# Grounding Large Language Model with Causal Knowledge Retrieval

Anonymous ACL submission


## Abstract

Large Language Models (LLMs) often require grounding on external knowledge to generate accurate and faithful outputs. However, this process can easily fail with inaccurate semantic similarity searches: it tends to retrieve information that only appears similar to the query without actually aiding in the response, thus acting as noise or even misguiding the generation. Addressing this issue, we propose the Causal Inference Score (CIS), which measures how likely a knowledge candidate will help answer the user's question by computing the debiased textual entailment confidence between the question and the candidate using an LLM. For cost-efficient inference, we further propose a knowledge distillation method to transfer CIS estimation to a lightweight BERT model. Extensive experiments show that simply altering the similarity measure to CIS can lead to significant improvements, increasing answer accuracy by up to 20.5% and F1 by 23.3%, outperforming recent works that involve complex multistage pipelines.


## 1 Introduction

In recent years, Large Language Models (LLMs) have demonstrated impressive capabilities in various natural language processing tasks (Brown et al., 2020; et al., 2024; Vaswani et al., 2017; Devlin et al., 2019). However, LLMs are reported to suffer from "hallucination" that produces plausible but factually incorrect information in the responses (Bender et al., 2021; Ji et al., 2023; Zellers et al., 2019). To mitigate this issue, retrieval-augmented generation (RAG) is proposed to integrate external knowledge from trusted knowledge sources before model generation, trying to override the outdated or wrong knowledge stored in the model parameters with the explicit contextual knowledge (Lewis et al., 2020b,a; Gao et al., 2024; Shapkin et al., 2024). Typically, an RAG system

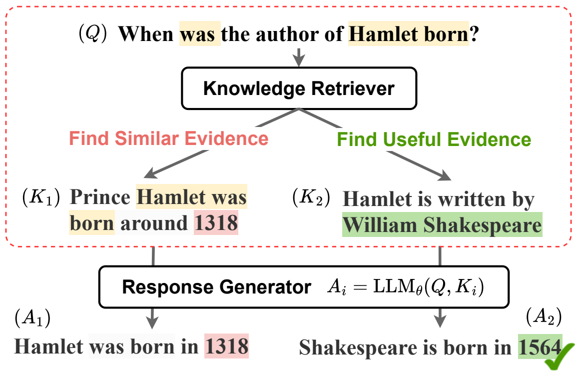


Figure 1: **Similarity vs. Causal Relevance.** High similarity, i.e., semantic overlap, between a question and related knowledge does not guarantee the utility of that knowledge for generating correct responses. A more effective metric would measure the degree to which the knowledge causally answers the question (or part of it).

uses a retriever model to first find knowledge evidences based on the input query, and then utilizes a generator model, usually an LLM, to generate the response (Cai et al., 2022; Ramesh et al., 2023; Zhang et al., 2024b).

Despite its effectiveness, RAG can easily fail and generate unfaithful responses when facing inaccurate retrieval results. With current similarity search methods, retrievers often find information that is semantically similar but not substantively useful for answering the user's question (Guu et al., 2020; Salemi and Zamani, 2024). This can easily mislead the LLM, causing it to generate inaccurate responses that stray from the original question. As shown in Fig. 1, consider the question: "When was the author of Hamlet born?" If we use similarity measures, we might find the text: "Prince Hamlet was born around 1318", which is very similar to the question but can mislead the LLM to produce an incorrect answer. This highlights the urgent need for a more effective retrieval method that focuses on accurately measuring the causal relevance between the question and text snippets, aiming to discover truly useful knowledge rather than merely similar information (Feder et al., 2022; Li et al., 2023).

To address this issue, we explore causal knowledge retrieval, which seeks external knowledge that directly addresses the question rather than merely resembling it. To this end, we propose a novel metric, i.e., *Causal Inference Score* (CIS), to measure the causal relevance using pre-trained language models, as they are extensively trained to learn causal entailments between texts. Given the question $Q$ and candidate text knowledge $K$, CIS is calculated by first determining the entailment confidence from $Q$ to $K$ through the likelihood of the model generating $K$ given $Q$, i.e., $p_\theta(K|Q)$. This metric can be biased when the LLM is overly familiar with the knowledge $K$, so we further mitigate this self-confirmation bias by scaling the entailment with the model probability of the knowledge, i.e., $\text{CIS}_\theta(Q \to K) = p_\theta(K|Q)/p_\theta(K)$. By emphasizing the causal relationship, we ensure that the retrieved information is what the LLM believes can effectively address the query and allows the model to leverage this knowledge more effectively for answer generation. In RAG applications, CIS replaces traditional similarity scores in the retriever while keeping all other components unchanged. Addressing the high inference cost of LLMs, we further propose a knowledge distillation method to effectively transfer the causal entailment capabilities learned by large auto-regressive language models into compact, inference-efficient bidirectional models, thereby enabling efficient and accurate document retrieval.

To evaluate the effectiveness of CIS, we conducted experiments on three question-answering datasets (HotpotQA, 2Wiki, and MuSiQue) and two information retrieval datasets (TREC-DL2019 and TREC-DL2020). Results show that replacing traditional similarity metrics with CIS significantly improves performance, increasing QA accuracy by $7.8\%$ to $10.75\%$ and NDCG@K for retrieval by at least $9.81\%$ relative to BM25. Moreover, even using weaker LLMs (e.g., GPT-2) for metric calculation yields notable gains, highlighting its broader applicability.

## 2 Related Work

**RAG for Multi-Hop QA.** RAG is a widely used framework for LLMs and has garnered considerable attention for various tasks such as question-answering (QA) and summarization. RAG (Lewis et al., 2020b) integrates a sequence-to-sequence model with external knowledge bases, significantly enhancing the performance of QA and summarization tasks. Breaking down a complex query into a series of simpler sub-queries (Khattab et al., 2022; Press et al., 2022; Pereira et al., 2022; Khot et al., 2022; Sun et al., 2023b) often necessitates multiple calls to LLMs, which can be computationally expensive. Adaptive-RAG (Jeong et al., 2024) addresses this issue by using a classifier to evaluate the problem's complexity and select the most suitable retrieval strategy accordingly. RQ-RAG (Chan et al., 2024) focuses on enhancing model performance by optimizing search queries through techniques like rewriting, decomposition, and disambiguation. Nevertheless, relying on multiple accesses to LLMs for each query is inefficient, and retrieving all dynamically relevant documents with a single query is unreliable.

**Retriever in RAG.** Traditional retrieval methods in RAG systems rely on similarity measures to find relevant documents, but struggle with queries involving logical or causal relationships, as they focus on shared words or phrases rather than deeper connections. To address this issue, we propose an enhanced causal retrieval approach that captures implicit connections and causal relationships by measuring term co-occurrence probabilities relative to their independent occurrences, enabling a more nuanced retrieval process.

In our approach, a causal reasoning score is calculated between the query and each document, and the documents with the highest causal reasoning score are considered highly relevant, indicating a stronger causal relationship with the query. These documents are then used by LLM to generate precise answers. This approach improves the quality of retrieved documents by ensuring that the documents are not only semantically relevant but also causally relevant, thereby improving the accuracy and relevance of the final answers generated (Jain et al., 2023; Zhang et al., 2024a).

## 3 Background

In an advanced RAG system, the process begins with a user-input query $Q$, which is processed by a retrieval module $\psi(\cdot)$ to extract relevant information text $\mathcal{B}$ from a comprehensive information repository $\mathcal{D}$. These retrieved texts are then used by a generation module $\gamma(\cdot)$ to produce the final output $R$. This workflow can be expressed as $\mathcal{B} =$

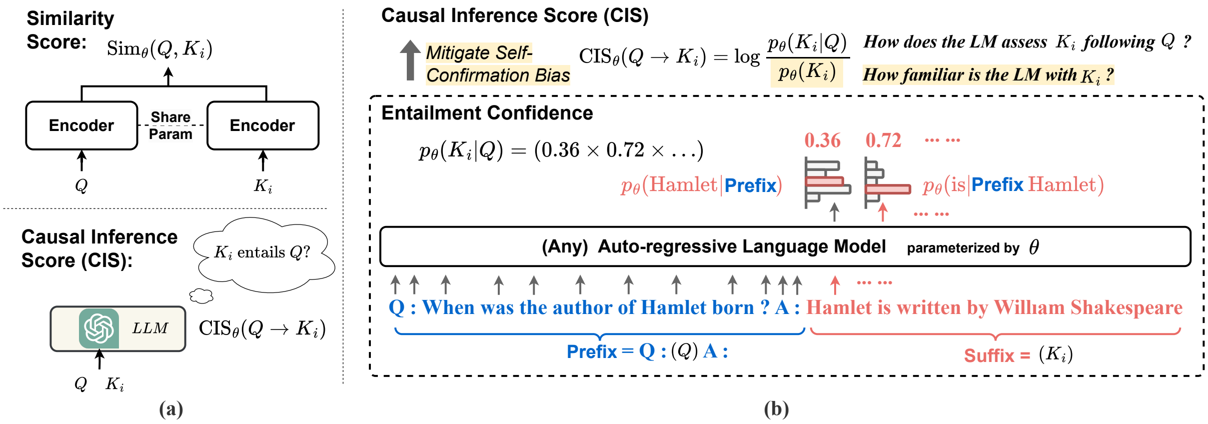


Figure 2: Diagram of CIS. **(a) CIS vs. Similarity Score.** Traditional similarity scores treat the question and knowledge as equal entities, focusing on shared semantic overlap. However, CIS considers the directional and causal relationship between the question and knowledge, assessing how well the question leads to or infers the knowledge. **(b) CIS Calculation.** First, we compute the entailment confidence by calculating the likelihood of the model generating $K_i$ given $Q$ as the prefix, denoted as $p_\theta(K_i \mid Q)$. This can be further improved by placing the question and the knowledge into a QA prompt to leverage the model's QA capabilities. To avoid self-confirmation bias where the model might be overly familiar with $K_i$, we scale this confidence by the probability of the model directly generating $K_i$, i.e., $p_\theta(K_i)$.

$\psi(Q, \mathcal{D})$, and $R = \gamma(Q, \mathcal{B})$. The retrieval module $\psi(\cdot)$ may involve various systems, such as an independent retrieval system like DPR (Karpukhin et al., 2020) and a commercial search engine like Google. While the generation module $\gamma(\cdot)$ is typically a sophisticated language model that has been pre-trained. The quality of the generated result $R$ is directly influenced by the accuracy of the retrieved information segments $\mathcal{B}$, making precise retrieval a critical component. Unfortunately, many retrieval modules struggle to pinpoint exact segments and often retrieve semantically similar ones, which may not always ensure the accuracy of the final output.

To identify the most relevant information segments $\mathcal{B} = \{K_1, K_2, K_3, \ldots\}$ from the repository $\mathcal{D}$, an effective retrieval strategy is crucial. In this paper, we propose a strategy that selects the information segment collection $\mathcal{B}$ by calculating causal inference scores. Specifically, we use an autoregressive model to compute these scores, enabling us to filter and select the highest-scoring segments $K$ as supporting information. This approach improves inference efficiency and ensures that the generated content accurately reflects the source information. Detailed descriptions of this process will be provided in the following sections.

# 4 Methodology

As depicted in Figure 2, CIS is designed to better capture the causal relationships between a query $Q$ and the candidate documents $K_i$. Unlike traditional similarity-based methods, our approach leverages the power of autoregressive language models to assess how well a document's content entails the query.

## 4.1 Causal Inference Score (CIS)

To improve the retrieval accuracy and mitigate the self-confirmation bias present in traditional methods,we leverage the CIS. The CIS aims to capture the causal relationship between the query $Q$ and the document $K_i$. This is achieved by leveraging an autoregressive language model to assess how well the document content entails the query. The CIS is defined as follows:

$$\mathrm{CIS}_\theta(Q \to K_i) = \log \frac{p_\theta(K_i|Q)}{p_\theta(K_i)}$$

where $\theta$ represents the parameters of the language model. The term $p_\theta(K_i|Q)$ measures how the language model assesses $K_i$ following the query $Q$, and $p_\theta(K_i)$ represents how familiar the language model is with $K_i$. This approach allows us to quantify the causal influence of the document on the query, leading to more accurate retrieval results. The theoretical basis of this method is explained in Appendix B. A positive CIS value indicates a strong correlation between the query and the document, implying a potential causal relationship. A CIS value of zero indicates that the query and document are independent. A negative CIS value indicates that the query and document are unlikely to appear at the same time. The CIS score is then calculated to replace the similarity scores in the retriever. We keep top-$k$ documents with the highest CIS score as the grounding knowledge for LLM generation.

**Entailment Confidence.** To estimate the degree to which the LLM believes the current knowledge candidate $K_i$ should entail the question $Q$, we concatenate the document $K_i = \{w_1, w_2, \ldots, w_n\}$ with the query $Q$ and use the pre-trained language model (PLM) to compute the conditional probability $p_\theta(K_i|Q)$. The PLM uses the query $Q$ as the prefix to sequentially predict the probability of each word in the document. The conditional probability is computed as:

$$p_\theta(K_i|Q) = p_\theta(w_1|Q)p_\theta(w_2|Q, w_1)$$
$$p_\theta(w_3|Q, w_1, w_2) \ldots p_\theta(w_n|Q, w_1, \ldots, w_{n-1})$$

In practice, we further consider a variant where we put the question and knowledge into a QA prompt, forming "Q: $Q$ A: $K$". This approach aims to more explicitly measure how the knowledge can partly address the question. Experimental results show that this can bring (limited) improvements.

**Correction for Self-confirmation Bias.** The language model can be overly familiar with specific text fragments in the knowledge candidates as they commonly appear, leading to an excessively high entailment confidence for those fragments. This issue, which we call self-confirmation bias, is harmful. We correct this bias by scaling the entailment confidence with document likelihood $p_\theta(K_i)$.

Given a document $K_i = \{w_1, w_2, \ldots, w_n\}$ , this likelihood is calculated by predicting the probability of each word in the sequence given its preceding context. This involves calculating the probability of each word $w_i$ in the document given the sequence of all previous words $w_1, w_2, \ldots, w_{i-1}$. Thus, the overall document probability is computed as the product of these conditional probabilities:

$$p_\theta(K_i) = p_\theta(w_1)p_\theta(w_2|w_1)$$
$$p_\theta(w_3|w_1, w_2) \ldots p_\theta(w_n|w_1, w_2, \ldots, w_{n-1})$$

### 4.2 Knowledge Distillation for Efficient Inference

While our proposed causal inference score $\text{CIS}_\theta(Q \rightarrow K_i) = \log \frac{p_\theta(K_i|Q)}{p_\theta(K_i)}$ is a plug-and-play method, it suffers from high computational cost during inference. Although the term $p_\theta(K_i)$ can be computed offline, evaluating $p_\theta(K_i|Q)$ requires multiple forward passes through LLMs.

To address this challenge, we introduce an innovative methodology that distills the causal entailment capabilities from a computationally expensive causal large language model into inference-efficient bidirectional lightweight language models, such as BERT (Devlin et al., 2018). This approach is both computationally efficient and straightforward to implement, making it suitable for large-scale information retrieval tasks.

For retrieval-related datasets, we generate training data by leveraging an internal unidirectional large language model, which acts as a data generator for the lightweight model. The training process involves generating supervised fine-tuning samples for each instance in the form of the triplet $< Q_i, K_j, \text{CIS}_\theta(Q_i \rightarrow K_j) >$.

During training, we fine-tune BERT using a pointwise learning-to-rank approach that predicts the relevance score for each query-document pair. In this framework, the relevance estimation for each query-document pair is treated as an independent task. The query and document are concatenated into a single input sequence (separated by the special token [SEP] and passed through BERT. The output embedding of the [CLS] token is then used to compute the relevance score via a feed-forward layer. The training objective is to minimize the difference between the predicted score and the ground-truth CIS score, using a loss function $\ell$ defined as:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \ell(S_{i,j}, \text{CIS}_\theta(Q_i \rightarrow K_j))$$

where $\ell$ is Mean Squared Error (MSE) for the regression task, $S_{i,j}$ represents the predicted relevance score between the query $Q_i$ and document $K_j$ by BERT.

## 5 Experiments

### 5.1 Evaluation on Retrieval Augmented Generation

#### 5.1.1 Settings

**Datasets and Metrics.** We assess the effectiveness of our proposed framework using three open-source multi-hop QA datasets:

- **HotpotQA** (Yang et al., 2018) requires models to combine information from multiple

| Methods | MuSiQue | | | | | HotpotQA | | | | | 2Wiki | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | EM | F1 | Acc | Step | Time | EM | F1 | Acc | Step | Time | EM | F1 | Acc | Step | Time |
| **Single-step Approach** | 13.80 | 22.80 | 15.20 | 1.00 | 1.00 | 34.40 | 46.15 | 36.40 | 1.00 | 1.00 | 41.60 | 47.90 | 42.80 | 1.00 | 1.00 |
| **Adaptive Retrieval** | 6.40 | <u>15.80</u> | 8.00 | **0.50** | <u>0.55</u> | 23.60 | 32.22 | 25.00 | **0.50** | <u>0.55</u> | 33.20 | 39.44 | 34.20 | **0.50** | <u>0.55</u> |
| **Self-RAG** | 1.60 | 8.10 | 12.00 | <u>0.73</u> | **0.51** | 6.80 | 17.53 | 29.60 | <u>0.73</u> | **0.45** | 4.60 | 19.59 | 38.80 | <u>0.93</u> | **0.49** |
| **Adaptive-RAG** | <u>23.60</u> | 31.80 | **26.00** | 3.22 | 6.61 | 42.00 | 53.82 | 44.00 | 3.55 | 5.99 | 40.60 | 49.75 | 46.40 | 2.63 | 4.68 |
| **Multi-step Approach** | 23.00 | <u>31.90</u> | 25.80 | 3.60 | 7.58 | <u>44.60</u> | <u>56.54</u> | **47.00** | 5.53 | 9.38 | <u>49.60</u> | <u>58.85</u> | **55.40** | 4.17 | 7.37 |
| **Causal Retrieval (Ours)** | **27.09** | **38.27** | <u>25.95</u> | 2.00 | 2.09 | **50.42** | **58.24** | <u>44.20</u> | 2.00 | 2.58 | **53.16** | **59.09** | <u>51.61</u> | 2.00 | 1.51 |

Table 1: Results of question answering using Llama3 (8B) as LLM on different datasets. We emphasize the best result in bold and underline the second best score.

paragraphs to answer complex questions, emphasizing reasoning and synthesis. The test set contains 7,405 samples.

- **2Wiki** (Ho et al., 2020) leverages Wikipedia articles to test multi-hop QA, requiring models to link and integrate knowledge from multiple articles, often spanning diverse topics. The test set contains 12,576 samples.

- **MuSiQue** (Trivedi et al., 2022) evaluates the ability to handle complex queries by integrating information from multiple documents, challenging multi-document reasoning. The validation set contains 2,417 samples.

We evaluate question answering performance using F1, Exact Match (EM), and Accuracy (Acc). F1 measures word overlap between the predicted and ground truth answers, EM checks for exact matches, and Acc assesses whether the predicted answer contains the ground truth. For knowledge retrieval, we report recall and precision.

**Baselines.** We compare our approach with state-of-the-art methods as follows: **1) Single-Step Approach-Based Methods:** Adaptive Retrieval (Mallen et al., 2023), Self-RAG (Asai et al., 2024), and Adaptive-RAG, which adaptively performs retrieval based on query complexity (Jeong et al., 2024). **2) Multi-Step Approach:** The most advanced state-of-the-art method (Trivedi et al., 2023), which uses iterative access to both the retriever and LLM with Chain-of-Thought reasoning (Wei et al., 2022) for every query.

Note that we simply replace the similarity score in single-step methods with CIS, without adding the complexities of multistep processes. While integrating these additional steps might improve results, we have not included them in our current evaluation to ensure a straightforward comparison.

#### 5.1.2 Overall Results

Table 1 shows the performance of different methods on the question-answering task using retrieval-augmented generation. The results demonstrate the effectiveness of our proposed CIS. Compared to single-step methods, we found that simply replacing the similarity metric with CIS can improve F1 by $11.19\%$ to $15.47\%$ and EM by $11.56\%$ to $16.02\%$. Our method also outperforms multi-step methods on these metrics.

Furthermore, we observe additional improvements when fine-tuning BERT on data distilled from LLaMA 3-8B, particularly in terms of F1 and accuracy across datasets. In addition to method optimization, we also investigate the role of carefully designed QA prompts in improving LLM answer generation. Well-crafted prompts help ensure that the retrieved content is effectively utilized, leading to more accurate answers. A detailed discussion on this can be found in Appendix E. Further insights into the advantages of our approach are provided in Appendix F, where the QA case study demonstrates enhanced answer generation. Moreover, Appendix C presents the error analysis of CIS, highlighting its limitations and potential future directions.

#### 5.1.3 CIS with Different LLMs

We also explore the impact of using different LLMs to compute CIS. Results in Table 2 indicate that even a weaker model like GPT-2 (Radford et al., 2019) can still lead to significant improvements across datasets, highlighting the robustness of our approach. Similarly, fine-tuning BERT on training data distilled from LLaMA 3-8B achieves competitive results and often surpasses causal models such as GPT-2 and LLaMA 3-8B, demonstrating the effectiveness of distillation-based fine-tuning in this setting.

Interestingly, we observe that the benefits become more pronounced with higher top-k settings, as distillation enables BERT to capture finer-

| Top-k | LLMs | Methods | MuSiQue | | | HotpotQA | | | 2Wiki | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | EM | F1 | Acc | EM | F1 | Acc | EM | F1 | Acc |
| - | - | **Adaptive-RAG** | 23.60 | 31.80 | 26.00 | 42.00 | 53.82 | 44.00 | 40.60 | 49.75 | 46.40 |
| 3 | **gpt-2** | **Causal Retrieval** | 22.46 | 32.05 | 20.02 | **48.07** | **56.14** | **41.09** | 42.99 | 46.24 | 42.01 |
| | **Llama-3** | **Causal Retrieval** | 25.07 | **35.29** | **26.29** | 45.98 | 53.88 | 39.05 | 54.68 | 61.34 | 53.26 |
| | **BERT** | **Knowledge Distill** | **30.81** | 34.29 | 24.57 | 46.12 | 51.54 | 38.40 | **56.80** | **62.74** | **56.20** |
| 4 | **gpt-2** | **Causal Retrieval** | 24.86 | 34.33 | 22.80 | 46.29 | 54.40 | 39.27 | 44.81 | 48.79 | 43.76 |
| | **Llama-3** | **Causal Retrieval** | 27.80 | **38.11** | 28.44 | 47.42 | 55.45 | 40.41 | 54.12 | 60.82 | 52.78 |
| | **BERT** | **Knowledge Distill** | **33.46** | 36.48 | 26.35 | **51.25** | **56.00** | **42.79** | **56.60** | **62.45** | **56.20** |
| 6 | **gpt-2** | **Causal Retrieval** | 27.26 | 37.89 | 24.91 | 40.56 | 48.53 | 34.88 | 46.47 | 51.33 | 45.09 |
| | **Llama-3** | **Causal Retrieval** | 30.03 | **40.83** | **29.60** | **53.08** | **61.48** | **45.59** | 51.71 | 58.40 | 50.22 |
| | **BERT** | **Knowledge Distill** | **34.29** | 39.70 | 27.83 | 49.70 | 55.44 | 40.99 | **53.20** | **58.83** | **52.60** |

Table 2: Results of question answering with different LLMs (Llama 3-8B, gpt-2 1.5B, BERT-340M) and different top-k compared to Adaptive-RAG. Knowledge Distillation refers to the process in which BERT-340M is distilled using Llama 3-8B. We highlight in bold the best results of different LLMs in the same top-k.

grained relevance signals. This suggests that with proper guidance from larger LLMs, lightweight models can effectively balance efficiency and performance, making them viable alternatives for retrieval-augmented tasks.

#### 5.1.4 Impact of Top-k Values

We also experimented with different top-k values. Intuitively, providing more relevant text to the LLM should increase the likelihood of obtaining the correct answer. Our experimental results, shown in Table 2, largely confirm this expectation, as EM, F1 score, and accuracy generally improve with higher top-k values. However, we observed an exception in the 2Wiki dataset, where increasing top-k led to a decline in these metrics. We believe this occurs because, beyond a certain threshold, the retrieved content might exceed the model's input length limit. As a result, the model may truncate or underprocess the input, negatively impacting answer accuracy.

Further analysis of retrieval performance across different top-k values and retrieval strategies is provided in Appendix D. These findings highlight the importance of carefully selecting the retrieval strategy and top-k value to achieve optimal performance.

#### 5.1.5 Cost Analysis

In Table 1, the time consumption includes two main parts: first, calculating all CIS between each question and dozens of relevant or irrelevant texts provided in the dataset; second, combining the text with the highest CIS with the question into a prompt, and inputting the prompt into the large model to generate the answer. As the number of text paragraphs increases, each question needs to be compared with all these texts and the CIS of all texts are calculated, so the amount of calculation increases significantly, resulting in a significant increase in processing time. Compared with the single-step method, even if our method is more time-consuming when the number of texts is small, the time consumption will exceed the multi-step method as the number of texts increases.

To address this issue, preliminary optimization experiments in Appendix A compare retrieval time under different numbers of tokens after fine-tuning BERT. Further optimization details are provided in Section 4.2 Knowledge Distillation for Efficient Inference under Methodology.

### 5.2 Evaluation on Information Retrieval

#### 5.2.1 Settings

**Datasets and Metrics.** We conduct evaluations on the TREC Deep Learning 2019 (DL19) and 2020 (DL20) passage ranking test collections (Craswell et al., 2019, 2020), which serve as prominent benchmarks in information retrieval research. These collections include 43 queries in DL19 and 54 queries in DL20, accompanied by dense, graded human relevance judgments. Both datasets are derived from the MS MARCO v1 (Bajaj et al., 2018) passage corpus, comprising 8.8 million passages. For each query, the top 100 passages retrieved using BM25 (Lin et al., 2021) are re-ranked, ensuring consistency with experimental setups adopted in prior studies (Sun et al., 2023a; Ma et al., 2023; Qin et al., 2024).

**Baselines.** We evaluate our method against a range of baselines. **1) Supervised Methods:** monoBERT (Nogueira and Cho, 2020), a cross-

| Method | LLM | Size | TREC-DL2019 NDCG@1 | NDCG@5 | NDCG@10 | TREC-DL2020 NDCG@1 | NDCG@5 | NDCG@10 |
|---|---|---|---|---|---|---|---|---|
| BM25 | NA | NA | 54.26 | 52.78 | 50.58 | 57.72 | 50.67 | 47.96 |
| **Supervised Methods** | | | | | | | | |
| monoBERT | BERT | 340M | 79.07 | 73.25 | 70.50 | 78.70 | 70.74 | 67.28 |
| monoT5 | T5 | 220M | 79.84 | 73.77 | 71.48 | 77.47 | 69.40 | 66.99 |
| monoT5 | T5 | 3B | 79.07 | 73.74 | 71.83 | 80.25 | 72.32 | 68.89 |
| RankT5 | T5 | 3B | 79.07 | 75.66 | 72.95 | 80.86 | 73.05 | 69.63 |
| **Unsupervised LLM Methods** | | | | | | | | |
| LRL | text-davinci-003 | 175B | - | - | **65.80** | - | - | 62.24 |
| RankGPT | gpt-3 | 175B | 50.78 | 50.77 | 49.76 | 50.00 | 48.36 | 48.73 |
| RankGPT | text-davinci-003 | 175B | 69.77 | 64.73 | 61.50 | 69.75 | 58.76 | 57.05 |
| UPR | FLAN-T5-XXL | 11B | 62.79 | 62.07 | 62.00 | 64.20 | 62.05 | 60.34 |
| RG | FLAN-T5-XXL | 11B | 67.05 | 65.41 | 64.48 | 65.74 | 66.40 | 62.58 |
| UPR | FLAN-UL2 | 20B | 53.10 | 57.68 | 58.95 | 64.81 | 61.50 | 60.02 |
| RG | FLAN-UL2 | 20B | **70.93** | **66.81** | 64.61 | **75.62** | 66.85 | **65.39** |
| **Ours** | | | | | | | | |
| Knowledge Distill | BERT | 340M | 69.88 | 66.29 | 62.22 | 70.79 | **67.00** | 64.62 |
| Causal Retrieval | gpt-2 | 1.5B | 64.45 | 58.54 | 57.58 | 68.15 | 62.99 | 60.77 |
| Causal Retrieval | Llama-3 | 8B | 68.66 | 65.21 | 63.21 | 69.95 | 66.31 | 65.11 |

Table 3: Results are reported on the TREC-DL2019 and TREC-DL2020 datasets by re-ranking the top 100 documents initially retrieved using BM25. Knowledge Distillation refers to the process in which BERT-340M is distilled using Llama 3-8B. The highest performance is highlighted in bold, while the second-best is marked with an underline.

encoder re-ranker built on BERT-large for relevance estimation; monoT5 (Nogueira et al., 2020), which leverages T5 in a sequence-to-sequence framework to compute relevance scores using pointwise ranking loss; and RankT5 (Zhuang et al., 2023), an extension of T5 that incorporates listwise ranking loss to enhance performance. **2) Unsupervised LLM Methods:** Unsupervised Passage Re-ranker (UPR) (Sachan et al., 2022), which employs query generation in a pointwise manner; Relevance Generation (RG) (Liang et al., 2023), a relevance-focused pointwise approach; RankGPT (Sun et al., 2023a), a listwise ranking method using GPT-based large language models (LLMs); and Listwise Reranker with a Large Language Model (LRL) (Ma et al., 2023), a listwise approach similar to RankGPT but with a distinct prompt design.

#### 5.2.2 Overall Results

As shown in Table 3, our proposed causal retrieval method achieves strong performance on the TREC-DL2019 and TREC-DL2020 tasks. On TREC-DL2019, causal retrieval using LLaMA-3 attains an NDCG@1 score of 68.66, exceeding BM25 by more than 26% and outperforming most unsupervised methods, including RankGPT (text-davinci-003). On TREC-DL2020, the NDCG@1 score further improves to 69.95 with causal retrieval using LLaMA-3. Moreover, fine-tuning the BERT model with knowledge distilled from LLaMA-3 raises the NDCG@1 score to 70.79, demonstrating that BERT significantly benefits from knowledge distillation while maintaining high computational efficiency. These results highlight the effectiveness of causal retrieval, particularly in scenarios with limited labeled data, as it combines strong retrieval capabilities with efficient computation.

| top-k | Method | EM | F1 | Acc |
|---|---|---|---|---|
| 3 | Causal Retrieval | **54.68** | **61.34** | **53.26** |
| | w/o $p_\theta(K)$ | 31.49 | 33.30 | 31.08 |
| | w/o prompt | 53.79 | 59.43 | 51.10 |
| 4 | Causal Retrieval | **54.12** | **60.82** | **52.78** |
| | w/o $p_\theta(K)$ | 34.95 | 36.60 | 34.42 |
| | w/o prompt | 53.79 | 60.78 | 52.23 |
| 6 | Causal Retrieval | 51.71 | **58.40** | **50.22** |
| | w/o $p_\theta(K)$ | 39.51 | 41.76 | 38.64 |
| | w/o prompt | **52.01** | 57.66 | 49.42 |

Table 4: Ablation study on 2Wiki using Llama3-8B.

#### 5.2.3 Effectiveness of Causal Retrieval

Causal Retrieval methods, leveraging causal inference principles with LLaMA-3 and GPT-2, demonstrate their potential as scalable alternatives to traditional ranking techniques. Despite using smaller

LLMs such as GPT-2 (1.5B) and LLaMA-3 (8B), our methods achieve competitive or superior performance compared to larger unsupervised models. As shown in Table 3, Causal Retrieval with LLaMA-3 improves NDCG@1 by 26.5% over BM25 on TREC-DL2019, surpassing several larger models like FLAN-T5-XXL (11B). On TREC-DL2020, it achieves an NDCG@1 score 21% higher than BM25, closely matching supervised methods. Additional analysis in Appendix G further illustrates how our methods prioritize semantically relevant documents, often outperforming baselines such as BM25 in retrieving meaningful results in challenging scenarios.

#### 5.2.4 Knowledge Distillation for Efficient Retrieval

Knowledge distillation plays a key role in enhancing the efficiency of retrieval models without compromising performance. As shown in Table 3, fine-tuning BERT (340M) with distilled knowledge from LLaMA-3 achieves strong results, with an NDCG@1 of 69.88 on TREC-DL2019 and 70.79 on TREC-DL2020. By transferring the causal retrieval capabilities of LLaMA-3 to BERT, we demonstrate that compact models can achieve competitive performance even in scenarios with limited computational resources.

The effectiveness of this distillation process lies in its ability to retain the semantic understanding and ranking capabilities of larger models while reducing overfitting and enhancing generalization. For example, on TREC-DL2020, the distilled BERT model outperforms many unsupervised methods and approaches the performance of supervised models like monoT5. This highlights the potential of knowledge distillation as a practical solution for deploying high-performing retrieval systems in resource-constrained environments.

#### 5.2.5 Analysis of Model Limitations

As shown in Table 3, while Causal Retrieval achieves a balance between efficiency and effectiveness, it still falls short of some larger models like FLAN-T5-XXL and RankGPT. One key limitation is that smaller models, despite leveraging causal inference, struggle to fully capture the complex semantic relationships that larger models learn through extensive parameterization. Additionally, the knowledge distillation process, while effective in transferring insights, may lead to some information loss, preventing distilled models from fully replicating the performance of their larger counterparts.

Moreover, Causal Retrieval's reliance on causal assumptions, while improving robustness, may impose constraints that limit its ability to leverage deep contextual representations. This trade-off means that while it performs well in many cases, it does not always surpass the best supervised methods. Future improvements could involve refining causal modeling techniques or integrating hybrid approaches that combine causal inference with more expressive neural ranking architectures.

### 5.3 Ablation Study

We analyzed the impact of two types of removed components on model performance. Specifically, w/o $p_\theta(K)$ means not excluding the relevant text provided by the large model itself, while w/o prompt means removing the prompt provided to the model.

As shown in Table 4, the Causal Retrieval method consistently outperforms the ablation methods across all top-k values. Removing $p_\theta(K)$ leads to a significant drop in EM, F1, and Acc by about 40%, while removing the prompt results in a smaller, but noticeable, 5% decline. This suggests that the texts provided by the large model are crucial for retrieval accuracy, and the prompt plays an important, though less critical, role in guiding model generation.

## 6 Conclusion

This paper explores causal knowledge retrieval to enhance grounding in large language models. Specifically, we enhance retrieval-augmented generation by prioritizing causal relevance between questions and text snippets during the retrieval process. A novel CIS is introduced to measure this relevance, utilizing the capabilities of autoregressive models to model textual entailments. A knowledge distillation method is further introduced to enable cost-effective CIS calculation. Our comprehensive experiments demonstrate that simply replacing traditional similarity metrics with our causal relevance metric can significantly reduce the retrieval of redundant documents and enhance performance. This improvement boosts the quality of retrieved documents and increases answer accuracy.

## 7 Limitations

Despite its effectiveness, the causal retrieval method has limitations. First, its high computational cost makes it less suitable for real-time or resource-constrained applications. Second,while reducing redundancy, the method may overlook diverse documents that contribute to query understanding. Future work will explore hybrid metrics combining causal and similarity-based approaches.

## References


Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Deng Cai, Yan Wang, Lemao Liu, and Shuming Shi. 2022. Recent advances in retrieval-augmented text generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3417–3419, New York, NY, USA. Association for Computing Machinery.

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2020. Overview of the trec 2020 deep learning track. In *TREC*.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen Voorhees. 2019. Overview of the trec 2019 deep learning track. In *TREC 2019*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

OpenAI et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6095–6107.

Nihal Jain, Dejiao Zhang, Wasi Ahmad, Zijian Wang, Feng Nan, Xiaopeng LI, Ming Tan, Ramesh Nallapati, Baishakhi Ray, Parminder Bhatia, Xiaofei Ma, and Bing Xiang. 2023. Contraclm: Contrastive learning for causal language model. In *ACL 2023*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7036–7050, Mexico City, Mexico. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

O. Khattab, Keshav Santhanam, Xiang Lisa Li, David Leo Wright Hall, Percy Liang, Christopher Potts, and Matei A. Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *ArXiv*, abs/2212.14024.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020a. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Filippo Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Dandan Li, Ziyu Guo, Qing Liu, Li Jin, Zequn Zhang, Kaiwen Wei, and Feng Li. 2023. Click: Integrating causal inference and commonsense knowledge incorporation for counterfactual story generation. *Electronics*, 12(19).

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model. *Preprint*, arXiv:2305.02156.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2023. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage re-ranking with bert. *Preprint*, arXiv:1901.04085.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. *Cornell University - arXiv,Cornell University - arXiv*.

Jayr Alencar Pereira, Robson do Nascimento Fidalgo, Roberto de Alencar Lotufo, and Rodrigo Nogueira. 2022. Visconde: Multi-document qa with gpt-3 and neural reranking. In *European Conference on Information Retrieval*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518, Mexico City, Mexico. Association for Computational Linguistics.

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

Raviteja Anantha Ramesh, Tharun Bethi, Danil Vodianik, and Srinivas (Vasu) Chappidi. 2023. Context tuning for retrieval augmented generation. In *EACL Workshop*.

Devendra Singh Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen tau Yih, Joëlle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Conference on Empirical Methods in Natural Language Processing*.

Alireza Salemi and Hamed Zamani. 2024. Evaluating retrieval quality in retrieval-augmented generation. *Preprint*, arXiv:2404.13781.

Anton Shapkin, Denis Litvinov, Yaroslav Zharov, Egor Bogomolov, Timur Galimzyanov, and Timofey Bryksin. 2024. Dynamic retrieval-augmented generation. *Preprint*, arXiv:2312.08976.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023a. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore. Association for Computational Linguistics.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2023b. Recitation-augmented language models. In *The Eleventh International Conference on Learning Representations*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *ArXiv*, abs/1905.12616.

Yuzhe Zhang, Yipeng Zhang, Yidong Gan, Lina Yao, and Chen Wang. 2024a. Causal graph discovery with retrieval-augmented generation based large language models. *Preprint*, arXiv:2402.15301.

Zihan Zhang, Meng Fang, and Ling Chen. 2024b. Retrievalqa: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. *Preprint*, arXiv:2402.16457.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2023. Rankt5: Fine-tuning t5 for text ranking with ranking losses. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2308–2313, New York, NY, USA. Association for Computing Machinery.

## A Knowledge Distillation for Efficient Document Retrieval

We present the experimental setup and results of fine-tuning the BERT model for document retrieval using the TREC-DL2019 and TREC-DL2020 datasets. Training data is generated by retrieving the top 100 documents for each query using BM25. For each query-document pair, we compute the CIS value using LLaMA3 and then fine-tune `bert-large-uncased` based on the method described in Section 4.2.

After fine-tuning, we compare the computation time for calculating the CIS values between the fine-tuned BERT model and the original model, using different token counts to evaluate performance. The results demonstrate that the fine-tuned BERT model, trained using distilled knowledge from LLaMA3, significantly reduces computation time compared to the original BERT model while maintaining similar retrieval performance, as shown in Table 3.

Figure 3 presents the computation time (in seconds) for different methods and token counts. The fine-tuned BERT model outperforms causal retrieval in terms of computation time, especially as the token count increases. For instance, when the token count is 500, the fine-tuned BERT model requires only 0.22 seconds, while causal retrieval takes much longer. This performance advantage becomes more significant as the token count increases, further demonstrating the efficiency of the fine-tuned BERT model.

These results validate the effectiveness of distilling knowledge from a large, computationally expensive model (LLaMA3) into a lightweight BERT model, confirming that this approach is not only efficient but also maintains high retrieval accuracy.

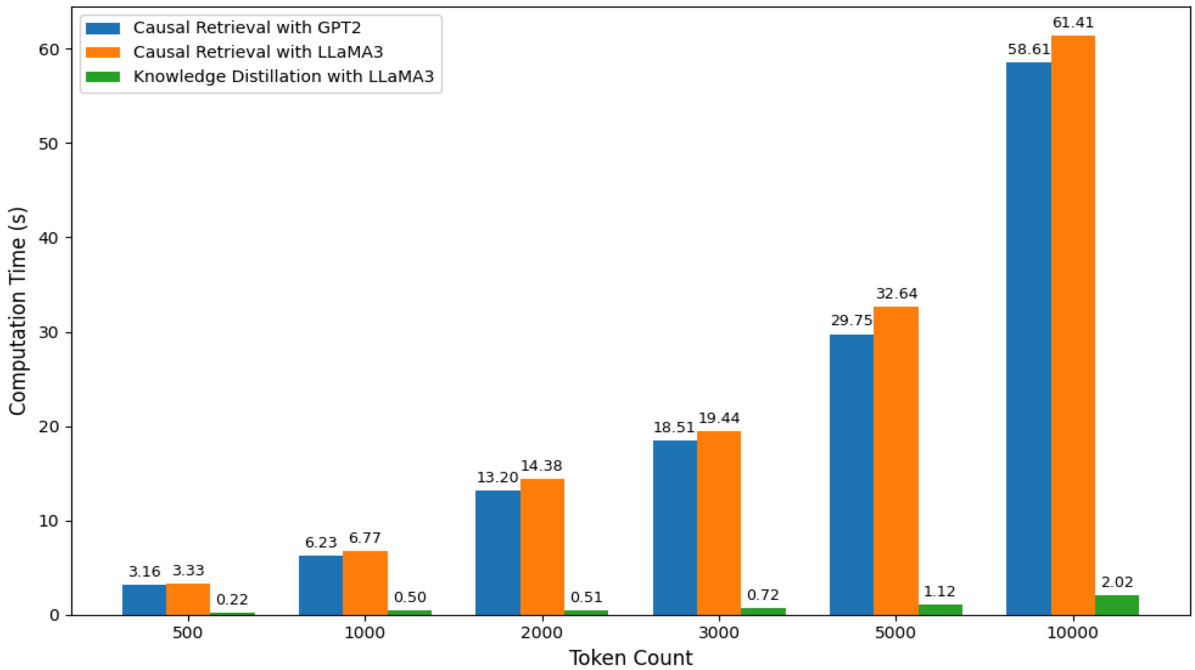


Figure 3: The bar chart illustrates the computation time for different methods (Causal Retrieval and Knowledge Distillation) across various token counts (500, 1000, 2000, 3000, 5000, 10000). It is evident that the computation time for causal retrieval increases linearly with the token count, whereas fine-tuning BERT using training data distilled from LLaMA 3-8B demonstrates significantly lower computation times and a smaller growth rate. The exact computation time values are annotated above each bar for better comparison.

## B An Information Theory Look at CIS

The CIS can be understood from the perspective of information theory, similar to pointwise mutual information (PMI). In information theory, the PMI between two random variables $X$ and $Y$ is defined as:

$$\mathrm{PMI}(X,Y) = \log \frac{p(X,Y)}{p(X)p(Y)} = \log \frac{p(Y|X)}{p(Y)} = \log \frac{p(X|Y)}{p(X)}$$

This measures the association between $X$ and $Y$, quantifying how much knowing one of the variables reduces uncertainty about the other. It is symmetric, reflecting the mutual dependence between the two variables.

In contrast, $\text{CIS}_\theta(Q \rightarrow K)$ is inherently asymmetric: it captures how well $K$ can be inferred given $Q$, but not necessarily the reverse. This asymmetry is intentional and crucial for our goal: to prioritize knowledge that causally and directionally addresses the query, rather than merely finding an overlap. This makes CIS a powerful tool for RAG, where the goal is not just to find related information, but to find information that directly supports answering the query.

## C Error Analysis of CIS

We performed a thorough human error analysis of CIS and show representative cases in Table 5. The breakdown of 20 error cases reveals that 50% are due to answer extraction issues, including wrong extraction (25%) and partial matching with the true answer (25%). The remaining 50% are information processing problems, such as redundant information (20%), partial answers (15%), and ambiguous questions (5%). For instance, in the question "In which year and in which country did the first moon landing take place?", the system retrieves the correct document about the 1969 Apollo 11 moon landing but mixes in irrelevant details, like the Soviet Union's unmanned missions, resulting in a cluttered answer. This highlights the need for improvements in answer extraction and information processing to enhance accuracy and relevance. Enhancing question understanding, extracting the most pertinent information, and filtering out redundancy through improved algorithms are crucial for providing concise and accurate answers.

| **Breakdown of 20 failure cases** | |
|---|---|
| Incorrect Retrieval | 2 |
| Incorrect Answer Extraction | 5 |
| Partial Answer | 3 |
| Ambiguous Question | 1 |
| Redundant Information | 4 |
| Partial match with the groundtruth | 5 |

Table 5: The error analyses of causal knowledge retrieval experiment. Randomly select the experimental results, input the question, retrieved documents, correct answer, and the output answer of the large model into GPT4.0 to determine the reason for the wrong answer.

## D Knowledge Retrieval Results

The ability of LLMs to accurately answer domain-specific queries depends heavily on including all necessary information in the prompt context. LLMs that are prone to hallucinating questions have difficulty providing correct answers when critical information is missing. In the absence of relevant data, LLMs may default to using their existing knowledge base, which often results in incorrect responses. To evaluate how the retrieved knowledge covers this necessary information, we conducted experiments regarding different retrieval strategies. The results are shown in Table 6.

In the experiments, the Causal Retrieval methods, especially the Llama3-8B model, showed significant performance improvements over traditional base methods such as BM25 and Dense. For example, in the Top-k = 6 setting, the Recall rate of Llama3-8B increased by 7.55%, and the Precise rate increased by $6.41\%$. The GPT-2 model also showed advantages, with its Recall rate increased by $5.77\%$ and Precise rate increased by $4.83\%$ in the Top-k = 6 setting. Similarly, fine-tuning BERT using training data distilled from LLaMA 3-8B demonstrated strong performance, achieving a Recall rate of 60.90 and a Precise rate of 58.29 on the 2Wiki dataset, which outperformed BM25 by $16.78\%$ and $13.17\%$, respectively.

The recall of the causal retrieval approach is 13.6% higher than that of BM25 and 10.1% higher than that of the Dense approach, showing its effectiveness. Notably, BERT also surpassed Dense by $6.80\%$

| Top-k | Retrieval Strategies | MuSiQue | | HotpotQA | | 2Wiki | |
|---|---|---|---|---|---|---|---|
| | | Recall rate | Precise rate | Recall rate | Precise rate | Recall rate | Precise rate |
| 3 | BM25 | 49.88 | 50.70 | 49.65 | 50.42 | 44.11 | 43.67 |
| | Dense | 51.33 | <u>52.31</u> | 51.53 | 52.53 | 47.72 | 47.05 |
| | Hybrid | 46.16 | 45.80 | 49.01 | 49.77 | 49.45 | 50.47 |
| | Hybrid+Rerank | 48.34 | 47.72 | 54.17 | <u>56.05</u> | 54.19 | 55.56 |
| | Causal Retrieval(GPT-2) | <u>51.98</u> | 50.69 | <u>55.22</u> | **57.29** | 49.56 | 46.01 |
| | Causal Retrieval(Llama3-8B) | **53.10** | **53.32** | **56.52** | 54.78 | <u>63.75</u> | <u>60.66</u> |
| | Knowledge Distill(BERT) | 47.96 | 50.32 | 53.50 | 52.09 | **65.26** | **62.02** |
| 4 | BM25 | 50.73 | 51.17 | 50.73 | 51.32 | 44.25 | 43.79 |
| | Dense | 52.65 | 53.62 | 52.97 | 54.06 | 48.32 | 47.62 |
| | Hybrid | 49.62 | 48.15 | 53.13 | 53.76 | 54.95 | 53.69 |
| | Hybrid+Rerank | 52.77 | 53.74 | 55.54 | **57.21** | 48.67 | 48.06 |
| | Causal Retrieval(GPT-2) | <u>54.11</u> | <u>53.77</u> | 57.60 | 55.27 | 50.23 | 48.49 |
| | Causal Retrieval(Llama3-8B) | **57.54** | **54.12** | **58.61** | <u>56.34</u> | <u>63.33</u> | <u>60.16</u> |
| | Knowledge Distill(BERT) | 53.96 | 52.71 | <u>58.11</u> | 56.21 | **64.83** | **61.77** |
| 6 | BM25 | 51.88 | 52.74 | 48.99 | 49.87 | 46.12 | 45.52 |
| | Dense | 54.35 | 55.12 | 57.02 | 58.23 | 49.54 | 49.12 |
| | Hybrid | 53.11 | 54.14 | 58.39 | 59.47 | 55.01 | 54.01 |
| | Hybrid+Rerank | 55.63 | 56.78 | 61.70 | <u>62.27</u> | 55.34 | 56.32 |
| | Causal Retrieval(GPT-2) | <u>57.20</u> | <u>56.85</u> | <u>62.17</u> | 60.85 | 53.21 | 50.92 |
| | Causal Retrieval(Llama3-8B) | **59.43** | **59.15** | **64.50** | **62.52** | **60.93** | <u>57.75</u> |
| | Knowledge Distill(BERT) | 54.75 | 56.12 | 58.81 | 55.72 | <u>60.90</u> | **58.29** |

Table 6: Question answering results of different top-k using different search strategies. Knowledge Distillation refers to the process in which BERT-340M is distilled using Llama 3-8B. The best result is in bold, and the second-best result is underlined.

in Recall rate and $10.72\%$ in Precise rate, further validating the importance of fine-tuning for improving retrieval quality.

## E Question Answering Prompts

After identifying the document with the highest CIS $K^*$, we use a structured instruction to guide the LLM in generating the final answer:

> You are an information specialist. You are given a question and a document is retrieved based on the question. Your task is to answer the question using only the information from the document. Your answer must be brief and directly address the question. Do not add any additional text or context. Here are some examples of how to respond based on provided information.
>
> Question: example of a problem
> Retrieved document: documentation examples
> Answer: example answer
>
> Question: {question}
> Retrieved document: {search_document}
> Answer:

## F QA Case Study

We conduct a case study to compare our causal retrieval approach with traditional RAG. In Table 7, we present specific reasoning examples from multi-hop datasets.

In the HotpotQA dataset, the causal retrieval approach accurately identifies the correct publisher "PUBG Corp" by retrieving documents related to "Battle Royale Game" that include information about the game publisher. In contrast, traditional RAG sometimes generates incorrect responses by retrieving information related to "Battle Royale Game" but not directly relevant, such as "Epic Games," leading to incorrect answers. This demonstrates that the causal retrieval approach can more accurately answer complex questions by identifying the causal relationship between the question and the text fragment using pointwise mutual information, whereas traditional RAG often produces inaccurate results due to information confounding and lack of context.

| Datasets | Queries | Traditional RAG | Causal Retrieval (Ours) |
|---|---|---|---|
| **MuSiQue** | When did Australian troops fight in the area where Stag Arms's headquarter locates? | **Retrieval strategy:** Hybrid **Reasoning:** The Australian troops fought in August 1915 because the retrieved document mentioned this date in a relevant context, although it does not explicitly confirm the location as Stag Arms's headquarters. So the answer is: August 1915. | **Retrieval strategy:** Causal Retrieval **Reasoning:** The Australian troops fought in October 1944 because it is explicitly mentioned that they were involved in combat in New Britain, which is identified as the location of Stag Arms's headquarters. So the answer is: October 1944. |
| **HotpotQA** | Which company published the battle royale game that Summit1g started streaming? | **Retrieval strategy:** Hybrid **Reasoning:** Based on the retrieved information, while Summit1g streamed "PlayerUnknown's Battlegrounds", the publisher's details were not specified, leading to an incorrect assumption that the publisher might be Epic Games, known for another popular battle royale game. So the answer is: Epic Games. | **Retrieval strategy:** Causal Retrieval **Reasoning:** Summit1g's streamed game "PlayerUnknown's Battlegrounds" is published by PUBG Corp, directly supporting the correct answer. So the answer is: UBG Corp. |
| **2Wiki** | Where was the place of death of the director of the film Yaarukkaga Azhudhaan? | **Retrieval strategy:** Hybrid **Reasoning:** The retrieved information identifies Jayakanthan as the director of Yaarukkaga Azhudhaan. However, the place of his death was not specified in this context, leading to an incorrect conclusion. So the answer is: New Delhi. | **Retrieval strategy:** Causal Retrieval **Reasoning:** By cross-referencing detailed information about Jayakanthan, it is established that he died in Madras. This specific fact is crucial for answering the question accurately. So the answer is: Madras. |

Table 7: Case study with Llama3-8B, where we present the factual error in red and the accurate information in blue.

## G An example of Causal Retrieval in Information Retrieval (IR)

To verify the effectiveness of the causal retrieval method in information retrieval tasks, we demonstrate the process using a specific query. This approach first uses the BM25 model to retrieve the top 100 documents related to the query from a document corpus. Then, the causal retrieval method is applied to model the causal relationship between the query and each document, calculating their CIS. Finally, the documents are re-ranked based on their CIS values to optimize the relevance and interpretability of the retrieval results. Below is an example query from TREC-DL2019 used for retrieval and re-ranking:

**Query:** *how is the weather in jamaica*

### G.1 Retrieval Result

Table 8 shows the top 3 documents retrieved using BM25, along with their corresponding BM25 scores, CIS, and relevance scores.

| Rank | Document | BM25 score | CIS | relevance |
|---|---|---|---|---|
| 1 | D2301225 | 8.30 | 1.94 | 1 |
| 2 | D441607 | 8.21 | 4.89 | 3 |
| 3 | D1318068 | 8.05 | 2.94 | 2 |

Table 8: Top 3 documents retrieved by BM25, re-ranked by CIS. A higher relevance score indicates a stronger relationship between the document and the query.

**Document D2301225:**
We had it down to Jamaica or the Bahamas and having read the post from dlmcdon214 with the same dilemma about which to chose, we've decided on Jamaica. Now, all experts out there - what is the weather usually like around Christmas/New Year in Jamaica? Also, we're undecided yet about Negril or Ocho Rios - which please? Thanks in advance to anyone who helps out!!! Mentioned in this post Jamaica Caribbean Bahamas Caribbean Ocho Rios Jamaica Report inappropriate content Related: What are the most popular tours in Negril? Re: Christmas Weather in Jamaica Jul 6, 2005, 5:33 AMHi thewoolleys,I hear the weather is still in the mid to late 20s in dec..... I going to mobay on dec 8 for my wedding staying at the wyndham rose hall. Cant wait. Report inappropriate contentthewoolleyskent england Level Contributor303 posts Save Reply2. Re: Christmas Weather in Jamaica Jul 6, 2005, 5:48 AMThanks Janlo - have a great holiday and a fab wedding - wishing you lots of happiness in wedded bliss!!!! Hubby and I got married in Las Vegas 18 months ago after 15 years together - it was, and still is, the best thing we've ever done!We're looking at staying in either Club Hotel Rui Negril or the new Riu opening in Ocho Rios...

BM25 score: 8.30
CIS: 1.94
relevance: 1

**Document D441607:**
This is Jamaica weather! Most of our days are filled with warmth and sunshine, even during the rainy season. Jamaica has a tropical climate with hot and humid weather at sea level. The higher inland regions have a more temperate climate. (Bring a light jacket just in case you travel to the mountains where temperatures can be 10 degrees cooler or in case you go on a windy boat ride). Our average annual temperature is between 80-860̆0b0F (27-300̆0b0C). The coolest months are January and February and the temperature starts going back up in March. July and August are typically the hottest months. Temperature variations between summer and winter is about 10 degrees. The rainiest months in Jamaica are normally May-June and September-October (lasts until November sometimes). Enjoying the Jamaica weather ...even when it's raining!This so-called rainy season is characterized by brief afternoon showers followed by sunshine. Look at it as a welcome break from the tropical heat! ( The family in the photo seem to agree! )Jamaica's average annual rainfall is 50.7 inches (1,288 mm). However, the distribution of rainfall is quite uneven across the island. (You may want to grab a map of Jamaica to find your bearings...

BM25 score: 8.21
CIS: 4.89
relevance: 3

**Document D1318068:**
Weather experience please! 8 Sep 2010, 00:28Hi everyone,My husband and I are traveling to JA next week leaving on Sept 16. I'm worried about what the weather is going to be like. I've looked at weather.com which says 80's and scattered or isolated T-storms, and Storm carib gives daily information which doesnt really help me out for next week. I just want to know if anyone has traveled there during September and if so, how was the weather? Raining all day, some of the day, would never go in September etc.. Please help!!!!! I'm kinda freaking out about it and ready to cancel the trip!Report inappropriate content Related: What are the most popular tours in Negril? Weather experience please! !8 Sep 2010, 00:46don't cancel!!!! i'm going on the 21st...and sure, you may get your normal afternoon rain for an hour or so, but WHO CARES?!? it's where

you ARE that's important. i don't know how long you are going for, but i highly doubt you will have an ENTIRE week of constant rain...and from what i've seen on the wunderground website, it doesn't look like there is any hurricane threat for the week we will be there. have a dirty banana, or a red stripe and have a GREAT time!Report inappropriate contentforce10JCHouston...

BM25 score: 8.05
CIS: 2.94
relevance: 2