuracy on the validation dataset after running 500 epochs. The batch size was

[3]https://huggingface.co/datasets/wikimedia/wikipedia

[4]https://huggingface.co/datasets/togethercomputer/RedPajama-Data-V2

set to 64, and the learning to 1e-5 with AdamW optimizer. The experiments were conducted using either on NVIDIA A6000 GPUs or 3090 GPUs.

5 Results

5.1 Retrieval Accuracy

Table 1: **Comparison of various models** In Task 1, a model needs to retrieve the corresponding effect sentence for given cause text as a query whereas in Task 2, the model receive effect sentence and retrieve corresponding cause sentence.

Model	e-CARE			e-CARE + CoLA			e-CARE + wiki$_S$			e-CARE + wiki$_M$			e-CARE + wiki$_L$			e-CARE + wiki$_{XL}$			e-CARE + wiki$_{XXL}$		
	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10
Task 1. Cause to Effect																					
BM25	0.089	0.218	0.127	0.086	0.202	0.120	0.096	0.211	0.132	0.089	0.197	0.119	0.074	0.146	0.096	0.046	0.085	0.058	0.031	0.052	0.036
DPR	**0.379**	0.660	0.468	**0.359**	0.634	0.445	0.362	0.634	0.448	0.321	0.561	0.397	0.257	0.458	0.318	0.151	0.287	0.189	0.088	0.169	0.111
CAWAI	0.373	0.641	0.452	0.354	0.603	0.428	**0.364**	0.629	0.443	**0.346**	0.591	0.417	**0.303**	0.503	0.362	**0.220**	0.337	0.256	0.166	0.248	0.190
Task 2. Effect to Cause																					
BM25	0.094	0.206	0.126	0.088	0.190	0.118	0.096	0.211	0.130	0.091	0.189	0.120	0.079	0.145	0.098	0.049	0.089	0.060	0.028	0.055	0.036
DPR	**0.381**	0.660	0.466	**0.361**	0.630	0.443	0.371	0.639	0.453	0.342	0.573	0.415	0.282	0.474	0.343	0.178	0.320	0.219	0.102	0.202	0.130
CAWAI	0.378	0.641	0.458	0.352	0.608	0.434	**0.374**	0.631	0.451	**0.344**	0.584	0.418	**0.301**	0.495	0.359	**0.219**	0.337	0.254	0.161	0.249	0.188

Model	e-CARE + RedPajama$_S$			e-CARE + RedPajama$_M$			e-CARE + RedPajama$_L$			e-CARE + RedPajama$_{XL}$			e-CARE + RedPajama$_{XXL}$		
	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10
Task 1. Cause to Effect															
BM25	0.102	0.227	0.137	0.089	0.192	0.120	0.070	0.138	0.091	0.049	0.093	0.061	0.036	0.057	0.042
DPR	**0.355**	0.627	0.442	0.304	0.544	0.379	0.219	0.416	0.277	0.129	0.262	0.167	0.071	0.156	0.096
CAWAI	0.354	0.615	0.432	**0.320**	0.544	0.387	**0.256**	0.419	0.306	**0.207**	0.312	0.236	0.154	0.233	0.176
Task 2. Effect to Cause															
BM25	0.101	0.212	0.133	0.094	0.182	0.120	0.074	0.134	0.092	0.046	0.092	0.059	0.032	0.056	0.039
DPR	0.362	0.631	0.444	0.317	0.547	0.380	0.240	0.427	0.297	0.161	0.293	0.198	0.095	0.183	0.119
CAWAI	**0.365**	0.618	0.442	**0.333**	0.551	0.396	**0.275**	0.431	0.321	**0.207**	0.320	0.240	0.161	0.231	0.182

CAWAI achieves comparable performance to DPR We first measure the accuracy in causal retrieval tasks when the size of the pool is 2,136. (Table 1). The models are initialized with the same weights: `bert-base-uncased` [5].

In various tasks, CAWAI shows comparable performance to DPR when the retrieval pool is small. In e-CARE, e-CARE + CoLA, and e-CARE + Wiki$_S$ (with retrieval pool sizes of 2,136, 3,836, and 4,136, respectively), CAWAI consistently achieves similar Hit@1 and MRR@10 performance to DPR. Across these smaller retrieval pools, CAWAI performs on par with DPR.

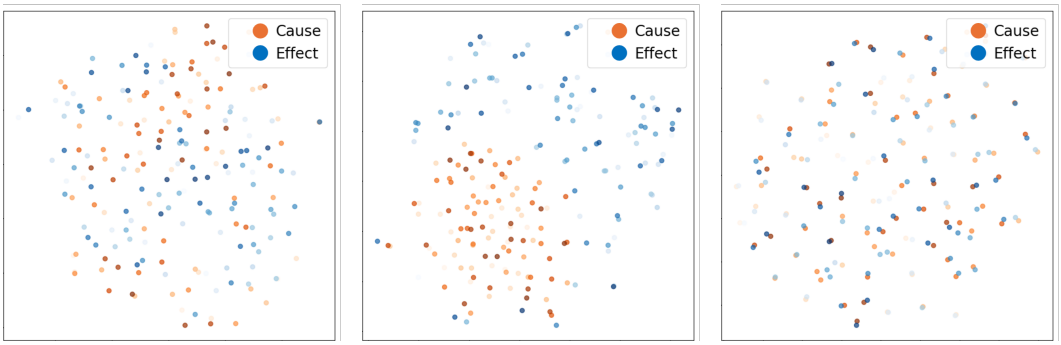
CAWAI shows superior generalization performance compared to DPR Next, we increase the size and diversity of the retrieval pool. As shown in Table 1, CAWAI demonstrates superior performance compared to DPR, particularly in Hit@1 scores, where the differences range from 0.105 to 0.126 as the dataset size increase. The model's capability to identify causal cues across diverse contexts also suggests its robustness in handling complex queries that go beyond simple semantic similarity, highlighting its potential for applications where understanding deeper relationships between events is essential. The experiments with RedPajama corpus also shows similar results (Table 1, second panel).

5.2 Ablation study

Next, we examine whether the results depend on the presence of semantic loss. Table 2 presents how the performance changes on different configurations under varying weights (β) of the semantic loss (Eq. 4) Upon integration of the semantic loss, there is a noticeable improvement in accuracy compared (row 1 vs row 2) indicating that a emphasis on semantic loss correlates positively with model performance. The best results is obtained with we set β = 2. The examples corresponding to each model can be found in Appendix Table 5.

Table 2: **The Accuracy comparison on varying hyperparameter of semantic loss** This table shows the accuracy results based on a difference parameters of semnatic loss. Task 1 involves retrieving the corresponding effect sentence given a query (cause), while Task 2 is the reverse, retrieving the corresponding cause sentence given the effect.

Loss	e-CARE			e-CARE + CoLA			e-CARE + wiki$_S$			e-CARE + wiki$_M$			e-CARE + wiki$_L$			e-CARE + wiki$_{XL}$		
	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10
Task 1. Cause to Effect																		
$\text{loss}_c + \text{loss}_e$	0.320	0.604	0.405	0.300	0.561	0.378	0.312	0.586	0.395	0.279	0.544	0.360	0.228	0.436	0.290	0.125	0.245	0.159
$\text{loss}_c + \text{loss}_e + 0.1\ \text{loss}_s$	0.376	0.632	0.453	0.360	0.593	0.431	0.368	0.622	0.444	0.346	0.576	0.417	0.298	0.489	0.358	0.241	0.330	0.206
$\text{loss}_c + \text{loss}_e + 2\ \text{loss}_s$	0.390	0.646	0.468	0.372	0.609	0.444	0.384	0.635	0.459	0.360	0.592	0.430	0.307	0.505	0.368	0.227	0.341	0.261
$\text{loss}_c + \text{loss}_e + 5\ \text{loss}_s$	0.385	0.644	0.464	0.372	0.609	0.443	0.381	0.635	0.457	0.360	0.593	0.431	0.317	0.507	0.374	0.221	0.345	0.258
Task 2. Effect to Cause																		
$\text{loss}_c + \text{loss}_e$	0.319	0.594	0.404	0.298	0.556	0.377	0.308	0.308	0.392	0.282	0.519	0.356	0.232	0.426	0.292	0.132	0.266	0.171
$\text{loss}_c + \text{loss}_e + 0.1\ \text{loss}_s$	0.373	0.632	0.453	0.353	0.601	0.429	0.366	0.622	0.445	0.349	0.583	0.421	0.310	0.508	0.371	0.220	0.357	0.262
$\text{loss}_c + \text{loss}_e + 2\ \text{loss}_s$	0.382	0.649	0.462	0.365	0.608	0.439	0.380	0.634	0.456	0.362	0.600	0.432	0.324	0.519	0.382	0.232	0.366	0.273
$\text{loss}_c + \text{loss}_e + 5\ \text{loss}_s$	0.392	0.646	0.466	0.371	0.610	0.441	0.385	0.636	0.458	0.360	0.594	0.431	0.315	0.513	0.374	0.237	0.365	0.277



(a) Bert base-uncased (b) DPR (c) CAWAI

Figure 2: **t-SNE visualization of** `BCOPA-CE` **validation set**. In each figure, red represents the cause embeddings, and blue represents the premise embeddings, which are the effect of the cause embeddings. Each gradation indicates that pairs of cause-effect relationships share the same gradation.

6 Analysis

6.1 t-SNE Visualization

To further demonstrate the effectiveness of our approach, we conducted an embedding visualization using the `BCOPA-CE` validation set. The case where the cause sentences are used as queries are shown in Figure 2. The reverse scenario is depicted in Figure 3 in Appendix. When querying with the cause, in CAWAI we input the cause in to Cause Encoder while premise into Effect Encoder. Prior to fine-tuning, the embeddings appear randomly scattered ((a)). As the original structure of DPR is optimized for question-passage retrieval pairs, the space is separated with no additional semantic meaning shared between cause and effect ((b)). This separation suggests that DPR may not fully capture the semantic relationships required for effective causal mapping. Unlike general question-answering tasks, where the context provides a direct passage for retrieval, causal tasks require a wide understanding of temporal and logical dependencies between events. In CAWAI, we observe that each cause and its corresponding effect are mapped closely in a similar space ((c)), indicating that the model has learned to associate causes and effects more effectively after fine-tuning. This demonstrates that CAWAI effectively learns an embedding space that captures causal relationships.

7 Conclusion

We proposed a novel retrieval, CAWAI that incorporates cause-effect relationships into the retrieval process. Our experiments demonstrate that this approach outperforms traditional DPR models in causal tasks under real-world setting, confirming the effectiveness of modeling causal dependencies for retrieval.

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

8 Acknowledgement

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NO.2022M3J6A1084845).

References

T. Computer. Redpajama: an open dataset for training large language models, 2023. URL https://github.com/togethercomputer/RedPajama-Data.

L. Du, X. Ding, K. Xiong, T. Liu, and B. Qin. e-CARE: a new dataset for exploring explainable causal reasoning. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.33. URL https://aclanthology.org/2022.acl-long.33.

J.-J. Erker, F. Mai, N. Reimers, G. Spanakis, and I. Gurevych. Triple-encoders: Representations that fire together, wire together. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5317–5332, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.290. URL https://aclanthology.org/2024.acl-long.290.

M. Han and Y. Wang. Doing good or doing right? exploring the weakness of commonsense causal reasoning models. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 151–157, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.20. URL https://aclanthology.org/2021.acl-short.20.

V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. Dense passage retrieval for open-domain question answering. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550.

J. Lee, Z. Dai, X. Ren, B. Chen, D. Cer, J. R. Cole, K. Hui, M. Boratko, R. Kapadia, W. Ding, Y. Luan, S. M. K. Duddu, G. H. Abrego, W. Shi, N. Gupta, A. Kusupati, P. Jain, S. R. Jonnalagadda, M.-W. Chang, and I. Naim. Gecko: Versatile text embeddings distilled from large language models, 2024. URL https://arxiv.org/abs/2403.20327.

K. Luo, M. Qin, Z. Liu, S. Xiao, J. Zhao, and K. Liu. Large language models as foundations for next-gen dense retrieval: A comprehensive empirical assessment, 2024. URL https://arxiv.org/abs/2408.12194.

V. Magesh, F. Surani, M. Dahl, M. Suzgun, C. D. Manning, and D. E. Ho. Hallucination-free? assessing the reliability of leading ai legal research tools, 2024. URL https://arxiv.org/abs/2405.20362.

S. E. Robertson and H. Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389, 2009. URL https://api.semanticscholar.org/CorpusID:207178704.

M. van Opijnen and C. Santos. On the concept of relevance in legal information retrieval. *Artificial Intelligence and Law*, 25(1):65–87, Mar 2017. ISSN 1572-8382. doi: 10.1007/s10506-017-9195-8. URL https://doi.org/10.1007/s10506-017-9195-8.

Z. Wang, Q. V. Do, H. Zhang, J. Zhang, W. Wang, T. Fang, Y. Song, G. Y. Wong, and S. See. Cola: contextualized commonsense causal reasoning from the causal inference perspective. *arXiv preprint arXiv:2305.05191*, 2023.

L. Ye, Z. Lei, J. Yin, Q. Chen, J. Zhou, and L. He. Boosting conversational question answering with fine-grained retrieval-augmentation and self-check, 2024. URL https://arxiv.org/abs/2403.18243.

M. Zečević, M. Willig, D. S. Dhami, and K. Kersting. Causal parrots: Large language models may talk causality but are not causal, 2023. URL https://arxiv.org/abs/2308.13067.

Y. Zhang, Y. Zhang, Y. Gan, L. Yao, and C. Wang. Causal graph discovery with retrieval-augmented generation based large language models. *arXiv preprint arXiv:2402.15301*, 2024.

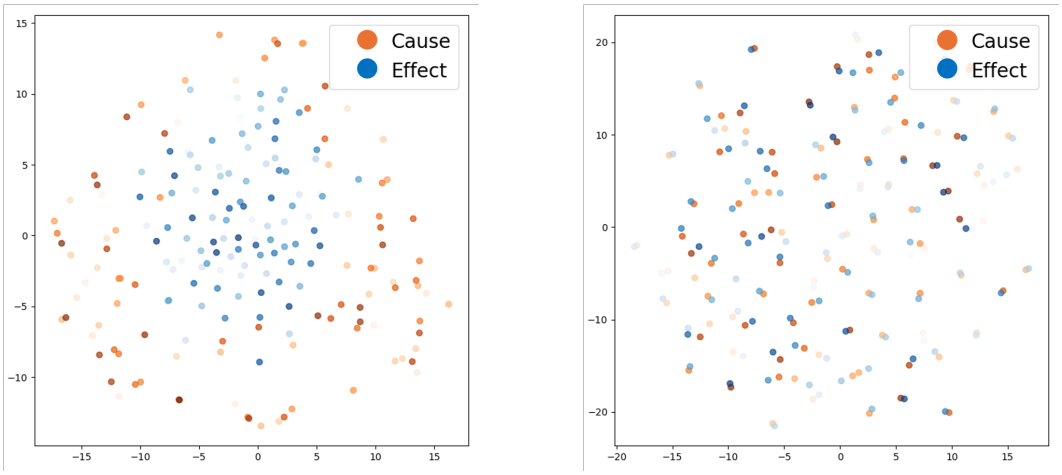
Appendix

Table 3: **The Accuracy comparison on Semantic encoder backbone models** This table shows the Hit@accuracy results based on a Semantic encoder backbone. Task 1 involves retrieving the corresponding effect sentence given a query (cause), while Task 2 is the reverse, retrieving the corresponding cause sentence given the effect.

Backbone	e-CARE			e-CARE + CoLA			e-CARE + wiki$_S$			e-CARE + wiki$_M$			e-CARE + wiki$_L$			e-CARE + wiki$_{XL}$		
	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10
Task 1. Cause to Effect																		
Sentence Bert	0.313	0.585	0.393	0.296	0.541	0.368	0.299	0.556	0.375	0.268	0.478	0.330	0.217	0.368	0.261	0.150	0.235	0.176
Task 2. Effect to Cause																		
Sentence Bert	0.306	0.581	0.388	0.282	0.546	0.360	0.294	0.553	0.371	0.267	0.486	0.332	0.218	0.385	0.267	0.156	0.247	0.181

Table 4: **The Accuracy using Semantic encoder as passage encoder.** Task 1 involves retrieving the corresponding effect sentence given a query (cause), while Task 2 is the reverse, retrieving the corresponding cause sentence given the effect.

Passage encoder	e-CARE			e-CARE + CoLA			e-CARE + wiki$_S$			e-CARE + wiki$_M$			e-CARE + wiki$_L$		
	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10
Task 1. Cause to Effect															
Semantic encoder	0.283	0.550	0.362	0.261	0.512	0.335	0.233	0.457	0.300	0.177	0.349	0.226	0.124	0.234	0.157
Task 2. Effect to Cause															
Semantic encoder	0.293	0.570	0.372	0.272	0.524	0.346	0.238	0.484	0.310	0.180	0.363	0.233	0.125	0.241	0.157



(a) DPR (b) CAWAI

Figure 3: **t-SNE visualization of `BCOPA-CE` validation set**. In each figure, red represents the premise embeddings, and blue represents the effect embeddings, which are the effect of the premise embeddings.

Table 5: **Example of Model Responses comparison on dataset e-CARE+wiki$_S$**

Input Cause	"Tom had a kidney transplant."
DPR	Top 1: "He received a liver transplant in 2012." (Wrong) Top 2: "Doctors suggested transplanting his kidney as soon as possible but at first there was a lack of potential compatible donor." (Wrong) Top 3: "His parents elected to donate his organs for transplant, a decision which was credited with saving five lives." (Wrong)
CAWAI	Top 1: "The doctor gave him a lot of medicine to fight allergy rejection." (Correct) Top 2: "The transplant would not save his life, but it might give him better breathing." (Wrong) Top 3: "The transplant revitalized his immune system." (Wrong)
Input Cause	"John suffered from indigestion."
CAWAI with no semantic loss	Top 1: "He had a serious stomachache after eating it." (Wrong) Top 2: "His stomach mucos membrane was damaged." (Wrong) Top 3: "He choked because of the blockage." (Wrong)
CAWAI	Top 1: "He needed to start water fasting to rest his digestive tract." (Correct) Top 2: "The doctor advised him to drink green tea." (Wrong) Top 3: "He had indigestion." (Wrong)