Debiasing LLMs in Knowledge-Intensive Tasks via Information-Gain Guided Front-Door Adjustment

Xin Miao, Yongqi Li, Hankun Kang, Mayi Xu, Jintao Wen, Yuanyuan Zhu, Ming Zhong, Jiawei Jiang, and Tieyun Qian(✉)

School of Computer Science, Wuhan University, Wuhan, China
{miaoxin,liyongqi,kanghankun,xumayi,23jtwen,yyzhu,
clock,jiawei.jiang,qty}@whu.edu.cn

Abstract. Large language models (LLMs) have achieved impressive performance on knowledge-intensive tasks; however, their predictions can be biased due to spurious correlations acquired during pretraining. Recently, debiasing approaches based on front-door adjustment have made substantial progress. Nevertheless, when chain-of-thought (CoT) is used as the mediator, these methods may still be influenced by latent biases, as the CoT generated by LLMs can also be biased. To address this issue, we propose a new paradigm that constructs mediators from sentences with high information gain. Specifically, we first provide a theoretical analysis of mediator properties and prove that valid mediators are variables that yield information gain for the task. Building on this insight, we introduce the Information-Gain Front-Door adjustment (IGFD) framework, which extracts key sentences as mediators from a document via information-gain computation, thereby mitigating biases introduced when LLMs generate mediators. Experiments on document-level multi-hop question answering show that IGFD consistently improves LLM performance and outperforms recent causal-based debiasing methods across both open- and closed-source models. The code and data are publicly available at: https://github.com/xin-miao-cs/IGFD.

Keywords: LLM Debiasing · Causal Inference · Front-Door Adjustment · Information Gain · Knowledge-Intensive Tasks · Prompt Engineering

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in language understanding and generation by pretraining on massive corpora. With Chain-of-Thought (CoT) prompting [28], they further achieve strong reasoning performance [7]. Nevertheless, LLMs still struggle with knowledge-intensive tasks [39], where external evidence must be incorporated for reasoning. Simply

H. Jung et al. (Eds.): DASFAA 2026, LNCS 16537, pp. 374–389, 2026.
https://doi.org/10.1007/978-981-92-0369-7_24

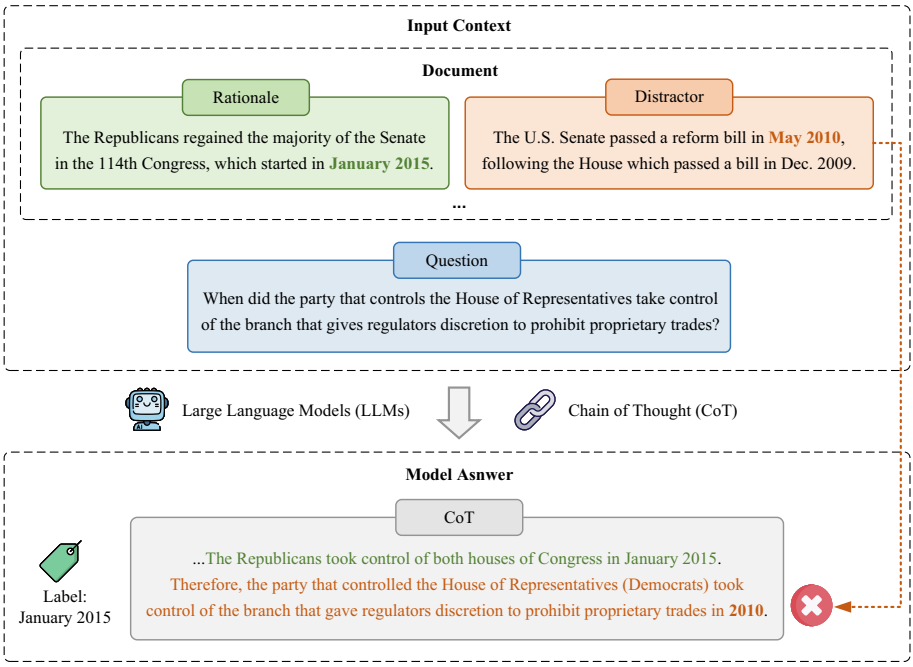

Fig. 1. An example of internal bias in LLMs. Distracting content may amplify the models internal biases, leading to an incorrect reasoning trajectory and answer. The example is taken from our experiments with GPT-4o-mini [8] on MuSiQue [25].

appending external knowledge to the prompt does not ensure that LLMs can reliably identify and exploit the relevant information [22]. A key reason is that, as statistical learners [4], LLMs inevitably pick up spurious correlations [17] from training data, which can manifest as biases such as position bias and stereotype bias [20,40]. In knowledge-intensive settings, these internal biases are often amplified by the presence of distracting evidence. As illustrated in Fig. 1, irrelevant content can steer the model toward an incorrect reasoning trajectory, ultimately producing a wrong answer. This phenomenon poses a serious challenge to the reliable deployment of LLMs in knowledge-intensive scenarios [2,33].

From a causal perspective [17,19], such bias can be viewed as spurious correlationsstatistical regularities that do not reflect the underlying semantic or causal relationships. Motivated by this view, causality-inspired debiasing methods have made notable progress in recent years. Since frequent fine-tuning of LLMs is often expensive and inefficient [36], most existing work focuses on debiasing at inference time. Early studies [24] identify biased samples via causal-invariant interventions, summarize common bias patterns, and then prompt models to ignore them. However, exhaustively identifying the internal biases of LLMs remains infeasible, as LLMs are largely black-box models [3]

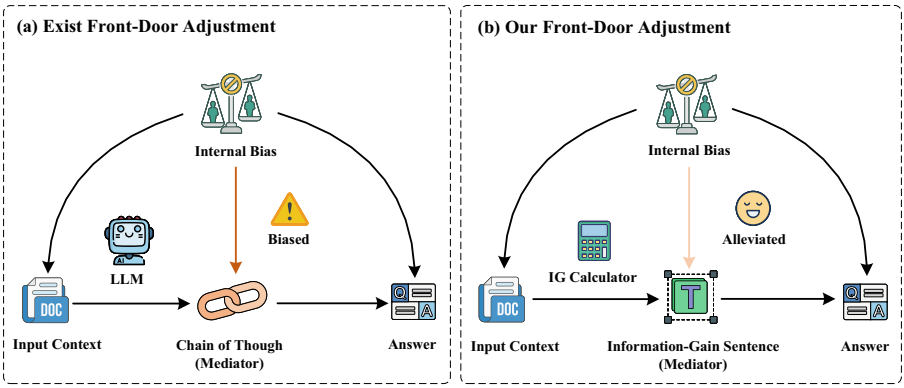

Fig. 2. Comparison of front-door adjustment paradigms. (a) Prior methods use CoT as the mediator; however, CoT is generated by the LLM and thus can inherit internal biases. (b) IGFD instead extracts sentences with high information gain with respect to the question as mediators, reducing the bias contamination in LLM generation.

To tackle this issue, recent studies have adopted front-door adjustment [16, 18] as a debiasing strategy. Front-door adjustment estimates the causal effect from inputs to outputs via an intermediate variable (i.e., a mediator), thereby avoiding the need to explicitly enumerate or identify specific biases, as illustrated in Fig. 2 (a). A mediator is a component of the input that causally influences the output. For example, "nicotine" mediates the causal relationship between "cigarettes" and "lung cancer" [19]. Consequently, identifying valid mediators is crucial for effective debiasing. Most existing methods treat chain-of-thought (CoT) as the mediator: they first sample diverse CoTs as candidates and then either select the most stable one via counterfactual interventions [31,39], or aggregate predictions through voting [37]. While effective, these approaches still rely on CoTs generated by LLMs and are therefore susceptible to the models internal biases. In other words, CoT mediators may themselves be biased, which can limit further improvements under this paradigm (Fig. 2(a)).

To mitigate this limitation, we propose the Information-Gain guided Front-Door adjustment (IGFD) framework. IGFD extracts high-information-gain sentences as mediators, thereby reducing the influence of LLMs' internal biases in generation. Figure 2 (b) illustrates the IGFD paradigm in comparison with prior approaches. The key difference is the level at which the mediator is constructed. Prior methods operate at the model-generation level by treating CoT as the mediator. In contrast, IGFD operates at the task level and selects sentences with high information gain (IG) with respect to the question as mediators. This design offers two advantages: (1) mediators are externally selected rather than generated by the model, making them less susceptible to internal biases; and (2) high-IG mediators highlight essential evidence (rationales) while suppressing irrelevant content (distractors), thereby improving interpretability.

Specifically, we formally prove the soundness of our information-gain guided paradigm. We then introduce a semantic decomposition strategy to accurately estimate the information gain of each sentence by decomposing both the sentence and the question into atomic semantic units, enabling fine-grained computation. Based on these estimates, IGFD extracts high-information-gain sentences as mediators and performs front-door adjustment to debias LLMs. We evaluate IGFD on multiple knowledge-intensive tasks with both open-source and closed-source LLMs, and the results show that it achieves state-of-the-art (SOTA) performance. Our contributions are summarized as follows:

1. (1) We propose an information-gain guided front-door adjustment paradigm for LLM debiasing and provide a proof that establishes its effectiveness.
2. (2) Building on this paradigm, we develop the IGFD framework, which automatically identifies mediators by computing sentence-level information gain and generates corresponding prompts to simulate front-door adjustment.
3. (3) We evaluate IGFD on multiple knowledge-intensive tasks with both open-source and closed-source LLMs. Compared with a range of recent causality-based debiasing baselines, IGFD achieves SOTA performance.

2 Related Work

2.1 LLMs for Knowledge-Intensive Tasks

Knowledge-intensive tasks require large-scale models to effectively leverage external information during inference [39]. As the input window of LLMs continues to expand, the application scenarios of knowledge-intensive tasks are also gradually broadening. One of the most typical scenarios is retrieval-augmented generation (RAG) [1,6,9], which retrieves large amounts of external knowledge to supplement the prompt and assist LLMs in reasoning. In addition, there are document-level semantic generation and understanding tasks across various domains, such as document-level machine translation [12,14,26,32] and relation extraction [11,23,41]. Despite significant progress, internal biases in LLMs can still make them error-prone when processing long contexts [2,33].

2.2 Debiasing LLMs via Causal Inference

Causal inference aims to estimate causal effects between variables through statistical measures [16–19]. Its solid theoretical foundation and practical value have led to growing interest in applying causality for LLM debiasing

For example, an early study [24] identifies the biased instances by detecting cases where LLMs fail to capture causally invariant semantic relationships in the context. It then uses LLMs to summarize the underlying bias types and incorporates these summaries into the prompt to regularize model behavior. Since hidden biases are difficult to enumerate and detect, front-door-adjustment-based methods have attracted increasing attention. Wu et al. [31] treat CoT as the mediator and select the most semantically consistent CoT via sampling and

counterfactual interventions. The recent work [39] follows a similar CoT-selection strategy, with the main difference being that the former adopts counterfactual intervention on the given document, while this method intervenes on CoTs. In addition, Zhang et al. [37] sample diverse CoTs and estimate the causal effect through a voting mechanism to achieve debiasing. Despite their effectiveness, using CoT as the mediator remains vulnerable to LLM internal biases in generation. In contrast, we effectively mitigate this issue through information gain.

3 Preliminaries

3.1 Structural Causal Model

The structural causal model (SCM) [17] can be defined as a directed acyclic graph (DAG) $\langle G := V, E \rangle$. Nodes V represent variables, and directed edges E represent the causal relationship. Parent nodes have a causal effect on child nodes. For instance, in Fig. 3 (a), the edge $X \rightarrow Y$ indicates that X triggers the LLM to generate Y. In this paper, X denotes the input context, including the given document and question. Y denotes the model answer. M serves as the mediator, which is CoT in existing methods, while in our approach, it refers to sentences with high information gain. Finally, U denotes internal biases in LLMs (unobserved), i.e., confounders, such as co-occurrence or length bias [10,13].

Bias acts as a confounder that simultaneously influences two variables, creating a spurious correlation between them, as shown by the dashed line in Fig. 3 (a). For example, if U represents length bias, it simultaneously affects the models encoding of the input X and its generated answer Y according to its length. Spurious correlations can confound the causal effects between variables. In front-door adjustment, when CoT is defined as the mediator M, it is also influenced by internal biases U within LLMs. Consequently, M has a spurious correlation with the models answer Y, as shown in Fig. 3 (b). However, previous studies overlook this issue and ignore the influence of U on M [31,37,39].

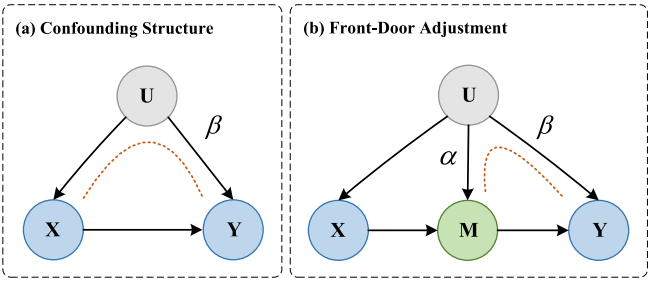

Fig. 3. The basic structural causal model: U denotes internal biases in LLMs (confounders), X denotes the input context, Y denotes the models answer, and M denotes the mediator. In addition, the arrows represent causal effects, while the dashed lines indicate spurious correlations. α, β and $\tilde{\beta}$ represent the degrees of bias influence.

3.2 Front-Door Adjustment

Front-door adjustment is a causal estimation method [18,19]. Unlike back-door adjustment, it does not require controlling for confounders, making it suitable for situations where confounders, i.e., biases, are unobservable. Taking Fig. 3 (b) as an example, to estimate the causal effect of X on Y, it can be decomposed into two sequential steps using the do-operator [16,17] as:

$$P(Y \mid do(X)) = \sum_{m} P(m \mid do(X))P(Y \mid do(m)), \tag{1}$$

where $P(Y \mid do(X))$ denotes the causal effect of X on Y. The first item estimates the effect from X to M, while the second captures the effect from M to Y. These estimates are then combined to derive the overall causal effect through the mediator. Prior to this, each item must be estimated separately. For $P(m \mid do(X))$, under ideal mediators, there are no confounders between X and M, i.e., $\alpha = 0$. Thereby, the statistical value equals the causal effect, i.e., $P(m \mid do(X)) \equiv P(m \mid X)$. Meanwhile, for $P(Y \mid do(m))$, X can be treated as an observable confounder between M and Y. Therefore, we can estimate the causal effect between them via conducting backdoor adjustment [18,19] on X:

$$P(Y \mid do(m)) = \sum_{x} P(Y \mid x, m)P(x). \tag{2}$$

Finally, the above two expressions are integrated to yield the complete front-door adjustment formula [19] as follows:

$$P(Y \mid do(X)) = \sum_{m} P(m \mid X) \sum_{x'} P(Y \mid x', m)P(x'). \tag{3}$$

To estimate the causal effect between X and Y, we need to identify the mediator M and then simulate Eq. 3 with LLMs to obtain debiasing answers.

3.3 Information Gain

Information gain is a metric derived from information theory [21] that measures the reduction in uncertainty about a variable after observing another variable. It quantifies how much knowing one variable contributes to predicting another. Formally, the IG of mediator M with respect to answer Y is defined as the difference between the entropy of Y before and after observing M:

$$IG(Y, M) = H(Y) - H(Y \mid M), \tag{4}$$

where $H(Y)$ is the entropy of Y, denoting its inherent uncertainty, and $H(Y \mid M)$ is the conditional entropy of Y given M, meaning the remaining uncertainty after M is given. A higher IG indicates that M provides more information for Y.

4 Methodology

4.1 Task Definition

This paper considers document-level question answering (QA) as the target knowledge-intensive task. We represent the dataset as $\{(x = (d, q), y)\}$, where x denotes the input context consisting of a given document d and the corresponding question q, and y is the associated answer. The object of the task is to predict y given x. Furthermore, the debiasing objective is to identify sentences $\mathbf{s}_i$ from the document d ($\mathbf{s}_i \in d$) that carry high information gain to the question q, and employ them as the mediator m to perform front-door adjustment as Equation 3. The entire process consists of first extracting $\mathbf{s}_i$ based on information gain, and then predicting y given $x' = (\mathbf{s}_i, q)$ based on front-door adjustment.

4.2 Formal Proof

This section presents a formal proof of the rationality of our information-gain guided paradigm. Using Fig. 3 as an illustrative example, we first assume that the mediator M enables effective debiasing, meaning that the amount of information required to answer the question is reduced. Therefore, the information required to obtain the answer Y in figure (a) is greater than that in figure (b). For clarity, we utilize the self-information [21] to quantify the information of each variable:

$$I(v) = -logP(v), \tag{5}$$

where $P(v)$ denotes the occurrence probability of the variable v; the smaller the probability, the more information is needed. Building upon this, we obtain the required information to answer the question in figures (a) and (b), forming the following inequality, which can be further expanded and derived as follows:

$$\begin{aligned}
&I(Y \mid X) + I(\beta) > I(M \mid X) + I(\alpha) + I(Y \mid M) + I(\tilde{\beta}) && (6)\\
&\Rightarrow -\log P(Y \mid X) - \log[1 - P(\beta)] > -\log P(M \mid X) \\
&\quad - \log[1 - P(\alpha)] - \log P(Y \mid M) - \log[1 - P(\tilde{\beta})], && (7)\\
&\Rightarrow \log P(Y \mid X) + \log[1 - P(\beta)] < \log P(M \mid X) \\
&\quad + \log[1 - P(\alpha)] + \log P(Y \mid M) + \log[1 - P(\tilde{\beta})], && (8)\\
&\Rightarrow \log P(Y \mid X) + \log[1 - P(\beta)] < \log P(M \mid X) \\
&\quad + \log[1 - P(\alpha)] + \log P(Y \mid M), && (9)\\
&\Rightarrow P(Y \mid X)[1 - P(\beta)] < P(M \mid X)P(Y \mid M)[1 - P(\alpha)], && (10)\\
&\Rightarrow [1 - P(\beta)] < \frac{P(M \mid X)P(Y \mid M)}{P(Y \mid X)}[1 - P(\alpha)], && (11)\\
&\Rightarrow P(\beta) > P(\alpha), && (12)
\end{aligned}$$

where in Eq. 7, $I(\beta) = -\log[1 - P(\beta)]$ since β denotes bias interference, and higher interference probability implies greater information. In Eq. 9, $\log[1-P(\tilde{\beta})]$

is less than 0 and therefore can be omitted. In Eq. 11, according to the chain rule, $P(M \mid X)P(Y \mid M) = P(Y \mid X)$. Finally, $P(\beta) > P(\alpha)$ indicates that the impact of internal bias on the mediator M is relatively weaker. We thus conclude that a mediator offering information gain can alleviate the influence of internal biases. Thus, we employ high-information-gain sentences as mediators.

4.3 Overall Framework

The pipeline of the IGFD framework is illustrated in Fig. 4, which comprises the following three procedures. (a) IGFD first performs semantic decomposition on a question to obtain atomic questions. For each atomic question, its mediator sentences need to be individually identified. During this process, each sentence is also decomposed into atomic sentences, and the information gain between each atomic sentence and the atomic question is calculated. A sentence's information gain for an atomic question is the maximum value among its atomic sentences. (b) The mediator sentences identified for each atomic question are then integrated to form the final mediators for a question. (c) During the inference stage, these mediators are treated as the main material, while the remaining sentences serve as material, explicitly guiding the model via prompts.

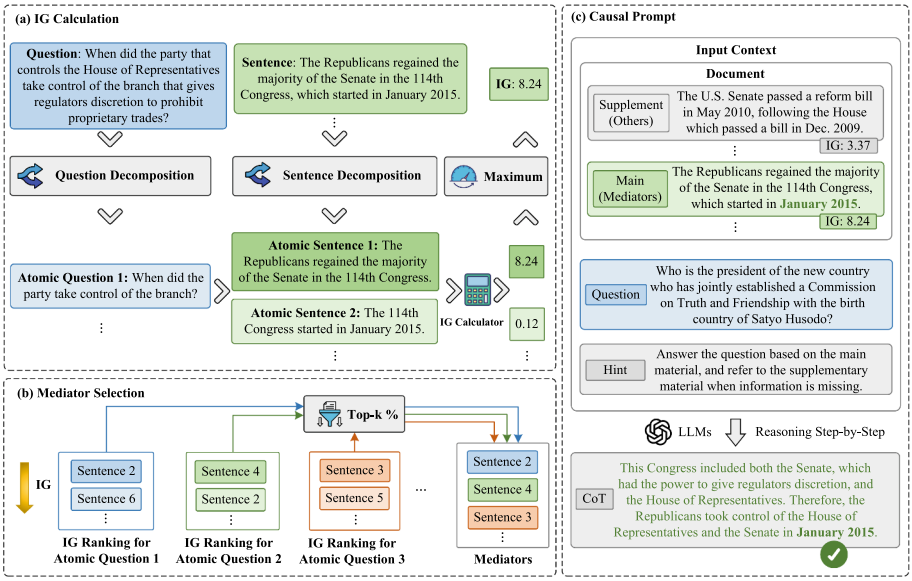

Fig. 4. The pipeline of the IGFD framework. (a) IG Calculation decomposes each sentence into atomic questions. (b) Mediator Selection integrates the top-k% information-gain sentences as mediators from each atomic question. (c) Causal Prompt explicitly describes the mediator sentences as materials to highlight essential evidence.

Simulate $\sum_m \boldsymbol{P(m \mid X)}$. This term reflects stratification according to the distribution of the mediator M. As we focus on semantic distribution, we use LLMs to perform semantic decomposition to obtain fine-grained atomic sentences:

$$\mathbf{s}_a = LLM(P_{dec}, s), \tag{13}$$

where $s \in d$ is a sentence in the document; P_{dec} denotes the prompt that guides LLMs to conduct semantic decomposition; $\mathbf{s}_a$ represents atomic sentences.

Simulate $\sum_{x'} \boldsymbol{P(Y \mid x', m)P(x')}$. This term represents the average effect of mediator m on the answer Y obtained by aggregating over input distribution x'. In this task, each question is independent, meaning that different instances cannot be merged. To address this, we semantically decompose a question into atomic questions by LLMs, to simulate different input distributions x':

$$\mathbf{q}_a = LLM(P_{dec}, q), \tag{14}$$

where $\mathbf{q}_a$ denotes the decomposed atomic questions. To simplify, we assume that all sub-questions are uniformly distributed; therefore, $P(x')$ is omitted.

Simulate $\sum_m \boldsymbol{P(m \mid X)} \sum_{x'} \boldsymbol{P(Y \mid x', m)P(x')}$. This term involves estimating the causal effect of X on Y by integrating all average effect of mediators. We first simulate the individual causal effect:

$$s_i = *top - k_{s \in d}(\max_{s_a \in \mathbf{s}_a} IG(q_a, s_a)), \tag{15}$$

where $IG(q_a, s_a)$ compute the information gain of atomic sentence s_a on atomic question q_a; the max operate extracts the maximum value among s's atomic sentences as the final IG value of sentence s to atomic question q_a; the $*top - k$ operate extracts the top k% information-gain sentences s_i from the document d as the mediators for atomic question q_a. We aggregate the mediator sentences of all atomic questions $\mathbf{q}_a$ and obtain the final mediators $\mathbf{s}_i$ for the question q.

To obtain debiasing predictions, we explicitly highlight the identified mediator sentences in the prompt and guide the LLMs to sample n CoTs:

$$\mathbf{c} = \sum_n LLM((\mathbf{s}_i, q), P_{cot}), \tag{16}$$

where P_{cot} denotes the prompt that guides LLMs to sample CoTs; $\sum_n$ represent sampling n times; $\mathbf{c}$ denote the sampled CoTs. Finally, CoTs are used to steer LLM reasoning, with the final debiased result obtained via majority voting:

$$\hat{y} = \arg\max \sum_{t=1}^{n} \mathbb{I}\big(LLM(\mathbf{c}_t)\big). \tag{17}$$

Although CoTs are still generated by LLMs, mediator selection is performed beforehand, allowing its advantages to facilitate the generation of CoT, resulting in more reasonable reasoning and thereby achieving effective debiasing.

4.4 Information Gain Calculation

This section provides a brief overview of how LLMs are used to implement the IG Calculation. According to Eq. 4, the first step is to compute the information entropy. For LLMs, since the true data distribution is unknown, the negative log-likelihood (NLL) provides a computable surrogate for entropy [15]. Specifically, the entropy of sentence s to question q can be estimated as:

$$\hat{H}_\theta(q \mid s) = -\sum_{t \in S(q)} \log_2 p_\theta(y_t \mid s, y_{<t}), \tag{18}$$

where $S(q)$ denotes the span of the question; t denotes its token; $p_\theta(y_t \mid s, y_{<t})$ represent the probability assigned by the model to token t. In implementation, this corresponds to performing a forward pass over the input prompt [*Material: s provide information gain for Question: q*], applying log-softmax over the output logits, gathering token log-probabilities according to ground-truth labels, and summing them over the question span. Building upon this operator, the information gain of the sentence s to the question q can be estimated as:

$$\mathrm{IG}(q, s) = \hat{H}_\theta(q \mid \varnothing) - \hat{H}_\theta(q \mid s), \tag{19}$$

where $\hat{H}_\theta(q \mid \varnothing)$ using an empty context in the input prompt, which indicates the initial information entropy of the question.

5 Experiments

5.1 Datasets and Evaluation

Following previous works [38,39], we selected the SciQ, HotpotQA, WikiHop, and MuSiQue datasets for evaluation, where a document is provided as external knowledge for question answering (QA). The detailed information for each dataset is as follows. **SciQ** [29] is a multiple-choice science QA dataset. We select the 1,000 longest-document instances from the training set. **HotpotQA** [35] is a multi-hop QA benchmark with open-ended and yes/no questions. We randomly select 1,000 instances from the official validation set (distractor version). **WikiHop** [30] is a multi-choice, multi-hop reasoning dataset. We randomly select 1000 instances from the training set. **MuSiQue** is also a multi-hop QA benchmark with longer documents. We select the more challenging part of the test set, which contains 1,165 instances that require more than three reasoning hops. More specifically, for SciQ and WikiHop, we prompt models to generate free-form answers instead of selecting from candidates. In addition, we employ Exact Match (EM) and F1 as evaluation metrics, following prior work [35].

5.2 Baselines and Backbone Models

We compare our framework with the following competitive baselines. **CoT without context (CoT w/o ctx)** [28] applies CoT prompting without providing the

document, reasoning purely based on the model's internal knowledge. **CoT** [28] guides the model to perform step-by-step reasoning to reach an answer. **CoT Self-Consistency (CoT-SC)** [27] prompts the model to generate multiple reasoning chains for a given query, and majority voting is used to determine the final answer. **Causal-guided Active Learning (CAL)** [24] identifies the biased instances by causal invariant semantic relationship and induces the bias patterns. **Debiasing CoT (DeCoT)** [31] introduces CoT as a mediator to conduct front-door adjustment by identifying stable sampled CoTs through entity intervention within the document. **Causal Prompting (CP)** [37] also employs CoT as a mediator, but it debiases CoTs by sampling and voting. **Conditional Front-Door Prompting (CFDP)** [39] is similar to DeCoT, but it identifies stable CoTs by performing entity intervention within the CoTs themselves.

To ensure diversity and comparability, we select two open-source and one closed-source LLMs as the backbones: Llama-3.1-8B [5], Qwen3-8B [34], and GPT-4o-mini [8]. These LLMs differ in training strategies and open-source versus closed-source design, providing a comprehensive foundation for evaluation.

5.3 Implementation Details

In Sect. 4.3, following prior work [39], the sample count n is set to 5. We employ Qwen3-8B as the IG calculator for the closed-source model. To improve efficiency, the open-source models are deployed in 8-bit quantized form on NVIDIA L20 GPUs. For stability, the CoT sampling temperature is set to 1, and 0 otherwise.

5.4 Main Results

As shown in Table 1, we draw the following conclusions. (1) **IGFD achieves SOTA performance**. Across nearly all datasets and backbone models, IGFD consistently outperforms existing causal debiasing frameworks. For example, with GPT-4o-mini, IGFD improves EM and F1 by 1.64% in EM and 1.61% on average over the second-best method, DeCoT. These results demonstrate the effectiveness and generalizability of our information-gain guided front-door adjustment strategy. (2) **IGFD yields larger gains on more knowledge-intensive settings**. In particular, on MuSiQue, which contains the longest documents, IGFD achieves the largest improvements over the naive CoT baseline. For instance, with Qwen3-8B, IGFD improves EM by 8.33% and F1 by 10.56% compared with CoT. (3) **CoT-mediated baselines provide limited benefits**. Methods that use CoT as the mediator, such as DeCoT, CP, and CFDP, offer only marginal improvements and perform comparably to the naive CoT sampling baseline, CoT-SC. For example, on GPT-4o-mini, DeCoT exceeds CoT-SC by only 0.14% EM and 0.37% F1 on average. This suggests that using CoT as the mediator may be affected by inner biases in LLM generation.

5.5 Ablation Study

To assess the effectiveness of our strategy, we conduct ablation studies, as reported in Table 2. We draw three conclusions. (1) **Each component con-**

Table 1. The comparison results of the IGFD framework and seven baselines across three backbone LLMs on four knowledge-intensive tasks. The best results are highlighted in **bold**, while the second-best results are underlined.

Method	SciQ		HotpotQA		WikiHop		MuSiQue		Average	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Llama-3.1-8B										
CoT w/o ctx	32.80	43.69	18.10	25.98	12.30	21.48	2.58	7.20	16.45	24.59
CoT	37.70	42.08	44.50	59.59	19.70	29.67	31.42	41.80	33.33	43.29
CoT-SC	51.80	59.20	46.90	60.99	21.40	30.62	35.45	46.96	38.89	49.44
CAL	32.90	37.05	44.30	59.46	21.70	30.71	32.27	42.66	32.79	42.47
DeCoT	49.60	57.81	42.00	58.18	20.60	30.94	31.07	42.50	35.82	47.36
CP	50.40	58.77	48.30	62.51	21.20	30.70	**36.48**	**47.46**	39.10	49.86
CFDP	49.20	56.55	43.80	59.03	20.00	30.02	31.24	42.13	36.06	46.93
Ours	**52.50**	**59.72**	**49.30**	**63.52**	**22.10**	**31.78**	36.05	46.44	**39.99**	**50.37**
Qwen3-8B										
CoT w/o ctx	53.20	64.33	17.70	26.00	11.50	18.74	4.21	11.28	21.65	30.09
CoT	78.20	87.91	50.60	64.59	21.80	31.82	34.25	44.87	46.21	57.30
CoT-SC	77.80	87.42	51.60	66.64	23.30	34.15	40.26	51.21	48.24	59.86
CAL	77.60	87.59	50.50	64.28	21.30	31.45	35.19	45.47	46.15	57.20
DeCoT	78.30	87.86	52.10	66.39	23.20	33.74	38.88	50.30	48.12	59.57
CP	77.10	87.05	51.81	66.78	23.60	34.34	39.06	50.34	47.89	59.63
CFDP	78.20	87.75	51.70	66.06	22.90	33.40	38.45	49.76	47.81	59.24
Ours	**78.60**	**88.23**	**52.80**	**67.73**	**25.30**	**35.51**	**42.58**	**55.43**	**49.82**	**61.73**
GPT-4o-mini										
CoT w/o ctx	48.50	63.40	25.60	37.03	14.90	26.44	8.07	18.10	24.27	36.24
CoT	72.70	85.08	46.80	62.56	21.50	32.98	37.17	49.99	44.54	57.65
CoT-SC	71.90	83.79	47.10	63.05	24.10	35.30	38.80	51.63	45.48	58.44
CAL	73.30	85.51	46.90	62.93	20.70	32.32	37.25	49.89	44.54	57.66
DeCoT	72.50	84.66	47.40	63.08	23.60	35.32	38.97	52.18	45.62	58.81
CP	71.70	83.76	47.80	63.50	23.90	35.71	38.28	50.95	45.42	58.48
CFDP	72.50	84.19	48.10	63.43	22.10	33.95	39.14	52.29	45.46	58.47
Ours	**73.70**	**85.82**	**48.80**	**65.17**	**25.60**	**36.30**	**40.94**	**54.39**	**47.26**	**60.42**

tributes. Across all datasets, removing any single component consistently degrades performance, confirming the utility of each design choice. (2) **Information gain is the most critical**. Eliminating the information-gain module causes the largest drop in overall performance among all ablations. (3) **IGFD remains effective without sampling and voting**. Even when the sampling-and-voting module is removed, our method still significantly outperforms the CoT baseline.

Table 2. The results of the ablation study with Qwen3-8B backbone. "w/o Dec." denotes the removal of the semantic decomposition strategy; "w/o IG" indicates that the information-gain calculation is removed and replaced it with random values; "w/o Vot." represents the removal of the sampling and voting strategy.

Method	SciQ		HotpotQA		WikiHop		MuSiQue		Average	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Qwen3-8B										
CoT	78.20	87.91	50.60	64.59	21.80	31.82	34.25	44.87	46.21	57.30
IGFD	**78.60**	**88.23**	**52.80**	**67.73**	**25.30**	**35.51**	**42.58**	**55.43**	**49.82**	**61.73**
w/o Dec.	78.30	88.10	51.50	66.45	23.30	33.45	37.64	50.16	47.69	59.54
w/o IG	78.40	88.14	51.70	66.44	23.10	33.22	37.08	49.15	47.57	59.24
w/o Vot.	78.30	88.11	51.60	67.03	23.80	33.87	39.23	51.15	48.23	60.04

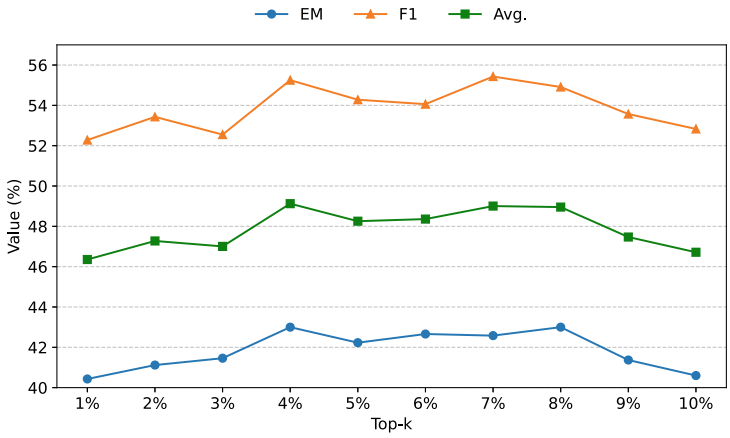

Fig. 5. Top-k hyperparameter study on MuSiQue with Qwen3-8B. The figure shows the trends of EM, F1, and their average (Avg.) as k increases.

5.6 Hyperparameter Study

We determine the value of top-k via a hyperparameter analysis, as shown in Fig. 5. We observe a consistent trend: as k increases, performance first improves and then declines. This behavior is intuitive. Selecting more mediator sentences initially helps capture additional relevant evidence; however, once k becomes too large, irrelevant sentences are increasingly included, which can distract the model. We observe similar patterns across other models and datasets. Based on this empirical finding, we set k to 7%, where the average value reaches its peak.

6 Conclusion

In this paper, we tackle the challenge of internal bias in large language models (LLMs), particularly in knowledge-intensive reasoning settings. Prior causality-based debiasing methods, especially those that perform front-door adjustment using Chain-of-Thought (CoT) as the mediator, remain vulnerable because the generated CoT itself can be influenced by internal biases in LLM generation. To address this limitation, we propose IGFD, a novel framework that replaces CoTs with highinformation-gain sentences extracted as mediators.

We provide a formal proof of the soundness of this information-gainguided paradigm and demonstrate its ability to effectively mitigate internal biases by extracting the most relevant and informative content. Specifically, our semantic decomposition strategy enables accurate estimation of sentence-level information gain, allowing the model to emphasize key reasoning cues while de-emphasizing distractors. Experiments on a range of knowledge-intensive tasks, covering both open- and closed-source LLMs, demonstrate that IGFD delivers state-of-the-art results and consistently outperforms existing causal debiasing approaches.

7 Limitation and Further Work

Although IGFD effectively mitigates internal biases in LLMs, it still inherits certain limitations from prior front-door adjustment methods [31,37,39]. Specifically, when simulating front-door adjustment, our approach continues to rely on sampling and voting to stabilize causal effect estimation, which inevitably increases inference-time computational overhead. Nevertheless, our ablation studies show that removing the voting strategy still yields substantial gains over strong baselines, suggesting that IGFD is intrinsically robust and efficient despite this limitation. Looking ahead, we plan to extend IGFD to broader reasoning settings, including multi-modal and multi-hop tasks, and to explore dynamic mediator selection strategies that further improve interpretability and robustness. Overall, our work offers a new perspective for future causality-based LLM research, suggesting that information-theoretic tools may enable more appropriate causal estimation for LLMs, since an LLM can be viewed as an opaque information-processing system with latent internal variables.

Acknowledgments. This work was supported by the grant from the National Natural Science Foundation of China (NSFC) project (No. 62276193) and the Fundamental Research Funds for the Central Universities, China (Grant No. 2042022dx0001).

References

1. Arslan, M., Ghanem, H., Munawar, S., Cruz, C.: A survey on rag with LLMs. Proc. Comput. Sci. **246**, 3781–3790 (2024)
2. Atil, B., Chittams, A., Fu, L., Ture, F., Xu, L., Baldwin, B.: LLM stability: a detailed analysis with some surprises. CoRR (2024)

3. Bai, X., Wang, A., Sucholutsky, I., Griffiths, T.L.: Explicitly unbiased large language models still form biased associations. Proc. Natl. Acad. Sci. **122**(8), e2416228122 (2025)
4. Chang, H., et al.: How do large language models acquire factual knowledge during pretraining? Adv. Neural. Inf. Process. Syst. **37**, 60626–60668 (2024)
5. Dubey, A., et al.: The llama 3 herd of models. arXiv e-prints pp. arXiv–2407 (2024)
6. Fan, W., et al.: A survey on rag meeting LLMs: towards retrieval-augmented large language models. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 6491–6501 (2024)
7. Huang, J., Chang, K.C.C.: Towards reasoning in large language models: a survey. In: Findings of the Association for Computational Linguistics: ACL 2023, pp. 1049–1065 (2023)
8. Hurst, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
9. Jin, B., Yoon, J., Han, J., Arik, S.O.: Long-context LLMs meet rag: Overcoming challenges for long inputs in rag. arXiv preprint arXiv:2410.05983 (2024)
10. Kang, C., Choi, J.: Impact of co-occurrence on factual knowledge of large language models. In: Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 7721–7735 (2023)
11. Li, X., Chen, K., Long, Y., Zhang, M.: LLM with relation classifier for document-level relation extraction. arXiv preprint arXiv:2408.13889 (2024)
12. Li, Y., Li, J., Jiang, J., Zhang, M.: Enhancing document-level translation of large language model via translation mixed-instructions. arXiv preprint arXiv:2401.08088 (2024)
13. Lin, L., Wang, L., Guo, J., Wong, K.F.: Investigating bias in LLM-based bias detection: disparities between LLMs and human perception. In: COLING (2025)
14. Liu, B., et al.: Improving LLM-based document-level machine translation with multi-knowledge fusion. arXiv preprint arXiv:2503.12152 (2025)
15. Meister, C., Wiher, G., Pimentel, T., Cotterell, R.: On the probability–quality paradox in language generation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 36–45 (2022)
16. Pearl, J.: Causal inference in statistics. an overview (2009)
17. Pearl, J.: Causality. Cambridge university press (2009)
18. Pearl, J., Glymour, M., Jewell, N.P.: Causal inference in statistics: A primer. John Wiley & Sons (2016)
19. Pearl, J., Mackenzie, D.: The book of why: the new science of cause and effect. Basic books (2018)
20. Shaikh, O., Zhang, H., Held, W., Bernstein, M., Yang, D.: On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics (2023)
21. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948)
22. Shi, F., et al.: Large language models can be easily distracted by irrelevant context. In: International Conference on Machine Learning, pp. 31210–31227. PMLR (2023)
23. Sun, Q., Huang, K., Yang, X., Tong, R., Zhang, K., Poria, S.: Consistency guided knowledge retrieval and denoising in LLMs for zero-shot document-level relation triplet extraction. In: Proceedings of the ACM Web Conference 2024, pp. 4407–4416 (2024)

24. Sun, Z., et al.: Causal-guided active learning for debiasing large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 14455–14469 (2024)
25. Trivedi, H., Balasubramanian, N., Khot, T., Sabharwal, A.: Musique: Multi-hop questions via single-hop question composition. Trans. Assoc. Comput. Linguist. **10**, 539–554 (2022)
26. Wang, L., et al.: Document-level machine translation with large language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 16646–16661 (2023)
27. Wang, X., et al.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)
28. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural. Inf. Process. Syst. **35**, 24824–24837 (2022)
29. Welbl, J., Liu, N.F., Gardner, M.: Crowdsourcing multiple choice science questions. arXiv preprint arXiv:1707.06209 (2017)
30. Welbl, J., Stenetorp, P., Riedel, S.: Constructing datasets for multi-hop reading comprehension across documents. Trans. Assoc. Comput. Linguist. **6**, 287–302 (2018)
31. Wu, J., et al.: 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), pp. 14073–14087 (2024)
32. Wu, M., Vu, T.T., Qu, L., Foster, G., Haffari, G.: Adapting large language models for document-level machine translation. arXiv preprint arXiv:2401.06468 (2024)
33. Wu, X., et al.: Lifbench: Evaluating the instruction following performance and stability of large language models in long-context scenarios. arXiv preprint arXiv:2411.07037 (2024)
34. Yang, A., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
35. Yang, Z., et al.: Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2018)
36. Zhai, Y., et al.: Investigating the catastrophic forgetting in multimodal large language models. arXiv preprint arXiv:2309.10313 (2023)
37. Zhang, C., Zhang, L., Wu, J., He, Y., Zhou, D.: Causal prompting: Debiasing large language model prompting based on front-door adjustment. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 25842–25850 (2025)
38. Zhang, C., Zhang, L., Zhou, D.: Causal walk: debiasing multi-hop fact verification with front-door adjustment. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 19533–19541 (2024)
39. Zhao, B., et al.: Unbiased reasoning for knowledge-intensive tasks in large language models via conditional front-door adjustment. arXiv preprint arXiv:2508.16910 (2025)
40. Zheng, L., et al.: Judging LLM-as-a-judge with mt-bench and chatbot arena. Adv. Neural. Inf. Process. Syst. **36**, 46595–46623 (2023)
41. Zhong, H., Wei, X., Zhang, H.: Leveraging LLM for enhancing document-level relation extraction with correction and completion. In: 2025 8th International Conference on Advanced Algorithms and Control Engineering (ICAACE), pp. 2364–2369. IEEE (2025)