Unbiased Reasoning for Knowledge-Intensive Tasks in Large Language Models via Conditional Front-Door Adjustment

Bo Zhao*
zbo@webmail.hzau.edu.cn
Huazhong Agricultural University
Wuhan, China

Yinghao Zhang*
yhzhang@mail.hzau.edu.cn
Huazhong Agricultural University
Wuhan, China

Ziqi Xu†
ziqi.xu@rmit.edu.au
RMIT University
Melbourne, Australia

Yongli Ren
yongli.ren@rmit.edu.au
RMIT University
Melbourne, Australia

Xiuzhen Zhang
xiuzhen.zhang@rmit.edu.au
RMIT University
Melbourne, Australia

Renqiang Luo
lrenqiang@jlu.edu.cn
Jilin University
Changchun, China

Zaiwen Feng†
zaiwen.feng@mail.hzau.edu.cn
Huazhong Agricultural University
Wuhan, China

Feng Xia
f.xia@ieee.org
RMIT University
Melbourne, Australia

Abstract

Large Language Models (LLMs) have shown impressive capabilities in natural language processing but still struggle to perform well on knowledge-intensive tasks that require deep reasoning and the integration of external knowledge. Although methods such as Retrieval-Augmented Generation (RAG) and Chain-of-Thought (CoT) have been proposed to enhance LLMs with external knowledge, they still suffer from internal bias in LLMs, which often leads to incorrect answers. In this paper, we propose a novel causal prompting framework, Conditional Front-Door Prompting (CFD-Prompting), which enables the unbiased estimation of the causal effect between the query and the answer, conditional on external knowledge, while mitigating internal bias. By constructing counterfactual external knowledge, our framework simulates how the query behaves under varying contexts, addressing the challenge that the query is fixed and is not amenable to direct causal intervention. Compared to the standard front-door adjustment, the conditional variant operates under weaker assumptions, enhancing both robustness and generalisability of the reasoning process. Extensive experiments across multiple LLMs and benchmark datasets demonstrate that CFD-Prompting significantly outperforms existing baselines in both accuracy and robustness. The source code and case study are available at: https://github.com/zbb79/CFD-Prompting.

CCS Concepts

• **Information systems** → **Question answering**; • **Computing methodologies** → **Causal reasoning and diagnostics**.

*Both authors contributed equally to this research.
†Corresponding author.

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761103

Keywords

Large Language Models, Causal Inference, Knowledge-Intensive Tasks, Chain-of-Thought, Question Answering

ACM Reference Format:
Bo Zhao, Yinghao Zhang, Ziqi Xu, Yongli Ren, Xiuzhen Zhang, Renqiang Luo, Zaiwen Feng, and Feng Xia. 2025. Unbiased Reasoning for Knowledge-Intensive Tasks in Large Language Models via Conditional Front-Door Adjustment. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3746252.3761103

1 Introduction

In recent years, Large Language Models (LLMs) have achieved remarkable progress in natural language processing. By pre-training on massive text corpora, they have demonstrated impressive capabilities in language understanding and generation. Techniques such as in-context learning [4] and Chain-of-Thought (CoT) [40] have enabled LLMs to make significant advancements in tasks such as question answering and behaviour simulation [23]. However, LLMs still face critical challenges in knowledge-intensive tasks, as accurate answers often require specific information that falls outside the distribution of their internal knowledge [47].

To efficiently access external knowledge, relying solely on fine-tuning incurs significant computational costs and is further constrained by limited timeliness [48]. A promising alternative is incorporating external knowledge directly into prompts [50], for example, through Retrieval-Augmented Generation (RAG) [19] or knowledge graphs [33], which enable LLMs to access more comprehensive and up-to-date information. However, simply injecting external knowledge into prompts does not guarantee that LLMs can identify and utilise relevant information [31]. Recent studies further find that internal bias in LLMs can lead to spurious correlations with the query, thereby preventing the models from effectively leveraging external information to generate accurate answers [24].

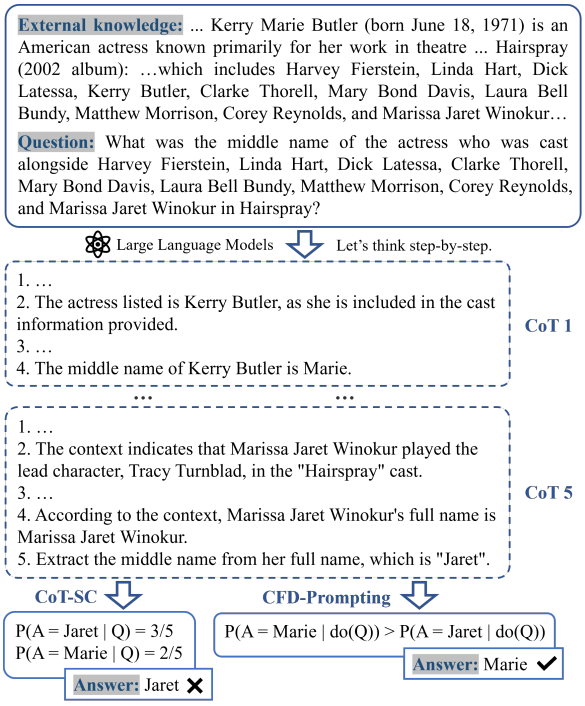

Figure 1: An example illustrating internal bias in LLMs. CoT-SC [38] selects the most frequent answer among sampled CoTs, which results in incorrect answer. In contrast, CFD-Prompting selects the answer with the highest causal effect and yields the correct result. This example is taken from a response by GPT-3.5 Turbo on HotpotQA.

To address this challenge, CoT self-consistency (CoT-SC) [38] improves reasoning by sampling multiple CoTs and selecting the most frequent answer. While effective against random errors, it fails to correct spurious correlations caused by internal bias in LLMs. As a stronger alternative, causality-based prompting methods estimate the causal effect of either the query or the CoT on the answer to select more reliable reasoning paths [44, 49]. As shown in Figure 1, we present an illustrative example comparing majority voting with our framework. The latter produces more accurate answers by ranking the causal effect of the query on the answer.

While prior work has explored both general reasoning and causality-based methods, their differences have not been formally characterised. We introduce a set of Structural Causal Models (SCMs) to clarify the assumptions and limitations of each method. As shown in Figure 2a, general methods perform direct reasoning without CoTs, where the latent confounder U induces spurious correlations between the query and the answer, often leading to incorrect answers. Figure 2b illustrates CoT and CoT-SC, which enhance performance through explicit reasoning steps but still suffer from bias due to U. Causal prompting (CP) mitigates this bias via the standard front-door adjustment [49], but it relies on strong assumptions, namely the absence of any observed confounders that interact with the CoT. To relax these constraints, Wu et al. [44] propose DeCoT for knowledge-intensive tasks (as shown in Figure 2c), which leverages

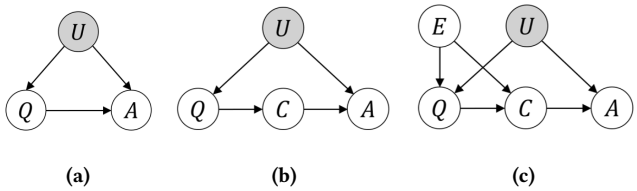

Figure 2: Three SCMs representing reasoning in LLMs: (a) general reasoning without CoTs; (b) CoT and CoT-SC incorporate explicit reasoning, and CP applies the standard front-door adjustment; (c) DeCoT and the proposed CFD-Prompting, which are specifically for knowledge-intensive tasks. Here, Q is the query, A is the answer, C is the CoT, U is the latent confounder, and E is the observed external knowledge.

external knowledge as an instrumental variable to estimate the average causal effect of the CoT on the answer. However, this method provides only a coarse estimate and may overlook fine-grained causal effects between the query and the answer. Thus, there is a clear need for a causal prompting framework that can provide an unbiased estimate of the causal effect of the query on the answer to improve performance on knowledge-intensive tasks.

In this work, we propose the Conditional Front-Door Prompting (CFD-Prompting) framework, which leverages conditional front-door adjustment to mitigate internal bias and generate more reliable answers for knowledge-intensive tasks. As a relaxed variant of the standard front-door criterion, it allows interactions between CoTs and external knowledge, making it better suited to such tasks. To implement this, we generate counterfactual external knowledge to simulate causal interventions on the query. CFD-Prompting adopts an encoder-based architecture that does not require access to model logits, enabling compatibility with closed-source LLMs. The contributions of this paper are summarised as follows:

- We present a causal analysis of LLM reasoning using structural causal models, offering a theoretical foundation for de-biasing answers in knowledge-intensive tasks.
- We propose CFD-Prompting, a general and logit-free causal prompting framework that supports both open-source and closed-source LLMs, and relaxes the assumptions of standard front-door methods by allowing interactions between CoTs and external knowledge.
- We conduct extensive experiments across multiple LLMs and benchmark datasets, demonstrating that CFD-Prompting consistently outperforms state-of-the-art prompting baselines in both accuracy and robustness.

2 Preliminaries

We use capital letters to denote variables and lowercase letters to denote their values. Due to space limitations, we refer readers to [25] for the fundamental definitions of causality, including directed acyclic graphs (DAGs), the Markov condition, faithfulness, d-separation, and d-connection.

2.1 Structural Causal Model

The structural causal model (SCM) [26] formalises causal relationships between variables using a directed acyclic graph (DAG) and a set of structural equations. In a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, $\mathcal{V}$ denotes the set of nodes (variables), and $\mathcal{E}$ denotes the set of directed edges, where an edge $\mathcal{V}_i \to \mathcal{V}_j$ indicates that $\mathcal{V}_i$ is a direct cause of $\mathcal{V}_j$.

A *path* π between nodes $\mathcal{V}_1$ and $\mathcal{V}_n$ is a sequence of distinct nodes $\langle \mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_n \rangle$ such that each consecutive pair $(\mathcal{V}_i, \mathcal{V}_{i+1})$ is adjacent in the graph. A node $\mathcal{V}$ is said to lie on the path π if it appears in the sequence. A path π is called *causal* if all edges along it follow the same direction, i.e., $\mathcal{V}_1 \to \mathcal{V}_2 \to \cdots \to \mathcal{V}_n$; otherwise, it is referred to as a non-causal path.

As illustrated in Figure 2a, we denote the query as Q, which includes both demonstrations and test examples provided to the LLM. The predicted answer generated by the LLM in response to the query is denoted as A. Since LLMs generate answers directly based on the given Q, we represent the direct causal effect from query to answer as $Q \to A$. However, during the pre-training phases, LLMs may internalise spurious correlations between surface-level patterns and output distributions. These correlations, often inherited from large-scale web corpora or task-specific fine-tuning datasets, can manifest as implicit biases in downstream reasoning [1, 13].

To model this phenomenon, we introduce an unobservable variable U, which captures the internal bias in LLMs. In this case, although $Q \to A$ holds as a structural dependency, the true causal relationship is confounded by U, which influences the generation of a reasonable answer. This is formally captured by the back-door path $Q \leftarrow U \to A$, indicating that the observed statistical association between Q and A is not purely causal. To estimate the true causal effect of Q on A, it is necessary to adjust for this confounder U. If the confounder U is observable, the causal effect could be adjusted using the back-door adjustment formula [25] as follows:

$$P(A \mid \mathrm{do}(Q)) = \sum_{u} P(A \mid Q, u) P(u). \tag{1}$$

However, in practice, since U is latent (i.e., unmeasured), alternative strategies such as front-door adjustment are required to obtain an unbiased estimate of the causal effect.

2.2 Front-door Adjustment

One prominent approach to addressing unobserved confounders is the standard front-door adjustment [25]. Unlike the back-door criterion, which requires blocking all back-door paths, the standard front-door criterion isolates the causal pathway through a suitable mediator, even in the presence of unobserved confounders. In the following, we outline the standard front-door criterion and describe how it can be adapted to mitigate internal bias in LLMs.

Definition 1 (Standard Front-Door Criterion [25]). *A set of variables Z_{SFD} is said to satisfy the (standard) front-door criterion relative to an ordered pair of variables (Q, A) in a DAG $\mathcal{G}$ if the following conditions hold: (1) Z_{SFD} intercepts all directed paths from Q to A; (2) there is no unblocked back-door path from Q to Z_{SFD}; (3) all back-door paths from Z_{SFD} to A are blocked by Q.*

Theorem 1 (Standard Front-Door Adjustment [25]). *If Z_{SFD} satisfies the standard front-door criterion relative to (Q, A), then the causal effect of Q on A is identifiable and is given by the following standard front-door adjustment formula:*

$$P(A|do(Q)) = \sum_{z_{\mathrm{SFD}},\, q} P(A \mid q, z_{\mathrm{SFD}}) P(z_{\mathrm{SFD}} \mid q) P(q). \tag{2}$$

The derivation of Equation 2 relies on the rules of do-calculus [25], which allows the systematic transformation of expressions involving the $do(\cdot)$ operator into observational probabilities under certain graphical conditions. The rules of do-calculus are as follows:

Theorem 2 (Rules of *do*-Calculus [25]). *Let $\mathcal{G}$ be the DAG associated with a structural causal model, and let $P(\cdot)$ denote the probability distribution induced by that model. For any disjoint subsets of variables Q, A, Z, and W, the following rules hold:*

- *Rule 1 (Insertion/deletion of observations): $P(A \mid do(A), Z, W) = P(A \mid do(Q), W)$, if $(A \perp\!\!\!\perp Z \mid A, W)$ in $\mathcal{G}_{\overline{Q}}$;*
- *Rule 2 (Action/observation exchange): $P(A \mid do(Q), do(Z), W) = P(A \mid do(Q), Z, W)$, if $(Y \perp\!\!\!\perp Z \mid Q, W)$ in $\mathcal{G}_{\overline{Q}\underline{Z}}$;*
- *Rule 3 (Insertion/deletion of actions): $P(A \mid do(Q), do(Z), W) = P(A \mid do(Q), W)$, if $(A \perp\!\!\!\perp Z \mid Q, W)$ in $\mathcal{G}_{\overline{Q},\overline{Z(W)}}$,*

where $Z(W)$ is the set of nodes in Z that are not ancestors of any node in W in $\mathcal{G}_{\overline{Q}}$.

Here, $\mathcal{G}_{\overline{Q}}$ denotes the DAGs obtained by removing all incoming edges into Q, while $\mathcal{G}_{\underline{Q}}$ denotes the graph obtained by removing all outgoing edges from Q. This notation generalises to any variable or set of variables, not limited to Q.

The standard front-door criterion provides a theoretical foundation for identifying causal effects even in the presence of unobserved confounders. In the context of LLMs, this insight motivates treating the CoT as a valid front-door variable for estimating the causal effect of the query on the answer . As illustrated in Figure 2b, C satisfies the conditions of the standard front-door criterion with respect to the causal effect of Q on A. This allows the causal effect $P(A \mid \mathrm{do}(Q))$ to be decomposed into two components: the effect of Q on C, and the effect of C on A conditional on Q. Specifically, the front-door adjustment formula is given by:

$$P(A \mid do(Q)) = \sum_{c} P(c \mid Q) \sum_{q} P(A \mid c, q) P(q). \tag{3}$$

CP is a causality-based prompting framework grounded in the SCM shown in Figure 2b. It assumes no observed variables interact with the CoT, simplifying the DAG to satisfy the standard front-door criterion [49]. However, this assumption limits its applicability to knowledge-intensive tasks, where external knowledge E often acts as an observed confounder influencing both the query and the CoT. In such cases, the CoT no longer meets the conditions for a valid front-door variable.

To address the limitations of standard front-door adjustment in knowledge-intensive tasks, we adopt the conditional front-door adjustment, which accounts for observed confounders through conditioning. This motivates our proposed CFD-Prompting framework, which treats the CoT as a conditional front-door variable and incorporates external knowledge to enable unbiased reasoning.

3 Method

In this section, we first outline the task and introduce the notations used throughout the paper. We then present our proposed

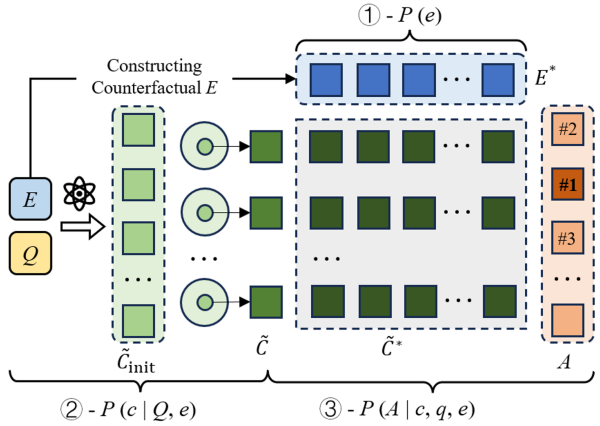

Figure 3: The overall architecture of CFD-Prompting. $\tilde{C}_{\text{init}}$ denotes the encoded CoT generated by the LLM given the query Q and external knowledge E; $\tilde{C}$ is the encoded CoT after applying k-means; $\tilde{C}^*$ represents the encoded CoT generated using the counterfactual variant of external knowledge E^*; #1 indicates the final answer selected based on the highest estimated causal effect, i.e., $P(A \mid \textbf{do}(Q))$.

framework, CFD-Prompting, which decomposes the overall effect into three components and enables unbiased answer selection via causal interventions. The overall architecture of CFD-Prompting is illustrated in Figure 3.

3.1 Task Description

We consider knowledge-intensive tasks in which an LLM is prompted with a query Q, generates a CoT C, and subsequently produces a final answer A, as illustrated in Figure 2c. This SCM, inspired by [44], assumes that both Q and C may be influenced by external knowledge E, while U represents a latent confounder that biases the estimation of the causal effect of Q on A.

To address this issue, we adopt a conditional front-door adjustment strategy to obtain an unbiased estimate of $P(A \mid do(Q))$. By recovering the unbiased causal effect, we are able to select the answer with the highest causal effect from the query, which we regard as the most reliable or correct answer.

3.2 Conditional Front-Door Adjustment

To mitigate the bias introduced by U, we leverage conditional front-door adjustment. The formal criterion is defined as follows:

Definition 2 (Conditional Front-Door Criterion [45]). *A set of variables Z_{CFD} is said to satisfy the conditional front-door criterion relative to an ordered pair of variables (Q, A) in a DAG $\mathcal{G}$ such that the following conditions hold: (1) Z_{CFD} intercepts all directed paths from Q to A; (2) there exists a set of variables W, called the conditioning variables of Z_{CFD}, such that all back-door paths from Q to Z_{CFD} are blocked by W; (3) all back-door paths from Z_{CFD} to A are blocked by $Q \cup W$.*

As illustrated in Figure 2c, C satisfies all conditions of the conditional front-door criterion relative to (Q, A), where the external knowledge E serves as the conditioning variable of C. Therefore, C can be used as a valid conditional front-door adjustment variable to identify the causal effect of Q on A.

We now apply Theorem 2 to derive $P(A \mid \text{do}(Q))$. The derivation proceeds as follows:

$$
\begin{aligned}
P(A|do(Q)) &= \sum_c P(c|do(Q))P(A|c, do(Q)) \\
&= \sum_c P(c|do(Q)) \sum_e P(A|\text{do}(Q), c, e)P(e|\text{do}(Q), c) \\
&= \sum_c P(c|do(Q)) \sum_e P(A|\text{do}(Q), \text{do}(c), e)P(e|\text{do}(Q), c), \\
&\quad \text{since } (A \perp\!\!\!\perp C|Q, E) \text{ in } \mathcal{G}_{\overline{Q}\underline{C}} \text{ (Rule 2 in Theorem 2)} \\
&= \sum_c P(c|do(Q)) \sum_e P(A|\text{do}(c), e)P(e|\text{do}(Q), c), \\
&\quad \text{since } (A \perp\!\!\!\perp Q|C, E) \text{ in } \mathcal{G}_{\overline{C\underline{Q}(E)}} \text{ (Rule 3 in Theorem 2)} \\
&= \sum_c P(c|do(Q)) \sum_{e,\,q} P(A|\text{do}(c), q, e)P(q|\text{do}(c), e)P(e|\text{do}(Q), c), \\
&= \sum_c P(c|do(Q)) \sum_{e,\,q} P(A|c, q, e)P(q|\text{do}(c), e)P(e|\text{do}(Q), c), \\
&\quad \text{since } (A \perp\!\!\!\perp C|Q, E) \text{ in } \mathcal{G}_{\underline{C}} \text{ (Rule 2 in Theorem 2)} \\
&= \sum_c P(c|do(Q)) \sum_{e,\,q} P(A|c, q, e)P(q|e)P(e|\text{do}(Q), c), \\
&\quad \text{since } (Q \perp\!\!\!\perp C|E) \text{ in } \mathcal{G}_{\overline{C(E)}} \text{ (Rule 3 in Theorem 2)} \\
&= \sum_c P(c|do(Q)) \sum_{e,\,q} P(A|c, q, e)P(q|e)\frac{P(e, c|do(Q))}{P(c|\text{do}(Q))}, \\
&\quad \text{since the chain rule of conditional probability} \\
&= \sum_c P(c|do(Q)) \sum_{e,\,q} P(A|c, q, e)P(q|e)\frac{P(c|Q, e)p(e)}{P(c|\text{do}(Q))} \\
&= \sum_{c,\,e} P(c|Q, e) \sum_{q,\,e} P(A|c, q, e)P(q|e)P(e)
\end{aligned}
$$

In our setting, Q denotes the query, which remains fixed throughout the reasoning process. That is, we do not perform any intervention over Q, and thus it can be treated as a constant rather than a random variable. Consequently, the term $P(q \mid e)$ and the summation over q can be omitted, simplifying the expression for the causal effect as follows,

$$
P(A \mid do(Q)) = \sum_{c,\,e} \underbrace{P(c \mid Q, e)}_{②} \underbrace{P(A \mid c, q, e)}_{③} \underbrace{P(e)}_{①} \tag{4}
$$

We begin by introducing the procedure for generating counterfactual external knowledge, which is a key step in estimating causal effects via conditional front-door adjustment. We then present Equation 4, which decomposes the causal effect from Q to A into three components, each of which can be estimated independently to recover the overall causal effect. Components ①, ②, and ③ are detailed in Sections 3.3, 3.4, and 3.5, respectively.

3.3 Constructing Counterfactual External Knowledge

In many LLM-based reasoning tasks, the query Q is fixed and not subject to direct causal intervention. This poses a fundamental challenge for causal effect estimation. Traditional causal inference

assumes the ability to intervene on the treatment, i.e., Q in our case. However, in practice, we typically only observe outputs conditioned on a single realisation of Q. As a result of this limitation, directly estimating the causal effect $P(A \mid do(Q))$ becomes infeasible under standard assumptions.

To overcome this limitation, we introduce counterfactual external knowledge as contextual background for reasoning. While Q remains fixed, modifying E allows us to simulate alternative reasoning environments. These changes induce meaningful shifts in the distribution of C, effectively mimicking how Q would behave under different contexts. In this way, we simulate causal intervention without directly altering Q.

To simulate the causal intervention, we construct counterfactual versions of the external knowledge that alter the context in which reasoning occurs. Concretely, given E, we first prompt the LLM to identify the top T entities most relevant to Q, denoted as $V = [v_1, v_2, \ldots, v_T]$. These entities are ranked by their relevance scores and assigned corresponding weights $W = [w_1, w_2, \ldots, w_T]$, where $w_1 \geq w_2 \geq \cdots \geq w_T$. Then, for each $v \in V$, we generate a counterfactual alternative v^*, forming the set $V^* = [v_1^*, v_2^*, \ldots, v_T^*]$. The weights of the counterfactual entities are preserved, i.e., w_t remains unchanged, allowing a controlled substitution in the reasoning context. Here, the subscript $t \in \{1, 2, \ldots, T\}$ indexes the entities based on their ranked relevance to Q.

Next, from the counterfactual entity list V^*, we enumerate all possible subsets of size $T-1$, resulting in $\binom{T}{T-1} = T$ combinations. For each combination, we replace the corresponding entities in the original external knowledge e to construct a counterfactual variant e^*, as follows:

$$E^* = [e_1^*, e_2^*, \ldots, e_T^*], \tag{5}$$

where $t \in \{1, 2, \ldots, T\}$ indexes the combinations, each formed by removing exactly one entity from the original set of T counterfactual entities.

We assign a probability to each e_t^* based on the product of the weights of its constituent entities. Let $W_t = \{w_{t,1}, \ldots, w_{t,T-1}\}$ be the weights of the entities in the t-th combination. Here, each W_t contains the weights of the $T-1$ entities, since each e_t^* is constructed by removing exactly one entity from the original set of T counterfactual entities. Then, the probability of e_t^* is defined as follows:

$$P(e_t^*) = \frac{\prod_{i=1}^{T-1} w_{t,i}}{\sum_{t=1}^{T} (\prod_{i=1}^{T-1} w_{t,i})}. \tag{6}$$

3.4 The calculation of $P(c \mid Q, e)$

We prompt the LLM to generate M CoTs based on the query Q and external knowledge E. These CoTs are then encoded into vector representations using a dedicated encoder. We then apply the K-means clustering algorithm [16] to partition the M CoTs into N clusters. The CoT closest to each cluster centroid is selected, resulting in N representative CoTs for downstream causal interventions.

Since the encoder operates in a representation space different from that of the LLM [49], the computed distances between CoTs may not accurately reflect the LLM's reasoning preferences. To mitigate this issue, we adopt contrastive learning [5, 21] to fine-tune the encoder so that its embedding space is aligned with the LLM representation space.

Specifically, we construct a training dataset $\mathcal{D}$, where each instance consists of Q and its associated CoT c. We treat each CoT $c \in \mathcal{D}$ as an anchor, and denote its representation as z_a. Then, we prompt the LLM to generate a semantically similar CoT c^+, whose representation is used as the positive sample z^+. Meanwhile, CoTs c^- from other instances in the same batch serve as negative samples, with their representations denoted as z^-. Following prior works [5, 49], we adopt the InfoNCE loss to fine-tune the encoder:

$$\mathcal{L} = -\log \frac{\exp(z_a^\top z^+/\tau)}{\exp(z_a^\top z^+/\tau) + \sum_{z^-} \exp(z_a^\top z^-/\tau)}, \tag{7}$$

where τ is a temperature parameter that controls the scaling of the similarities, and $\sum_{z^-}$ denotes the summation over all negative sample embeddings in the batch.

For each counterfactual external knowledge $e_t^* \in E^*$, we input it together with Q into the LLM and repeat this process P times to generate CoTs. These generated CoTs form the set $C_t^* = [c_{t,1}^*, c_{t,1}^*, \ldots, c_{t,p}^*]$. We then encode each CoT in this set using the encoder, resulting in the representations $\tilde{C}_t^* = [\tilde{c}_{t,1}^*, \tilde{c}_{t,2}^*, \ldots, \tilde{c}_{t,p}^*]$. Similarly, the CoT C is encoded into the representations $\tilde{C}$.

Next, we use cosine similarity to compute the similarity between $\tilde{c}_n$, the representation of the n-th CoT selected from the N clusters, and each representation in $\tilde{C}_t^*$, as follows:

$$d_{n,t,p} = \text{cosine}(\tilde{c}_n, \tilde{c}_{t,p}^*), \tag{8}$$

where $d_j \in [-1, 1]$. A value close to 1 indicates high semantic similarity, while a value near 0 suggests dissimilarity.

To assess the impact of the counterfactual external knowledge on the stability of reasoning, we define a similarity threshold s. For each generated CoT $c_{t,p}^*$ from set C_t^*, we calculate the similarity score $d_{n,t,p}$. If $d_{n,t,p} \geq s$, we consider $c_{t,p}$ to be logically consistent with the original CoT c_n, assigning it a score of 1. Conversely, if $d_{n,t,p} < s$, we deem it to differ significantly from c_n, and assign it a score of 0. Formally, we define the indicator function as:

$$\mathbb{I}_{\text{sim}}(c_n, c_{t,p}^*) = \begin{cases} 1 & \text{if } d_{n,t,p} \geq s, \\ 0 & \text{if } d_{n,t,p} < s. \end{cases} \tag{9}$$

We then compute the conditional probability $P(c \mid Q, e)$ by averaging the indicator scores over the P generated CoTs:

$$P(c \mid Q, e) \approx \frac{1}{P} \sum_{p=1}^{P} \mathbb{I}_{\text{sim}}(c_n, c_{t,p}^*). \tag{10}$$

This probability reflects the proportion of generated CoTs that remain semantically consistent with the original CoT under the given E and Q.

3.5 The calculation of $P(A \mid c, q, e)$

Through Section 3.3 and 3.4, we identify CoTs generated using counterfactual external knowledge that are semantically consistent with the original CoTs, forming the subset:

$$C_{t,\text{sub}}^* = [c_{t,1}^*, c_{t,1}^*, \ldots, c_{t,r}^*] \tag{11}$$

where $C_{t,\text{sub}}^* \subseteq C_t^*$ and $r \in \{1, 2, \ldots, R\}$. R is the number of CoTs that satisfy $d_{n,t,p} > s$, indicating semantic consistency with c_n.

For each $c^*_{t,r} \in C^*_{t,\text{sub}}$, we compute the answer using a reasoning function:

$$a_{t,r} = f(\tilde{c}^*_{t,r}), \tag{12}$$

where $f(\cdot)$ denotes the LLM-based reasoning process that produces an answer given a CoT representation.

We compare $a_{t,r}$ against the reference answer a_n (i.e, $a_n = f(\tilde{c}_n)$). If $a_{t,r} = a_n$, it suggests that $c^*_{t,r}$ is insensitive to external knowledge changes, and we assign a score of 0. Otherwise, we assign a score of 1, indicating that $c^*_{t,r}$ reflects adaptive reasoning. We formalise this using an indicator function:

$$\mathbb{I}_{\text{ins}}(c^*_{t,r}) = \begin{cases} 0 & \text{if } a_{t,r} = a_n, \\ 1 & \text{if } a_{t,r} \neq a_n. \end{cases} \tag{13}$$

The probability $P(A \mid c, q, e)$ is defined as:

$$P(A \mid c, q, e) \approx \frac{1}{R} \sum_{j=1}^{R} \mathbb{I}_{\text{ins}}(c^*_{t,r}), \tag{14}$$

This equation captures the sensitivity of the CoT c_i to the external knowledge variations. A low value of $P(A \mid c, q, e)$ indicates consistent answers across contexts, while a high value suggests that the reasoning outcome is highly responsive to external knowledge.

3.6 The calculation of $P(A \mid do(Q))$

Following the previous derivation, Equation 4 can be simplified to the following form:

$$\begin{aligned} P(A \mid do(Q)) &= \sum_{c,\, e} P(c \mid Q, e) P(A \mid c, q, e) P(e) \\ &\approx \sum_{n=1}^{N} \sum_{t=1}^{T} \left[\frac{1}{P} \sum_{j=1}^{P} \mathbb{I}_{\text{sim}}(c_n, c^*_{t,p}) \cdot \frac{1}{R} \sum_{j=1}^{R} \mathbb{I}_{\text{ins}}(c^*_{t,r}) \cdot P(e^*_t) \right] \end{aligned} \tag{15}$$

4 Experiments

In this section, we evaluate the effectiveness and robustness of CFD-Prompting across four knowledge-intensive datasets using three different backbone LLMs. We first introduce the datasets and baseline methods used for comparison, followed by implementation details. We then present the main results, conduct robustness and hyper-parameter studies to assess stability under noisy conditions, and perform an ablation study to examine the contribution of key components in our framework. Due to space constraints, the case study illustrating the practical application of our framework is provided via the anonymous link in the abstract.

4.1 Dataset and Evaluation

To better evaluate the performance of our framework in handling complex knowledge-intensive tasks [27, 52], we follow previous works [44, 49] and select the SciQ, HotpotQA, WikiHop, and MuSiQue datasets for evaluation. These four datasets cover diverse knowledge domains [35, 41, 42, 46], include multi-hop reasoning and feature various question types. They comprehensively assess the performance of methods in knowledge retrieval, reasoning ability, and information processing. The detailed information for each dataset are as follows:

- **SciQ** [41] is a multiple-choice science QA dataset covering physics, chemistry, and biology. We evaluate comparison methods and ours on the test set with provided supporting evidence.
- **HotpotQA** [46] is a multi-hop QA benchmark with open-ended and yes/no questions, requiring reasoning across multiple supporting documents. We use the provided documents as external knowledge in our experiments.
- **WikiHop** [42] is a multi-choice, multi-hop reasoning dataset. We treat its queries as questions and prompt models to generate free-form answers instead of selecting from candidates.
- **MuSiQue** [35] emphasises multi-step reasoning and compositional question decomposition. We select instances requiring more than three reasoning hops for evaluation.

We use Exact Match (EM) and F1 score as the evaluation metrics to assess method performance [46]. Following previous work [22], we extract the text span immediately following the keyword "answer is" as the final predicted answer.

4.2 Comparison Methods and Backbone Models

We compare our framework with the following methods:

- **In-Context Learning (ICL)** [4]: Prompt LLMs with a few demonstration examples consisting of only questions and their corresponding answers, without any intermediate reasoning or explanatory context.
- **CoT without context (CoT w/o ctx)** [40]: Apply CoT prompting without providing any external context, generating reasoning purely based on the query itself.
- **CoT** [40]: Prompt LLMs with demonstration examples containing detailed reasoning chains, guiding the model step-by-step through the thought process required to reach an answer.
- **CoT self-consistency (CoT-SC)** [38]: An extension of CoT prompting where LLMs generate multiple reasoning chains for a given query, and majority voting is used to determine the final answer.
- **Context-aware Decoding (CAD)** [32]: Improve LLM generation by comparing output distributions with and without external context to enhance reasoning reliability.
- **De-biasing CoT (DeCoT)** [44]: Mitigate internal knowledge bias by using external knowledge as an instrumental variable to estimate the average causal effect of the CoT on the answer, enabling the selection of logically correct reasoning paths.
- **Causal Prompting (CP)** [49]: Estimates the causal effect of the query on the answer using the standard front-door adjustment, but it is primarily tailored for general reasoning tasks and does not account for the complexities of knowledge-intensive settings.

In our experiments, we select three pre-trained LLMs as backbone models to ensure diversity and comparability: Llama-2-7b-chat-hf (LLaMA-2) [34], Meta-Llama-3-8B-Instruct (LLaMA-3) [2], and GPT-3.5 Turbo [3]. These LLMs differ in terms of parameter scale, training strategies, and open-source versus closed-source design, providing a comprehensive foundation for evaluation.

Table 1: The comparison results of CFD-Prompting and seven methods across three backbone LLMs on four knowledge-intensive tasks. Best results are highlighted in bold.

Model	Method	SciQ		HotpotQA		WikiHop		MuSiQue		Average	
		EM ↑	F1 ↑	EM ↑	F1 ↑	EM ↑	F1 ↑	EM ↑	F1 ↑	EM ↑	F1 ↑
LLaMA-2	ICL	7.81	9.56	8.40	11.95	11.60	15.47	1.33	2.85	7.29	9.96
	CoT w/o ctx	14.72	22.63	8.20	14.23	4.40	6.26	0.40	3.07	6.93	11.55
	CoT	30.55	45.98	9.50	17.19	16.10	21.22	1.73	4.50	14.47	22.22
	CoT-SC	41.63	54.23	17.20	24.46	21.30	26.78	2.13	4.62	20.56	27.52
	CAD	31.79	40.11	18.00	29.45	16.40	20.03	1.46	6.62	16.91	24.05
	DeCoT	42.26	54.43	20.40	32.16	19.59	25.92	3.32	6.27	21.39	29.70
	CP	42.10	54.23	17.90	25.54	22.01	29.03	2.99	8.05	21.25	29.21
	Ours	**43.60**	**55.35**	**22.00**	**33.73**	**22.93**	**31.06**	**4.52**	**11.68**	**23.26**	**32.95**
LLaMA-3	ICL	24.10	41.45	3.20	20.15	6.80	25.96	3.99	6.65	9.52	23.55
	CoT w/o ctx	35.29	48.48	15.30	25.23	12.60	17.91	2.39	7.70	16.39	24.83
	CoT	50.32	67.52	31.30	47.85	22.40	32.27	12.63	20.15	29.16	41.95
	CoT-SC	60.86	77.57	36.30	54.80	26.60	37.12	21.48	31.20	36.31	50.17
	CAD	52.60	65.20	30.30	40.58	24.00	32.22	12.63	23.12	29.88	40.28
	DeCoT	62.18	79.33	43.30	59.51	24.25	35.17	21.14	28.85	37.72	50.71
	CP	61.59	77.30	42.74	58.87	25.20	35.22	22.13	24.18	37.91	48.89
	Ours	**63.12**	**79.65**	**47.80**	**62.41**	**27.20**	**37.55**	**24.35**	**34.12**	**40.62**	**53.43**
GPT-3.5 Turbo	ICL	65.95	81.01	41.20	52.13	21.30	31.43	23.01	33.20	37.87	49.44
	CoT w/o ctx	42.19	55.58	30.30	42.89	16.30	23.26	8.91	17.65	24.42	34.84
	CoT	66.97	80.32	43.90	60.30	26.72	36.20	26.20	36.40	40.95	53.30
	CoT-SC	68.55	82.37	51.00	66.23	28.18	38.22	33.38	44.27	45.28	57.77
	CAD	67.08	78.84	45.00	60.85	27.70	37.70	26.86	40.14	41.66	54.38
	DeCoT	70.98	84.08	51.30	67.79	31.09	40.14	34.59	47.36	46.99	59.84
	CP	70.93	83.76	51.10	66.53	29.56	39.45	33.78	47.82	46.34	59.39
	Ours	**71.83**	**85.12**	**53.40**	**68.67**	**32.00**	**41.17**	**36.17**	**48.01**	**48.35**	**60.74**

4.3 Implementation Details

We deploy the LLM using the vLLM framework [18]. Compared with traditional Transformer serving frameworks, vLLM achieves higher throughput and faster response speed by leveraging optimised dynamic batching and efficient KV-cache management. In our framework, we first generate $M = 30$ initial CoTs, which are then clustered into $N = 5$ groups. To construct counterfactual external knowledge, we extract $T = 5$ entities from the original context for replacement during counterfactual generation.

4.4 Main Results

Table 1 reports the performance of our proposed framework, CFD-Prompting, across three LLM backbones and seven comparison methods on four knowledge-intensive tasks. As the model size increases from LLaMA-2 to GPT-3.5 Turbo, LLMs demonstrate stronger reasoning capabilities, leading to overall improved performance across all methods. CFD-Prompting consistently outperforms all baselines under each backbone model. Notably, it achieves an average EM/F1 of 23.26/32.95 on LLaMA-2. On LLaMA-3, it yields 40.62/53.43, surpassing CP by (+2.71) EM and (+4.54) F1. On GPT-3.5 Turbo, CFD-Prompting reaches 48.35/60.74, setting new state-of-the-art results with consistent gains across datasets. These results demonstrate the strong generalisability and effectiveness of our framework across both smaller and larger LLMs.

Among the non-causality-based prompting baselines, CoT w/o ctx performs the worst due to its lack of external knowledge, which limits its capacity for multi-hop reasoning. ICL improves performance by providing in-context demonstrations, while CAD enhances answer generation by contrasting model outputs with and without external knowledge. CoT-SC extends standard CoT by aggregating multiple reasoning paths through majority voting, thereby mitigating sampling variance.

DeCoT addresses internal bias by using external knowledge as an instrumental variable to estimate the average causal effect of CoTs on answers. It treats a CoT as logically correct if its ACE is greater than zero. However, this binary threshold offers limited granularity. Our framework instead estimates the causal effect of Q on A, and ranks candidate answers accordingly. While CP applies the standard front-door adjustment, it assumes no observed confounders influence both query and CoT, a condition often violated in knowledge-intensive tasks. In contrast, CFD-Prompting leverages the conditional front-door adjustment, which allows for observed confounders such as external knowledge. This yields more accurate de-biasing and leads to consistent performance improvements across all benchmarks.

Table 2: The results of the robustness study on the SciQ using LLaMA-3. Best results are highlighted in bold.

Method	SciQ		SciQ-Injected		SciQ-Shuffled	
	EM ↑	F1 ↑	EM ↑	F1 ↑	EM ↑	F1 ↑
ICL	24.10	41.45	18.33	34.97	19.46	35.76
CoT w/o ctx	35.29	48.48	35.29	48.48	35.29	48.48
CoT	50.32	67.52	43.78	60.32	46.15	62.14
CoT-SC	60.86	77.57	56.74	74.32	58.59	75.70
CAD	52.60	65.20	51.47	63.28	48.98	60.83
DeCoT	62.18	79.33	61.60	78.84	61.53	78.00
CP	61.59	77.30	58.74	74.60	59.79	76.68
Ours	**63.12**	**79.65**	**62.29**	**79.39**	**62.17**	**78.90**

4.5 Robustness Study

To evaluate the effectiveness and stability of our framework under noisy conditions, we design a robustness experiment based on the SciQ dataset. We introduce two types of perturbations: (1) SciQ-Injected, where we inject irrelevant content (10% of the total) into the support documents to simulate noise interference; and (2) SciQ-Shuffled, where we randomly shuffle half of the support sentences to assess the model's adaptability to contextual disorder. We adopt LLaMA-3 as the backbone model and report the results in Table 2.

As shown in the table, all methods except CoT w/o ctx experience performance degradation under the perturbations. We observe that CoT w/o ctx maintains unchanged results, as it does not utilise external knowledge and is thus unaffected by modifications to the context. Both DeCoT and CFD-Prompting demonstrate notable robustness. This can be attributed to their use of counterfactual external knowledge construction, which effectively mitigates the impact of noisy or disordered information. Compared to DeCoT, which selects entities based on frequency heuristics [44], CFD-Prompting focuses on query-relevant entity selection combined with conditional front-door adjustment, resulting in more reliable performance. Specifically, CFD-Prompting achieves the highest F1 scores of 79.39 on SciQ-Injected and 78.90 on SciQ-Shuffled. Even under perturbations, the performance drop remains minimal, highlighting the robustness and stability of our framework.

4.6 Hyper-parameter Study

We further study the influence of the number of initially generated CoTs (M) and the number of clustering categories (N) on the performance of our framework. Table 3 summarises the experimental results, where the upper part reports the effect of varying M while fixing $N = 5$, and the lower part shows the effect of varying N while fixing $M = 30$. The results indicate that increasing M generally improves model performance, while a larger N enables finer-grained clustering, further enhancing the robustness of causal effect estimation. However, larger M and N values incur higher token consumption and inference costs. To balance efficiency and performance, we set $M = 30$ and $N = 5$ as the default configuration throughout all experiments.

Table 3: The performance of CFD-Prompting with different numbers of CoTs (M) and clusters (N) across four knowledge-intensive tasks.

	SciQ		HotpotQA		WikiHop		MuSiQue	
M	EM ↑	F1 ↑	EM ↑	F1 ↑	EM ↑	F1 ↑	EM ↑	F1 ↑
10	60.75	74.77	45.75	61.20	25.63	35.29	22.77	31.98
20	62.55	76.36	46.83	62.38	27.10	36.85	23.50	32.76
30	63.12	79.65	47.80	62.41	27.20	37.55	24.35	34.12
40	62.13	76.45	47.17	62.61	27.25	36.72	24.27	33.65
50	63.57	79.79	48.20	63.71	27.50	37.53	25.17	34.21
N	EM ↑	F1 ↑	EM ↑	F1 ↑	EM ↑	F1 ↑	EM ↑	F1 ↑
1	60.63	74.64	46.19	60.46	26.30	36.28	22.85	32.10
3	63.03	76.89	47.40	62.56	26.66	36.54	23.27	32.62
5	63.12	79.65	47.80	62.41	27.20	37.55	24.35	34.12
7	62.50	76.33	47.51	62.80	27.39	37.74	24.23	33.95
9	63.34	80.08	48.06	62.88	27.73	37.63	24.53	34.78

4.7 Ablation Study

We conduct an ablation study to assess the contribution of three key components in CFD-Prompting: (1) relevance-based entity weighting during counterfactual construction, (2) encoder fine-tuning via contrastive learning, and (3) CoT clustering using K-means. All experiments are run on the LLaMA-3 model across four knowledge-intensive datasets. For (1), we compare our default relevance-based selection with two variants: (a) random selection and (b) reversed-weight selection, where the least relevant entities are chosen. For (2), we ablate contrastive learning by removing encoder fine-tuning. For (3), we disable clustering by using the full set of CoTs without K-means. These settings allow us to isolate and quantify the impact of each design choice.

As shown in Table 4, removing any of the three components in CFD-Prompting leads to consistent performance degradation across all four datasets, validating their importance. The most substantial drop occurs when the relevance-based weighting strategy is ablated. Specifically, replacing it with random selection reduces average F1 from 53.43 to 51.42 (−2.01), while reversed weighting further lowers it to 49.79 (−3.64). Contrastive learning also proves essential. Removing encoder fine-tuning reduces the average F1 to 52.13, with notable drops on MuSiQue (−2.38) and HotpotQA (−1.50), suggesting that enhancing the encoder's ability to distinguish different CoT representations improves the accuracy of causal effect estimation. Finally, removing K-means clustering leads to moderate but consistent drops, indicating that promoting CoT diversity through clustering contributes to more stable and robust causal effect estimation.

5 Related Work

5.1 LLMs for Knowledge-Intensive Tasks

Knowledge-intensive tasks [27, 42] require large-scale models to effectively leverage external information during inference. A common solution is RAG [14, 29], which retrieves relevant knowledge from external corpora based on the input and supplements the

Table 4: The results of ablation study on four knowledge-intensive tasks using LLaMA-3. Best results are highlighted in bold.

	SciQ		HotpotQA		WikiHop		MuSiQue		Average	
Method	EM ↑	F1 ↑	EM ↑	F1 ↑	EM ↑	F1 ↑	EM ↑	F1 ↑	EM ↑	F1 ↑
CFD-Promoting	**63.12**	**79.65**	**47.80**	**62.41**	**27.20**	**37.55**	**24.35**	**34.12**	**40.62**	**53.43**
w Random weighting	60.57	77.87	44.80	60.06	26.30	36.28	22.50	31.45	38.54	51.42
w Reversed weighting	60.18	76.43	40.85	56.11	25.63	35.82	21.93	30.80	37.15	49.79
w/o Contrastive learning	62.24	79.33	45.98	60.91	26.49	36.54	23.27	31.74	39.50	52.13
w/o K-means clustering	62.47	79.25	46.12	60.92	26.51	36.84	23.64	32.07	39.69	52.27

model's internal reasoning process. Building on this idea, ICL [17] provides input-output exemplars directly within the prompt, allowing the model to generalise task-specific behaviours from observed patterns. To further improve reasoning quality, CoT prompting [40] encourages models to decompose complex problems into intermediate steps, enabling more systematic utilisation of retrieved or contextual information. Beyond prompting strategies, structured external resources such as knowledge graphs [15, 43] and optimisation techniques like Reinforcement Learning from Human Feedback (RLHF) [12] have also been developed to guide model inference and enhance the integration of external knowledge.

However, despite these advancements, large-scale models still exhibit internal bias that can distort their use of external information, often leading to flawed reasoning and incorrect answers.

5.2 Chain-of-Thought

In recent years, CoT prompting has been widely adopted to enhance the reasoning capabilities of LLMs. This technique has shown significant performance improvements in domains such as mathematical problem solving and symbolic reasoning [40]. However, traditional CoT methods are prone to reasoning bias, where errors introduced in the early steps of a reasoning chain can propagate through subsequent steps, ultimately leading to biased final answers [36]. This issue is particularly pronounced in knowledge-intensive tasks [28], where longer reasoning chains and greater reliance on factual information increase the risk of compounding errors.

To address this, CoT-SC [38] generates multiple distinct reasoning paths and aggregates their outcomes to reduce the influence of individual erroneous paths. While effective in mitigating random inference errors, CoT-SC still suffers from internal bias in LLMs, particularly in tasks involving complex knowledge structures [22].

5.3 De-biasing LLMs via Causal Inference

Causal inference aims to quantify the effect of a treatment on an outcome [25, 26]. Supported by a rich theoretical foundation, various methods have been developed to estimate causal effects even in the presence of unobserved confounders [6–11]. Building on these foundations, an increasing number of studies apply causal inference to understand and mitigate biases in LLMs.

For example, Li et al. [20] propose a causality-guided prompting framework to control the influence of social information on model predictions. Other works explore causal modelling for de-biasing tasks: Wang et al. [39] design a causal graph for relation extraction and analyse counterfactual by removing textual context; Zhou et al. [51] introduce Causal-Debias, which mitigates stereotypical associations via causal disentanglement; and Wang et al. [37] develop a structural causal model to address entity bias through tractable interventions across entities, text, and outputs.

While these studies lay important groundwork, their scope is largely limited to fairness-oriented tasks or fine-tuning strategies. In contrast, recent causality-based prompting methods, such as DeCoT [44], CP [49], and CAPITAL [30], aim to improve reasoning quality. DeCoT treats external knowledge as an instrumental variable to estimate the average causal effect of CoTs on answers, but offers only a coarse assessment. CP uses standard front-door adjustment, which relies on the strong assumption that no observed confounders exist between the prompt and the CoT, a condition often violated in knowledge-intensive tasks.

In contrast, the proposed CFD-Prompting estimates the causal effect of the query on the answer via conditional front-door adjustment, allowing for finer-grained de-biasing and delivering consistent state-of-the-art performance across benchmarks.

6 Conclusion

In this paper, we propose CFD-Prompting, a novel framework that leverages conditional front-door adjustment to mitigate internal bias in LLMs. CFD-Prompting constructs counterfactual external knowledge and aligns reasoning representations via contrastive learning, enabling more accurate estimation of the causal effect between the query and the answer. Unlike existing methods, CFD-Prompting relaxes restrictive assumptions and does not require access to model logits, making it applicable to both open- and closed-source LLMs. Extensive experiments on multiple knowledge-intensive benchmarks demonstrate that CFD-Prompting consistently improves reasoning accuracy and robustness, highlighting its effectiveness and generalisability in real-world applications.

Acknowledgments

This research was supported by the National Key Research and Development Program of China (Grant 2023YFF1000100), the Hubei Key Research and Development Program of China (Grants 2024BBB-055, 2024BAA008), the Major Science and Technology Project of Yunnan Province (Grant 202502AE090003), the Fundamental Research Funds for the Chinese Central Universities (Grant 2662025XX-PY005), and the research support package from the School of Computing Technologies at RMIT University.

GenAI Usage Disclosure

We used ChatGPT (OpenAI) solely for language polishing purposes during the preparation of this manuscript. No part of the research design, data analysis, code implementation, or scientific content generation involved the use of generative AI tools. All technical ideas, experimental results, and interpretations were produced independently by the authors.

References

[1] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), ACL/IJCNLP 2021*. 7319–7328. https://doi.org/10.18653/v1/2021.acl-long.568

[2] Guangsheng Bao, Hongbo Zhang, Linyi Yang, Cunxiang Wang, and Yue Zhang. 2024. LLMs with Chain-of-Thought Are Non-Causal Reasoners. *CoRR* abs/2402.16048 (2024). https://doi.org/10.48550/arXiv.2402.16048

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33, NeurIPS 2020*. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33, NeurIPS 2020*. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, Vol. 119. 1597–1607. http://proceedings.mlr.press/v119/chen20j.html

[6] Debo Cheng, Jiuyong Li, Lin Liu, Ziqi Xu, Weijia Zhang, Jixue Liu, and Thuc Duy Le. 2024. Disentangled Representation Learning for Causal Inference With Instruments. *IEEE Transactions on Neural Networks and Learning Systems* (2024), 1–14. doi:10.1109/TNNLS.2024.3512790

[7] Debo Cheng, Yang Xie, Ziqi Xu, Jiuyong Li, Lin Liu, Jixue Liu, Yinghao Zhang, and Zaiwen Feng. 2023. Disentangled Latent Representation Learning for Tackling the Confounding M-Bias Problem in Causal Inference. In *IEEE International Conference on Data Mining, ICDM 2023*. 51–60. https://doi.org/10.1109/ICDM58522.2023.00014

[8] Debo Cheng, Ziqi Xu, Jiuyong Li, Lin Liu, Thuc Duy Le, and Jixue Liu. 2023. Learning Conditional Instrumental Variable Representation for Causal Effect Estimation. In *Machine Learning and Knowledge Discovery in Databases: Research Track - European Conference, ECML PKDD 2023*, Vol. 14169. 525–540. https://doi.org/10.1007/978-3-031-43412-9_31

[9] Debo Cheng, Ziqi Xu, Jiuyong Li, Lin Liu, Jixue Liu, Wentao Gao, and Thuc Duy Le. 2024. Instrumental Variable Estimation for Causal Inference in Longitudinal Data with Time-Dependent Latent Confounders. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024*. 11480–11488. https://doi.org/10.1609/aaai.v38i10.29029

[10] Debo Cheng, Ziqi Xu, Jiuyong Li, Lin Liu, Jixue Liu, and Thuc Duy Le. 2023. Causal Inference with Conditional Instruments Using Deep Generative Models. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023*. 7122–7130. https://doi.org/10.1609/aaai.v37i6.25869

[11] Debo Cheng, Ziqi Xu, Jiuyong Li, Lin Liu, Jixue Liu, and Thuc Duy Le. 2024. Conditional Instrumental Variable Regression with Representation Learning for Causal Inference. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. https://openreview.net/forum?id=qDhq1icpO8

[12] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. Safe RLHF: Safe Reinforcement Learning from Human Feedback. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. https://openreview.net/forum?id=TyFrPOKYXw

[13] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence* 5, 3 (2023), 220–235. https://doi.org/10.1038/s42256-023-00626-4

[14] Yasuto Hoshi, Daisuke Miyashita, Youyang Ng, Kento Tatsuno, Yasuhiro Morioka, Osamu Torii, and Jun Deguchi. 2023. RaLLe: A Framework for Developing and Evaluating Retrieval-Augmented Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*. 52–69. https://doi.org/10.18653/v1/2023.emnlp-demo.4

[15] Nourhan Ibrahim, Samar AboulEla, Ahmed F. Ibrahim, and Rasha F. Kashef. 2024. A survey on augmenting knowledge graphs (KGs) with large language models (LLMs): models, evaluation metrics, benchmarks, and challenges. *Discover Artificial Intelligence* 4, 1 (2024), 76. https://doi.org/10.1007/s44163-024-00175-8

[16] Abiodun M. Ikotun, Absalom E. Ezugwu, Laith Abualigah, Belal Abuhaija, and Heming Jia. 2023. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information science* 622 (2023), 178–210. https://doi.org/10.1016/j.ins.2022.11.139

[17] Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP. *CoRR* abs/2212.14024 (2022). https://doi.org/10.48550/arXiv.2212.14024

[18] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023*. 611–626. https://doi.org/10.1145/3600006.3613165

[19] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33, NeurIPS 2020*. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[20] Jingling Li, Zeyu Tang, Xiaoyu Liu, Peter Spirtes, Kun Zhang, Liu Leqi, and Yang Liu. 2025. Prompting Fairness: Integrating Causality to Debias Large Language Models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*. https://openreview.net/forum?id=7GKbQ1WT1C

[21] Xiaozhuan Liang, Ningyu Zhang, Siyuan Cheng, Zhenru Zhang, Chuanqi Tan, and Huajun Chen. 2022. Contrastive Demonstration Tuning for Pre-trained Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 799–811. https://doi.org/10.18653/v1/2022.findings-emnlp.56

[22] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful Chain-of-Thought Reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023*. 305–329. https://doi.org/10.18653/v1/2023.ijcnlp-main.20

[23] Chenglong Ma, Ziqi Xu, Yongli Ren, Danula Hettiachchi, and Jeffrey Chan. 2025. PUB: An LLM-Enhanced Personality-Driven User Behaviour Simulator for Recommender System Evaluation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025*. 2690–2694. doi:10.1145/3726302.3730238

[24] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*. 9802–9822. https://doi.org/10.18653/v1/2023.acl-long.546

[25] Judea Pearl. 2009. *Causality*. Cambridge university press.

[26] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

[27] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick S. H. Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2020. KILT: a Benchmark for Knowledge Intensive Language Tasks. *CoRR* abs/2009.02252 (2020). https://arxiv.org/abs/2009.02252

[28] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and Narrowing the Compositionality Gap in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 5687–5711. https://doi.org/10.18653/v1/2023.findings-emnlp.378

[29] Hongjing Qian, Yutao Zhu, Zhicheng Dou, Haoqi Gu, Xinyu Zhang, Zheng Liu, Ruofei Lai, Zhao Cao, Jian-Yun Nie, and Ji-Rong Wen. 2023. WebBrain: Learning to Generate Factually Correct Articles for Queries by Grounding on Large Web Corpus. *CoRR* abs/2304.04358 (2023). https://doi.org/10.48550/arXiv.2304.04358

[30] Jing Ren, Wenhao Zhou, Bowen Li, Mujie Liu, Nguyen Linh Dan Le, Jiade Cen, Liping Chen, Ziqi Xu, Xiwei Xu, and Xiaodong Li. 2025. Causal Prompting for Implicit Sentiment Analysis with Large Language Models. *CoRR* abs/2507.00389 (2025). https://doi.org/10.48550/arXiv.2507.00389

[31] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large Language Models Can Be Easily Distracted by Irrelevant Context. In *International Conference on Machine Learning, ICML 2023*, Vol. 202. 31210–31227. https://proceedings.mlr.press/v202/shi23a.html

[32] Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting Your Evidence: Hallucinate Less with Context-aware Decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Short Papers, NAACL 2024*. 783–791. https://doi.org/10.18653/v1/2024.naacl-short.69

[33] Xingyu Tan, Xiaoyang Wang, Qing Liu, Xiwei Xu, Xin Yuan, and Wenjie Zhang. 2024. Paths-over-Graph: Knowledge Graph Empowered Large Language Model Reasoning. *CoRR* abs/2410.14211 (2024). https://doi.org/10.48550/arXiv.2410.14211

[34] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR* abs/2307.09288 (2023). https://doi.org/10.48550/arXiv.2307.09288

[35] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics* 10 (2022), 539–554. https://doi.org/10.1162/tacl_a_00475

[36] Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change). In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

[37] Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou, and Muhao Chen. 2023. A Causal View of Entity Bias in (Large) Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 15173–15184. https://doi.org/10.18653/v1/2023.findings-emnlp.1013

[38] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. https://openreview.net/forum?id=1PL1NIMMrw

[39] Yiwei Wang, Muhao Chen, Wenxuan Zhou, Yujun Cai, Yuxuan Liang, Dayiheng Liu, Baosong Yang, Juncheng Liu, and Bryan Hooi. 2022. Should We Rely on Entity Mentions for Relation Extraction? Debiasing Relation Extraction with Counterfactual Analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022*. 3071–3081. https://doi.org/10.18653/v1/2022.naacl-main.224

[40] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 35, NeurIPS 2022*. http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

[41] Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing Multiple Choice Science Questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text, NUT@EMNLP 2017*. 94–106. https://doi.org/10.18653/v1/w17-4413

[42] Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing Datasets for Multi-hop Reading Comprehension Across Documents. *Transactions of the Association for Computational Linguistics* 6 (2018), 287–302. https://doi.org/10.1162/tacl_a_00021

[43] Yilin Wen, Zifeng Wang, and Jimeng Sun. 2024. MindMap: Knowledge Graph Prompting Sparks Graph of Thoughts in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*. 10370–10388. https://doi.org/10.18653/v1/2024.acl-long.558

[44] Junda Wu, Tong Yu, Xiang Chen, Haoliang Wang, Ryan A. Rossi, Sungchul Kim, Anup B. Rao, and Julian J. McAuley. 2024. DeCoT: Debiasing Chain-of-Thought for Knowledge-Intensive Tasks in Large Language Models via Causal Intervention. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*. 14073–14087. https://doi.org/10.18653/v1/2024.acl-long.758

[45] Ziqi Xu, Debo Cheng, Jiuyong Li, Jixue Liu, Lin Liu, and Kui Yu. 2024. Causal Inference with Conditional Front-Door Adjustment and Identifiable Variational Autoencoder. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. https://openreview.net/forum?id=wFf9m4v7oC

[46] 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. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018*. 2369–2380. https://doi.org/10.18653/v1/d18-1259

[47] Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023. Revisiting Out-of-distribution Robustness in NLP: Benchmarks, Analysis, and LLMs Evaluations. In *Advances in Neural Information Processing Systems 36, NeurIPS 2023*. http://papers.nips.cc/paper_files/paper/2023/hash/b6b5f50a2001ad1cbccca96e693c4ab4-Abstract-Datasets_and_Benchmarks.html

[48] Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2023. Investigating the Catastrophic Forgetting in Multimodal Large Language Models. *CoRR* abs/2309.10313 (2023). https://doi.org/10.48550/arXiv.2309.10313

[49] Congzhi Zhang, Linhai Zhang, Jialong Wu, Yulan He, and Deyu Zhou. 2025. Causal Prompting: Debiasing Large Language Model Prompting Based on Front-Door Adjustment. In *Proceedings of the AAAI Conference on Artificial Intelligence, AAAI 2025*, Vol. 39. 25842–25850. doi:10.1609/aaai.v39i24.34777

[50] Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-Edit: A Knowledge-Enhanced Chain-of-Thought Framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*. 5823–5840. https://doi.org/10.18653/v1/2023.acl-long.320

[51] Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 2023. Causal-Debias: Unifying Debiasing in Pretrained Language Models and Fine-tuning via Causal Invariant Learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*. 4227–4241. https://doi.org/10.18653/v1/2023.acl-long.232

[52] Yin Zhu, Zhiling Luo, and Gong Cheng. 2023. Furthest Reasoning with Plan Assessment: Stable Reasoning Path with Retrieval-Augmented Large Language Models. *CoRR* abs/2309.12767 (2023). https://doi.org/10.48550/arXiv.2309.12767