# Counterfactual Multihop QA: A Cause-Effect Approach for Reducing Disconnected Reasoning

**Wangzhen Guo** **Qinkang Gong** **Yanghui Rao** **Hanjiang Lai***
School of Computer Science, Sun Yat-Sen University
{guowzh6, gongqk}@mail2.sysu.edu.cn {raoyangh, laihanj3}@mail.sysu.edu.cn

## Abstract

Multi-hop QA requires reasoning over multiple supporting facts to answer the question. However, the existing QA models always rely on shortcuts, e.g., providing the true answer by only one fact, rather than multi-hop reasoning, which is referred to as *disconnected reasoning* problem. To alleviate this issue, we propose a novel counterfactual multihop QA, a cause-effect approach that enables to reduce the disconnected reasoning. It builds upon explicitly modeling of causality: 1) the direct causal effects of disconnected reasoning and 2) the causal effect of true multi-hop reasoning from the total causal effect. With the causal graph, a counterfactual inference is proposed to disentangle the disconnected reasoning from the total causal effect, which provides us a new perspective and technology to learn a QA model that exploits the true multi-hop reasoning instead of shortcuts. Extensive experiments have been conducted on the benchmark HotpotQA dataset, which demonstrate that the proposed method can achieve notable improvement on reducing disconnected reasoning. For example, our method achieves 5.8% higher points of its $\text{Supp}_s$ score on HotpotQA through true multihop reasoning. The code is available at https://github.com/guowzh/CFMQA.

## 1 Introduction

Multi-hop question answering (QA) (Groeneveld et al., 2020; Ding et al., 2019; Asai et al., 2019; Shao et al., 2020) requires the model to reason over multiple supporting facts to correctly answer a complex question. It is a more challenging task than the single-hop QA since not only the correct answer but the explicit reasoning across multiple evidences should be provided.

Hence, recent work (Groeneveld et al., 2020; Fang et al., 2019) has shown that the multi-hop QA is always formulated as two sub-tasks: question answering and support identification. For example, Groeneveld et al. (2020) solved the multi-hop QA via 1) question answering which uses a BERT (Devlin et al., 2018) span prediction model to answer the questions, and 2) support identification which aims to identify the supporting sentences. Yavuz et al. (2022) proposed PATHFID based on fusion-in-decoder (FID) (Izacard and Grave, 2020) for question answering and support identification.

However, dividing the multi-hop QA into question answering and support identification doesn't mean that QA models answer the questions according to what the multi-hop QA wants. These QA models are all based on the large pre-training language models, e.g., BERT (Devlin et al., 2018) and XLNet (Yang et al., 2019), these black-box language models are rather opaque models in terms of the reasoning processes. It may results in one main problem of multihop QA models: *disconnected reasoning* (Trivedi et al., 2020), which allows the QA models to exploit the reasoning shortcuts (Jiang and Bansal, 2019; Lee et al., 2021) instead of multi-hop reasoning to cheat and obtain the right answer. Taking Fig. 1 as an example, to answer the question *"until when in the U.S. Senate"*, multi-hop QA requires to answer the question with a true *reasoning path* (e.g., the second paragraph → the third paragraph in Fig. 1). However, the black-box QA models can also infer the correct answer by just utilizing the types of problems, e.g., we can find the corresponding fact *"from 2005 to 2008"* in the contexts to answer this type of question *"until when"* without reasoning.

To address the above problem, two issues should be considered: 1) how to measure the disconnected reasoning? To the best of our knowledge,Trivedi et al. (2020) firstly defined an evaluation measure, **DiRe** in short, to measure how much the QA model can cheat via disconnected reasoning. A probing dataset should be constructed to measure the DiRe. And 2) how to reduce the disconnected reasoning?

*Corresponding Author



*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*
*Volume 1: Long Papers*, pages 4214–4226

One possible solution is to strengthen the training dataset via extra annotations or adversarial examples, which makes it cannot find the correct answers by only one supporting fact. For example, Jiang and Bansal (2019) constructed the adversarial examples to generate better distractor facts. Besides, counterfactual intervention (Lee et al., 2021; Ye et al., 2021) had also been explored to change the distribution of the training dataset. However, when these existing approaches decrease the disconnected reasoning, the original performance also drops significantly. It is still challenging to reduce disconnected reasoning while maintaining the same accuracy on the original test set.

Motivated by causal inference (Pearl and Mackenzie, 2018; Pearl, 2022; Niu et al., 2021), we utilize the counterfactual reasoning to reduce the disconnected reasoning in multi-hop QA and also obtain the robust performance on the original dataset. We formalize a causal graph to reflect the causal relationships between question ($Q$), contexts, and answer ($Y$). To evaluate the disconnected reasoning, contexts are further divided into two subsets: $S$ is a supporting fact and $C$ are the remaining supporting facts. Hence, we can formulate the disconnected reasoning as two natural direct causal effects of $(Q, S)$ and $(Q, C)$ on $Y$ as shown in Fig. 1. With the proposed causal graph, we can relieve the disconnected reasoning by disentangling the two natural direct effects and the true multi-hop reasoning from the total causal effect. A novel counterfactual multihop QA is proposed to disentangle them from the total causal effect. We utilize the generated probing dataset proposed by (Trivedi et al., 2020) and DiRe to measure how much the proposed multi-hop QA model can reduce the disconnected reasoning. Experiment results show that our approach can substantially decrease the disconnected reasoning while guaranteeing the strong performance on the original test set. The results indicate that the proposed approach can improve the true multi-hop reasoning capability.

The main contribution of this paper is threefold. Firstly, our counterfactual multi-hop QA model formulates disconnected reasoning as two direct causal effects on the answer, which is a new perspective and technology to learn true multi-hop reasoning. Secondly, our approach achieves notable improvement on reducing disconnected reasoning compared to various baselines. Thirdly, our causal-effect approach is model-agnostic and can be used for reducing disconnected reasoning in many multi-hop QA architectures.

## 2 Related Work

Multi-hop question answering (QA) requires the model to retrieve the supporting facts to predict the answer. Many approaches and datasets have been proposed to train QA systems. For example, HotpotQA (Yang et al., 2018) dataset is a widely used dataset for multi-hop QA, which consists of fullwiki setting (Das et al., 2019; Nie et al., 2019; Qi et al., 2019; Chen et al., 2019; Li et al., 2021; Xiong et al., 2020) and distractor setting (Min et al., 2019b; Nishida et al., 2019; Qiu et al., 2019; Jiang and Bansal, 2019; Trivedi et al., 2020).

In fullwiki setting, it first finds relevant facts from all Wikipedia articles and then answers the multi-hop QA with the found facts. The retrieval model is important in this setting. For instance, SMRS (Nie et al., 2019) and DPR (Karpukhin et al., 2020) found the implicit importance of retrieving relevant information in the semantic space. Entity-centric (Das et al., 2019), CogQA (Ding et al., 2019) and Golden Retriever (Qi et al., 2019) explicitly used the entity that is mentioned or reformed in query key words to retrieve the next hop document. Furthermore, PathRetriever (Asai et al., 2019) and HopRetriever (Li et al., 2021) can iteratively select the documents to form a paragraph-level reason path using RNN. MDPR (Xiong et al., 2020) retrieved passages only using dense query vectors many times. These methods hardly discuss the QA model's disconnected reasoning problem.

In the distractor setting, 10 paragraphs, two gold paragraphs and eight distractors, are given. Many methods have been proposed to strengthen the model's capability of multi-hop reasoning, using graph neural network (Qiu et al., 2019; Fang et al., 2019; Shao et al., 2020) or adversarial examples or counterfactual examples (Jiang and Bansal, 2019; Lee et al., 2021) or the sufficiency of the supporting evidences (Trivedi et al., 2020) or make use of the pre-trained language models (Zhao et al., 2020; Zaheer et al., 2020).

However, Min et al. (2019a) demonstrated that many compositional questions in HotpotQA can be answered with a single hop. It means that QA models can take shortcuts instead of multi-hop reasoning to produce the corrected answer. To relieve the issue, Jiang and Bansal (2019) added adversarial examples as hard distractors during training. Re-

cently, Trivedi et al. (2020) proposed an approach, DiRe, to measure the model's disconnected reasoning behavior and used the supporting sufficiency label to reduce the disconnected reasoning. Lee et al. (2021) selected the supporting evidence according to the sentence causality to the predicted answer, which guarantees the explainability of the behavior of the model. While the original performance also drops when reducing the disconnected reasoning.

**Causal Inference.** Recently, causal inference (Pearl and Mackenzie, 2018; Pearl, 2022) has been applied to many tasks of natural language processing and computer vision, and it shows promising results and provides strong interpretability and generalizability. The representative works include counterfactual intervention for visual attention (Rao et al., 2021), causal effect disentanglement for VQA (Niu et al., 2021), the back-door and front-door adjustments (Zhu et al., 2021; Wang et al., 2021; Yang et al., 2021). Our method can be viewed as a complement of the recent approaches that utilize the counterfactual inference (Lee et al., 2021; Ye et al., 2021) to identify the supporting facts and predict the answer.

## 3 Preliminaries

In this section, we use the theory of causal inference (Pearl and Mackenzie, 2018; Pearl, 2022) to formalize our multi-hop reasoning method. Suppose that we have a multi-hop dataset $D$ and each instance has the form of $(Q, P; Y)$, where $Q$ is a question and $P = \{s_1, s_2, \cdots, s_n\}$ is a context consisting of a set of $n$ paragraphs. And $Y$ is the ground-truth label. Given a question $Q$ with multiple paragraphs as a context $P$, the multi-hop QA models are required to identify which paragraphs are the supporting facts and predict an answer using the supporting facts.

**Causal graph**. In multi-hop QA, multiple supporting facts are required to predict the answer. While QA models may use only one fact to give the answer, which is referred to as disconnected reasoning. For example, given a question $q$ and only a paragraph $s$, the disconnected reasoning model can predict the correct answer or correctly determine whether the paragraph $s$ is the supporting fact. Hence, to define the causal graph of disconnected reasoning, the context $P$ is further divided into a paragraph $S$ and the remaining paragraphs $C$. Now the $(Q, P; Y)$ becomes $(Q, S, C; Y)$. That is each instance $(Q, P; Y)$ is converted into $n$ examples, i.e., $(q, s_1, C = \{s_2, \cdots, s_n\}; Y(s_1))$, $\cdots$, $(q, s_n, C = \{s_1, \cdots, s_{n-1}\}; Y(s_n))$, where $Y(s_i) = \{F_{s_i}; A\}$ includes the supporting fact and answer, where $A$ is the answer and $F_{s_i} = 1$ means the paragraph $s$ is the supporting fact otherwise it is not. For each example, we consider the disconnected reasoning with the fact $S$ (not $C$).

The causal graph for multi-hop QA is shown in Figure 1 (a), where nodes denote the variables and directed edges represent the causal-and-effect relationships between variables. The paths in Figure 1 (a) are as follows.

$(Q, S) \rightarrow K_1 \rightarrow Y$: $(Q, S) \rightarrow K_1$ denotes that the feature/knowledge $K_1$ is extracted from the question $(Q)$ and the paragraph $(S)$ via the QA model backbone, e.g., BERT. $K_1 \rightarrow Y$ represents the process that the label $Y$ is predicted by only using the $K_1$.

$(Q, C) \rightarrow K_2 \rightarrow Y$: Similarly, the feature/knowledge $K_2$ extracted from the question $(Q)$ and the remaining paragraphs $(C)$ is used to predict the label $Y$.

$(Q, S, C) \rightarrow (K_1, K_2) \rightarrow K \rightarrow Y$: This path indicates that the QA model predicts the label $Y$ based on both the $K_1$ and $K_2$.

Based on the above, the effect of $Q, S, C$ on Y can be divided into: 1) shortcut impacts, e.g., $(Q, S) \rightarrow K_1 \rightarrow Y$ and $(Q, C) \rightarrow K_2 \rightarrow Y$, and 2) reasoning impact, e.g., $(Q, S, C) \rightarrow (K_1, K_2) \rightarrow K \rightarrow Y$. The shortcut impacts capture the direct effect of $(Q, S)$ or $(Q, C)$ on $Y$ via $K_1 \rightarrow Y$ or $K_2 \rightarrow Y$. The reasoning impact captures the indirect effect of $(Q, S, C)$ on $Y$ via $K \rightarrow Y$.

Hence, to reduce the multi-hop QA model's disconnected reasoning proposed in (Trivedi et al., 2020), we should exclude shortcut impacts ($K_1 \rightarrow Y$ and $K_2 \rightarrow Y$) from the total effect.

**Counterfactual definitions**. Figure 1 (a) shows the causal graph. From causal graph to formula, we denote the value of $Y$, i.e., the answer $A$ (e.g., 2008) or the supporting paragraph $F_s$ (e.g., is the paragraph $s$ supporting fact?), would be obtained when question $Q$ is set to $q$, the paragraph $S$ is set to $s$ and the remaining paragraphs $c = P - s$ are used as the context $C$, which is defined as

$$Y_{q,s,c}(A) = Y(A \mid Q = q, S = s, C = c),$$
$$Y_{q,s,c}(s) = Y(s \mid Q = q, S = s, C = c).$$

For simplicity, we omit $A, s$ and unify both equations as $Y_{q,s,c} = Y_{q,s,c}(A)$ or $Y_{q,s,c} =$

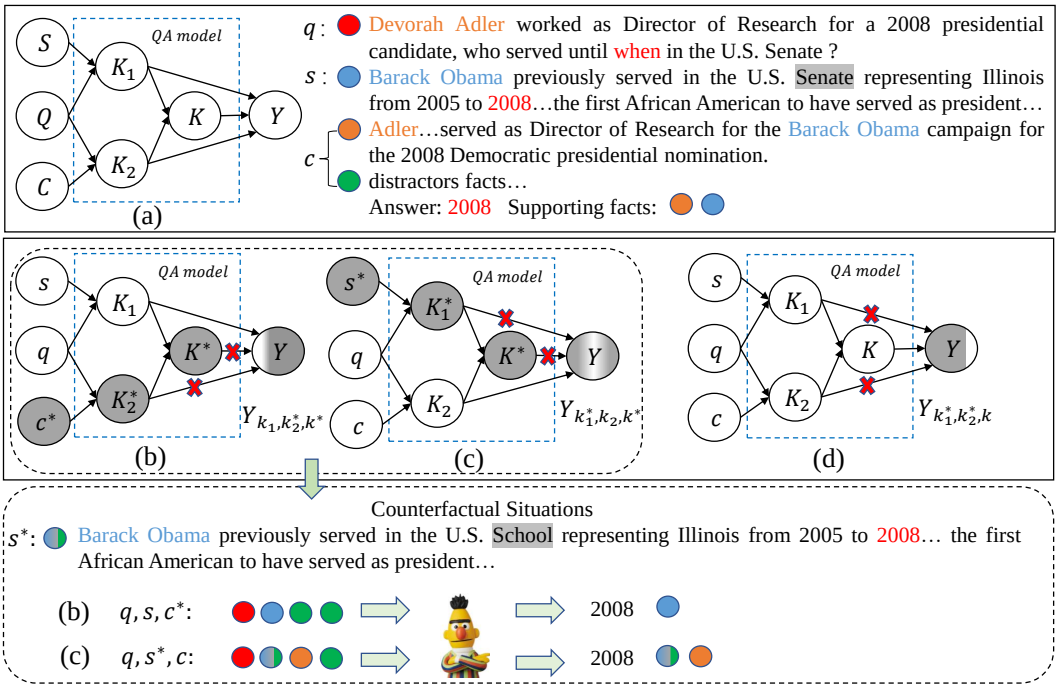


Figure 1: Illustration of disconnected reasoning in multi-hop QA, where red node denotes question $Q$, blue node is a supporting fact $S$, and orange and green nodes denote the remaining facts $C$. Deep gray nodes mean their variables are reference values instead of the given values, (e.g., $S = s^*$ instead of $S = s$). **(a):** Causal graph of multi-hop QA model; **(b):** is a possible scenario of disconnected reasoning, which uses only one fact $s$ to answer the question. **(c):** is another possibility of disconnected reasoning, e.g., the exclusive method to find whether $s$ is a supporting fact by a process of elimination of other facts $c$. **(d):** is the true multi-hop reasoning. All facts $s$ and $c$ are taken into considered to produce the answer.

$Y_{q,s,c}(s)$. Since the causal effect of $q, s, c$ on $Y$ via $K_1, K_2, K$ on $Y$, we have

$$Y_{k_1,k_2,k} = Y_{q,s,c} \tag{1}$$

in the following discussion.

To disentangle the shortcut impacts from the total causal effect, we use the counterfactual causal inference to block other effects. To model the $K_1 \rightarrow Y$, the counterfactual formulation

$$K_1 \rightarrow Y : Y_{k_1,k_2^*,k^*}, \tag{2}$$

which describes the situation where $K_1$ is set to the original value $k_1$ and $K$ and $K_2$ are blocked. The $k^*$ and $k_2^*$ are the counterfactual notations. The $k_1$ and $k^*, k_2^*$ represent the the two situations where the $k_1$ is under treatment in the factual scenario and $k^*, k_2^*$ are not under treatment (Pearl and Mackenzie, 2018) in the counterfactual scenario. The same definitions for other two effects as

$$K_2 \rightarrow Y : Y_{k_1^*,k_2,k^*}, \tag{3}$$

and

$$K \rightarrow Y : Y_{k_1^*,k_2^*,k}. \tag{4}$$

**Causal effects**. According to the counterfactual definitions, the total effect of $q, s, c$ on $Y$ can be decomposed into the natural direct effects of $K_1, K_2$ on $Y$ and the effect of $K$ on $Y$ as discussed before. The two natural direct effects cause the disconnected reasoning problem. The effect of $K \rightarrow Y$ is the desired multi-hop reasoning.

As shown in Figure 1 (b), the effect of $K_1$ on $Y$ with $K_2, K$ blocked and the effect of $K_2$ on $Y$ with $K_1, K$ blocked can be easily obtained by setting the $S$ or $C$ to counterfactual values (please refer to Section 4 for more details). While the effect of $K$ on $Y$ can not be obtained by changing the values $S$/$C$ to $S^*$/$C^*$. We follow (Niu et al., 2021) and total indirect effect (TIE) is used to express the effect of $K$ on $Y$, which is formulated as

$$Y_{k_1^*,k_2^*,k} = Y_{k_1,k_2,k} - Y_{k_1,k_2,k^*}. \tag{5}$$

## 4 Counterfactual Multihop QA

Following the former formulations, we propose to construct the counterfactual examples to estimate the natural direct effect of $K_1$ and $K_2$, as well as using parameters to estimate the total indirect effect

of $K$. And our calculation of $Y$ in Eq. (2), (3) and (5) is parametrized by a neural multi-hop QA model $\mathcal{F}$. Please note that $\mathcal{F}$ can be any multi-hop QA model and our method is model-agnostic.

### 4.1 Disentanglement of causal effect

$K_1 \rightarrow Y$. Specifically, in Eq. (2), the $Y_{k_1,k_2^*,k^*}$ describes the situation where $K_1$ is set to the factual value with $Q = q, S = s$ as inputs, and $K_2/K$ are set to the counterfactual values. Taking Figure 1 as an example, the QA model only considers the interaction between question $q$ and a given paragraph $s$. The remaining paragraphs $c$ are not given. It is the disconnected reasoning (Trivedi et al., 2020). To obtain the counterfactual values of $K_2$ and $K$, we can set the context $C$ as its counterfactual sample, and we have

$$Y_{k_1,k_2^*,k^*} = Y_{q,s,c^*} = \mathcal{F}(q, s, c^*). \tag{6}$$

In this paper, we randomly sample the remaining contexts from the training set to construct the counterfactual $c^*$. It represents the no-treatment or the prior knowledge of the remaining context. In the implementation, we randomly sample the remaining contexts in a mini-batch to replace the $c$ in the original triple example $(q, s, c)$ and obtain the corresponding counterfactual triple example $(q, s, c^*)$. With that, we can feed it into the QA model to get $Y_{q,s,c^*}$.

$K_2 \rightarrow Y$. Similarly, in Eq. (3), the $Y_{k_1^*,k_2,k^*}$ describes the situation where $K_2$ is set to the factual value with the inputs $Q = q$ and $C = c$. The $K_1$ and $K$ are set to the counterfactual values with the counterfactual sample $S = s^*$ as input, which is defined as

$$Y_{k_1^*,k_2,k^*} = Y_{q,s^*,c} = \mathcal{F}(q, s^*, c). \tag{7}$$

One may argue that how to predict the label when $S$ is set to the counterfactual values? As the example shown in Figure 1, even without the paragraph $s$ as input, the QA model still can infer the paragraph $s$ is supporting fact via wrong reasoning: since all paragraphs in $C$ do not include the words about time range, the rest paragraph $s$ should be the supporting fact to answer the "until when" question. It is the exclusive method. The wrong reasoning is caused by an incorrect interaction between the paragraph $s$ and the remaining context $c$. Hence, the $S$ is set to the counterfactual values that can correct such incorrect interactions.

Hence, in the implementation, we use the adversarial examples as suggested in (Jiang and Bansal, 2019) to construct the counterfactual $s^*$, which aims to remove the incorrect interaction and perform multi-hop reasoning. Specifically, we randomly perturb the 15% tokens of the paragraph $s$, 80% of which will be replaced by other random tokens, 10% of which will be replaced by the mask token (e.g. $[MASK]$) of the tokenizer, and 10% of which will keep unchanged. After that, we obtain another counterfactual triple example $(q, s^*, c)$, we can feed it into QA model to get $Y_{q,s^*,c}$.

$K \rightarrow Y$. In Eq. (5), $Y_{q,s,c}$ indicates that the question $q$, paragraph $s$ and remaining context $c$ are visible to the QA model $\mathcal{F}$:

$$Y_{k_1,k_2,k} = Y_{q,s,c} = \mathcal{F}(q, s, c). \tag{8}$$

The main problem is that $Y_{k_1,k_2,k^*}$ is unknown. It is also hard to use the counterfactual samples of $Q, S, C$ to obtain its value since the $q, s$ and $c$ should be the factual values for $k_1, k_2$. In this paper, we follow the work (Niu et al., 2021) and also assume the model will guess the output probability under the no-treatment condition of $k^*$, which is represented as

$$Y_{k_1,k_2,k^*} = \mathcal{C}, \tag{9}$$

where $\mathcal{C}$ is the output and a learnable parameter. Similar to Counterfactual VQA (Niu et al., 2021), we guarantee a safe estimation of $Y_{k_1,k_2,k^*}$ in expectation.

**Training objective**. From the above discussion, we can estimate the total causal effect:

$$Y \leftarrow Y_{q,s,c^*} + Y_{q,s^*,c} + Y_{q,s,c} - \mathcal{C}. \tag{10}$$

Note that $Y$ can be any ground-truth labels of answer span prediction, supporting facts identification or answer type prediction. See Appendix A.1 for further implementation details.

### 4.2 Training and Inference

**Training**. The model $\mathcal{F}$ is expected to disentangle the two natural direct effects and the true multi-hop effect from the total causal effect. To achieve this goal, we apply the Eq. (10) to train the QA model $\mathcal{F}$. Our training strategy follows the HGN (Fang et al., 2019):

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{start} + \mathcal{L}_{end} + \lambda\mathcal{L}_{sent} \\ + \mathcal{L}_{para} + \mathcal{L}_{type} + \mathcal{L}_{entity},\end{aligned} \tag{11}$$

| | Ans | | $\text{Supp}_\text{p}$ | | $\text{Supp}_\text{s}$ | | Ans + $\text{Supp}_\text{p}$ | | Ans + $\text{Supp}_\text{s}$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | original | dire↓ | original | dire↓ | original | dire↓ | original | dire↓ | original | dire↓ |
| BERT | 74.2 | 47.5 | 94.1 | 73.9 | 83.8 | 64.1 | 71.0 | 36.5 | 64.2 | 32.5 |
| +ours | 74.0 | **45.5** | **95.4** | **67.5** | 82.8 | **54.6** | **71.6** | **31.8** | 63.9 | **26.7** |
| XLNET | 76.2 | 50.3 | 96.5 | 75.0 | 86.6 | 64.8 | 74.4 | 39.1 | 68.0 | 34.6 |
| +ours | 75.9 | **49.8** | **96.6** | **74.1** | 86.6 | **63.6** | 74.1 | **38.0** | 67.9 | **33.6** |
| DFGN | 71.7 | 44.5 | 94.4 | 73.8 | 83.8 | 64.0 | 68.7 | 34.2 | 62.1 | 30.4 |
| +ours | **73.3** | 48.3 | **96.1** | **72.1** | **85.4** | **61.6** | 71.5 | 36.0 | 65.1 | 31.7 |
| HGN | 73.3 | 47.0 | 91.1 | 67.4 | 81.4 | 59.0 | 68.3 | 33.6 | 62.0 | 30.2 |
| +ours | 70.9 | **41.1** | **93.4** | 67.6 | **83.5** | **58.1** | 67.6 | **28.9** | 61.8 | **25.5** |

Table 1: F1 scores. The "original" denotes the model's performance on the development set of HotpotQA in the distractor setting, and the "dire" indicates that the model scores on the corresponding probing set, which measures how much disconnected reasoning the model can achieve. The smaller score of "dire" is better. We can see that the proposed method can reduce disconnected reasoning while maintaining the same accuracy on the original dataset.

| | Ans | | $\text{Supp}_\text{p}$ | | $\text{Supp}_\text{s}$ | | Ans + $\text{Supp}_\text{p}$ | | Ans + $\text{Supp}_\text{s}$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | original | dire↓ | original | dire↓ | original | dire↓ | original | dire↓ | original | dire↓ |
| BERT | 59.7 | 35.3 | 86.1 | 17.3 | 54.8 | 10.1 | 53.6 | 7.0 | 35.7 | 4.4 |
| +ours | **60.2** | **33.7** | **87.9** | **3.7** | 53.5 | **2.0** | **55.2** | **1.4** | **36.6** | **0.8** |
| XLNET | 62.0 | 38.0 | 91.8 | 14.5 | 58.9 | 8.5 | 58.6 | 6.5 | 40.1 | 4.2 |
| +ours | 61.7 | **37.6** | **92.1** | **6.9** | **59.3** | **3.9** | 58.5 | **2.6** | **40.9** | **1.6** |
| DFGN | 57.4 | 44.5 | 85.9 | 19.5 | 53.4 | 11.7 | 51.3 | 6.9 | 33.6 | 4.6 |
| +ours | **59.6** | **35.9** | **90.9** | **6.2** | **57.7** | **3.3** | **55.9** | **2.3** | **38.4** | **1.4** |
| HGN | 58.9 | 35.2 | 79.5 | 16.9 | 52.3 | 10.4 | 49.5 | 6.8 | 34.2 | 4.4 |
| +ours | 57.1 | **30.1** | **85.1** | **6.4** | **56.1** | **3.3** | **51.3** | **2.2** | **36.0** | **1.2** |

Table 2: EM scores on the development set and probing set of HotpotQA in the distractor setting.

where $\lambda$ is a hyper-parameter and each term of $\mathcal{L}$ is cross-entropy loss function. Specifically, for answer prediction, we utilize the Eq. (10) to obtain the predicted logits of the start and end position of the answer span, and respectively calculate the cross-entropy loss $\mathcal{L}_{start}$ and $\mathcal{L}_{end}$ with corresponding ground truth labels. As for supporting facts prediction, similarly, we use the Eq. (10) to calculate the predicted logits in sentence level and paragraph level, and then calculate $\mathcal{L}_{sent}$ and $\mathcal{L}_{para}$. We also apply our counterfactual reasoning method to identify the answer type (Qiu et al., 2019; Fang et al., 2019), which consists of yes, no, span and entity. We use the $[CLS]$ token as the global representation to predict the answer type under the Eq. (10) and calculate $\mathcal{L}_{type}$ with the ground truth label. Entity prediction ($\mathcal{L}_{entity}$) (Fang et al., 2019) is only a regularization term and the Eq. (10) is not applied to this term.

**Inference**. As illustrated in Section 3, our goal is to exclude the natural direct effect ($K_1 \rightarrow Y, K_2 \rightarrow Y$) and use the true multi-hop effect ($K \rightarrow Y$) to reduce the multi-hop QA model's disconnected reasoning, so we use Eq. (5) for inference:

$$\mathcal{F}(q, s, c) - \mathcal{C}. \tag{12}$$

The time consumption of our approach is equal to the existing methods.

## 5 Experiments

We extensively conduct the experiments on the HotpotQA (Yang et al., 2018) dataset in the distractor setting. See Appendix A.2 for more experimental results on the other multihop benchmark 2WikiMultihopQA (Ho et al., 2020).

**Metrics and Dataset:** To measure multi-hop QA models, we report two results: 1) *original* denotes the model's original performance (larger is better) and 2) *dire* denotes how much the QA model is cheated (smaller is better). Please note *original*−*dire* denotes the model's true multi-hop reasoning.

A probing dataset should be constructed to measure DiRe. To the best of our knowledge, only (Trivedi et al., 2020) provides the probing dataset

of the development set of HotpotQA. Thus we use the development set of HotpotQA to evaluate the *original* and *dire*. Specifically, the probing dataset for HotpotQA in the distractor setting divides each example of the original dataset into two instances, both of which only contain one of two ground truth supporting paragraphs respectively. If the multi-hop QA model can arrive at the correct test output on two instances, it means that the model performs disconnected reasoning on the original example. Please refer to (Trivedi et al., 2020) for more details.

Following the Dire (Trivedi et al., 2020), we report the metrics for HotpotQA: answer span ($\mathrm{Ans}$), supporting paragraphs ($\mathrm{Supp_p}$), supporting sentences ($\mathrm{Supp_s}$), joint metrics ($\mathrm{Ans}+\mathrm{Supp_p}$, $\mathrm{Ans}+\mathrm{Supp_s}$). We show both EM scores and F1 scores to compare the performance between baselines and our counterfactual multi-hop reasoning method.

**Baselines:** First, we simply use the BERT (Devlin et al., 2018) to predict the answer, supporting sentences and supporting paragraphs as the baseline. $BERT + ours$ denotes that we apply our counterfactual multi-hop reasoning method based on BERT as the backbone. The proposed approach is model-agnostic and we also implement it on several multi-hop QA architectures, including DFGN (Qiu et al., 2019), HGN (Fang et al., 2019) and XLNet in Dire (Trivedi et al., 2020; Yang et al., 2019). Our proposed algorithm also can be implemented on other baselines.

### 5.1 Quantitative Results

For fairness, we conduct the experiments under the same preprocessing of the dataset following HGN (Fang et al., 2019), which select top K relevant paragraphs corresponding to each question example. And the experimental results are shown in Table 1 and Table 2. The main observation can be made as follows:

*Our method can significantly reduce disconnected reasoning.* Compared to BERT baseline, our proposed counterfactual multi-hop reasoning method can reduce the disconnected reasoning of answer prediction and supporting facts identification in both the paragraph level and sentence level. In particular, we can see big drops of 9.5 F1 points on $\mathrm{Supp_s}$ (from 64.1 to 54.6) and 13.6 EM points on $\mathrm{Supp_p}$ (from 17.3 to 3.7) in disconnected reasoning (dire). Our method is better at reducing disconnected reasoning on the Exact Match (EM) evaluation metric. This is because EM is a stricter evaluation metric. For example, EM requires both of the supporting facts should be predicted correctly, while it has F1 scores even when only one supporting fact is predicted correctly. For dire evaluation where only one supporting fact is provided, our approach punishes this situation and achieves lower scores on EM metric of disconnected reasoning. It demonstrates that our method effectively reduces disconnected reasoning when the supporting facts are insufficient.

*Our method still guarantees comparable performance on the original dev set.* As seen from Table 1 and Table 2, the proposed method also maintains the same accuracy on the original set. It even shows a better performance on the supporting facts prediction in the paragraph level.

*Our method is model-agnostic and it demonstrates effectiveness in several multi-hop QA models.* Based on our causal-effect insight, our proposed approach can easily be applied to other multi-hop QA architectures including XLNET, DFGN, HGN (Trivedi et al., 2020; Qiu et al., 2019; Fang et al., 2019). As shown in Table 1 and Table 2, our proposed counterfactual reasoning method achieves better performance. Our method can reduce disconnected reasoning by introducing the proposed counterfactual approach in the training procedure. The dire scores of HGN (Fang et al., 2019) and XLNET (Trivedi et al., 2020) in $\mathrm{Ans}, \mathrm{Supp_p}, \mathrm{Supp_s}$ all drop to some extent. Besides, the performances on the original dev set are comparable simultaneously.

In summary, reducing the disconnected reasoning and guaranteeing the strong performance on the original development set indicate that the most progress of the model is attributed to the multi-top reasoning ($K \rightarrow Y$) capability. For intuitiveness, we also show the real multi-hop reasoning promoted by our proposed counterfactual reasoning approach, as shown in Fig. 2.

### 5.2 Ablation Study

As illustrated in Section 4, our goal is to exclude the shortcut impacts ($K_1 \rightarrow Y, K_2 \rightarrow Y$) to reduce the disconnected reasoning. Hence, we study the ablation experiments by excluding one of the shortcut impacts. We explore removing $K_1 \rightarrow Y$ or $K_2 \rightarrow Y$ that reduces the disconnected reasoning, as shown in Table 3. We can see that excluding one of them can decrease the amount of discon-

| | Ans | | $\text{Supp}_\text{p}$ | | $\text{Supp}_\text{s}$ | | Ans + $\text{Supp}_\text{p}$ | | Ans + $\text{Supp}_\text{s}$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | original | dire↓ | original | dire↓ | original | dire↓ | original | dire↓ | original | dire↓ |
| BERT | 74.2 | 47.5 | 94.1 | 73.9 | 83.8 | 64.1 | 71.0 | 36.5 | 64.2 | 32.5 |
| +$K_1 \rightarrow Y$ | 74.2 | 50.2 | 96.6 | 73.5 | 85.7 | 61.9 | 72.5 | 37.8 | 65.7 | 32.8 |
| +$K_2 \rightarrow Y$ | 74.6 | 48.8 | 96.4 | 68.2 | 85.6 | 57.7 | 72.8 | 34.5 | 66.0 | 30.1 |
| +ours(full) | 74.0 | **45.5** | 95.4 | **67.5** | 82.8 | **54.6** | 71.6 | **31.8** | 63.9 | **26.7** |

Table 3: F1 scores of ablation study on the development set and probing set of HotpotQA in the Distractor setting. The "+ $K_1 \rightarrow Y$" denotes that we only utilize the counterfactual examples $(q, s, c^*)$ to estimate the shortcut impact of $K_1 \rightarrow Y$, and similarly the "+ $K_2 \rightarrow Y$" represents that only the counterfactual examples $(q, s^*, c)$ are used to estimate the shortcut impact of $K_2 \rightarrow Y$.

| | Ans | | $\text{Supp}_\text{p}$ | | $\text{Supp}_\text{s}$ | | Ans + $\text{Supp}_\text{p}$ | | Ans + $\text{Supp}_\text{s}$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | original | dire↓ | original | dire↓ | original | dire↓ | original | dire↓ | original | dire↓ |
| random | **74.0** | **45.5** | **95.4** | 67.5 | **82.8** | 54.6 | **71.6** | 31.8 | **63.9** | 26.7 |
| uniform | 73.8 | 45.9 | 92.1 | **67.4** | 76.6 | **46.3** | 69.2 | 31.8 | 59.6 | **23.2** |

Table 4: F1 scores of ablation study of assumptions for counterfactual outputs $\mathcal{C}$ on the development set and probing set of HotpotQA in the distractor setting.

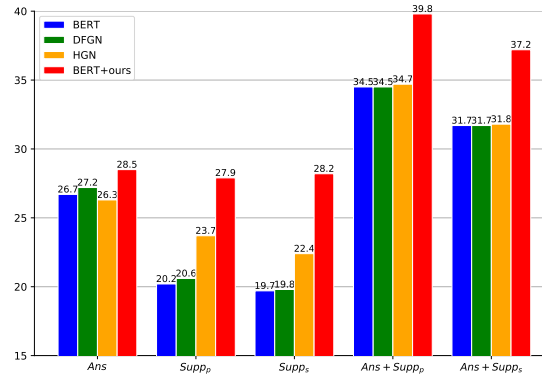


Figure 2: F1 scores of real multi-hop reasoning, which is denoted as the original scores minus the dire scores. We compare BERT, DFGN, HGN, and our method except XLNET, as they utilize the same pre-trained language model (i.e. bert-base-uncased).

nected reasoning to some extent on supporting facts identification except the answer span prediction. However, relieving the both impacts of $K_1 \rightarrow Y$ and $K_2 \rightarrow Y$ can achieve better performance on decreasing disconnected reasoning. Because the model can always exploit another shortcut if only one of the shortcuts is blocked.

We further conduct ablation studies to validate the distribution assumption for the counterfactual output of the parameter $\mathcal{C}$. Similar to CF-VQA (Niu et al., 2021), we empirically validate the two distribution assumptions, as shown in Table 4. The "random" denotes that $\mathcal{C}$ are learned without constraint and it means that $\mathcal{C}_{Ans} \in \mathbb{R}^n, \mathcal{C}_{supp} \in \mathbb{R}^2, \mathcal{C}_{type} \in \mathbb{R}^4$ respectively, and $n$ represents the length of the context. The "uniform" denotes that $\mathcal{C}$ should satisfy uniform distribution and it means that $\mathcal{C}_{Ans}, \mathcal{C}_{supp}$ and $\mathcal{C}_{type}$ are scalar. As shown in Table 4, the random distribution assumption performs better than the uniform distribution assumption.

## 6 Conclusion

In this work, we proposed a novel counterfactual reasoning approach to reduce disconnected reasoning in multi-hop QA. We used the causal graph to explain the existing multi-hop QA approaches' behaviors, which consist of the shortcut impacts and reasoning impacts. The shortcut impacts capture the disconnected reasoning and they are formulated as the natural direct causal effects. Then we constructed the counterfactual examples during the training phase to estimate the both natural direct effects of question and context on answer prediction as well as supporting facts identification. The reasoning impact represents the multi-hop reasoning and is estimated by introducing learnable parameters.

During the test phase, we exclude the natural direct effect and utilize the true multi-hop effect to decrease the disconnected reasoning. Experimental results demonstrate that our proposed counterfactual reasoning method can significantly drop the disconnected reasoning on the probing dataset and guarantee the strong performance on the original dataset, which indicates the most progress of the multi-hop QA model is attributed to the true multi-

hop reasoning. Besides, our approach is model-agnostic, and can be applied to other multi-hop QA architectures to avoid exploiting the shortcuts.

## Limitations

Our proposed method needs to construct counterfactual examples to estimate the natural direct effect of disconnected reasoning during the training phase, thus we need a little more GPU resources and computational time. However, the need of resource occupancy and time consumption of our approach does not increase during inference. Another limitation is that we use the learnable parameters to approximate the $Y_{k_1,k_2,k^*}$. In our future work, we will explore a better approach to model it.

## Acknowledgment

This work is supported by the National Natural Science Foundation of China (U1911203, U2001211, U22B2060), Guangdong Basic and Applied Basic Research Foundation (2019B1515130001, 2021A1515012172), Key-Area Research and Development Program of Guangdong Province (2020B0101100001).

## References

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2019. Learning to retrieve reasoning paths over wikipedia graph for question answering. *arXiv preprint arXiv:1911.10470*.

Jifan Chen, Shih-ting Lin, and Greg Durrett. 2019. Multi-hop question answering via reasoning chains. *arXiv preprint arXiv:1910.02610*.

Rajarshi Das, Ameya Godbole, Dilip Kavarthapu, Zhiyu Gong, Abhishek Singhal, Mo Yu, Xiaoxiao Guo, Tian Gao, Hamed Zamani, Manzil Zaheer, et al. 2019. Multi-step entity-centric information retrieval for multi-hop question answering. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 113–118.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. 2019. Cognitive graph for multi-hop reading comprehension at scale. *arXiv preprint arXiv:1905.05460*.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2019. Hierarchical graph network for multi-hop question answering. *arXiv preprint arXiv:1911.03631*.

Dirk Groeneveld, Tushar Khot, Ashish Sabharwal, et al. 2020. A simple yet strong pipeline for hotpotqa. *arXiv preprint arXiv:2004.06753*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Yichen Jiang and Mohit Bansal. 2019. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop qa. *arXiv preprint arXiv:1906.07132*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Kyungjae Lee, Seung-won Hwang, Sang-eun Han, and Dohyeon Lee. 2021. Robustifying multi-hop qa through pseudo-evidentiality training. *arXiv preprint arXiv:2107.03242*.

Shaobo Li, Xiaoguang Li, Lifeng Shang, Xin Jiang, Qun Liu, Chengjie Sun, Zhenzhou Ji, and Bingquan Liu. 2021. Hopretriever: Retrieve hops over wikipedia to answer complex questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13279–13287.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019a. Compositional questions do not necessitate multi-hop reasoning. *arXiv preprint arXiv:1906.02900*.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019b. Multi-hop reading comprehension through question decomposition and rescoring. *arXiv preprint arXiv:1906.02916*.

Yixin Nie, Songhe Wang, and Mohit Bansal. 2019. Revealing the importance of semantic retrieval for machine reading at scale. *arXiv preprint arXiv:1909.08041*.

Kosuke Nishida, Kyosuke Nishida, Masaaki Nagata, Atsushi Otsuka, Itsumi Saito, Hisako Asano, and Junji Tomita. 2019. Answering while summarizing: Multi-task learning for multi-hop qa with evidence extraction. *arXiv preprint arXiv:1905.08511*.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710.

Judea Pearl. 2022. Direct and indirect effects. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 373–392.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

Peng Qi, Xiaowen Lin, Leo Mehr, Zijian Wang, and Christopher D Manning. 2019. Answering complex open-domain questions through iterative query generation. *arXiv preprint arXiv:1910.07000*.

Lin Qiu, Yunxuan Xiao, Yanru Qu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2019. Dynamically fused graph network for multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6140–6150.

Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. 2021. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1025–1034.

Nan Shao, Yiming Cui, Ting Liu, Shijin Wang, and Guoping Hu. 2020. Is graph structure necessary for multi-hop question answering? *arXiv preprint arXiv:2004.03096*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2020. Is multihop qa in dire condition? measuring and reducing disconnected reasoning. *arXiv preprint arXiv:2005.00789*.

Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. 2021. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3091–3100.

Wenhan Xiong, Xiang Lorraine Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Wen-tau Yih, Sebastian Riedel, Douwe Kiela, et al. 2020. Answering complex open-domain questions with multi-hop dense retrieval. *arXiv preprint arXiv:2009.12756*.

Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. 2021. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9847–9857.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

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

Semih Yavuz, Kazuma Hashimoto, Yingbo Zhou, Nitish Shirish Keskar, and Caiming Xiong. 2022. Modeling multi-hop question answering as single sequence prediction. *arXiv preprint arXiv:2205.09226*.

Xi Ye, Rohan Nair, and Greg Durrett. 2021. Connecting attributions and qa model behavior on realistic counterfactuals. *arXiv preprint arXiv:2104.04515*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.

Chen Zhao, Chenyan Xiong, Corby Rosset, Xia Song, Paul Bennett, and Saurabh Tiwary. 2020. Transformer-xh: Multi-evidence reasoning with extra hop attention.

Beier Zhu, Yulei Niu, Xian-Sheng Hua, and Hanwang Zhang. 2021. Cross-domain empirical risk minimization for unbiased long-tailed classification. *arXiv preprint arXiv:2112.14380*.

# A Appendix

## A.1 Implementation Details

Specifically, given the question $Q = q$, and context $P = \{s_1, s_2, ..., s_m\}$ where $m$ is the number of paragraphs, we denote the remaining context $c_i = P - \{s_i\}, 1 \leq i \leq m$. To distinguish whether $s_i$ is supporting fact and get the answer distribution on $s_i$, we construct the $s_i^*$ and $c_i^*$ as illustrated in the subsection 4.1. We respectively encode $(q, s_i, c_i)$,$(q, s_i^*, c_i)$ and $(q, s_i, c_i^*)$ to get the contextualized representation $O \in R^{n\times d}, M_i \in^{n\times d}$ , $G_i \in R^{n\times d}$, where $n$ is the length of the question and context.

For supporting facts identification, similar to Dire (Trivedi et al., 2020), we use the start token $s_i^{start}$ of the paragraph as its representation and obtain their predicted logits under factual and counterfactual scenario:

$$\begin{aligned} Y_{q,s_i,c_i}(s_i) &= g(O[s_i^{start}]) \\ Y_{q,s_i^*,c_i}(s_i) &= g(M_i[s_i^{start}]) \\ Y_{q,s_i,c_i^*}(s_i) &= g(G_i[s_i^{start}]), \end{aligned} \tag{13}$$

where $g$ is a classifier instantiated as $MLP$ layer in practice. And $s_i^{start}$ and $s_i^{end}$ are denoted as the start and end position of the paragraph $s_i$ respectively. In sentence level, we use the their start positions in paragraph $s_i$ and operate in the same way.

As for answer span prediction, we concatenate the representation of $s_i$ in $M_i$ or $G_i$ $(1 \leq i \leq m)$

| | Ans | | $\text{Supp}_\text{p}$ | | $\text{Supp}_\text{s}$ | | Ans + $\text{Supp}_\text{p}$ | | Ans + $\text{Supp}_\text{s}$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | original | dire↓ | original | dire↓ | original | dire↓ | original | dire↓ | original | dire↓ |
| BERT | 59.7 | 11.6 | 86.4 | 47.5 | 83.1 | 42.4 | 56.1 | 5.0 | 54.5 | 4.6 |
| +ours | 58.9 | **9.8** | 84.6 | **34.4** | 74.4 | **22.7** | 54.1 | **3.3** | 48.7 | **2.3** |
| DFGN | 55.6 | 9.3 | 86.3 | 47.2 | 83.6 | 43.0 | 52.4 | 4.0 | 51.5 | 3.8 |
| +ours | 49.6 | **8.0** | 87.5 | **47.0** | 84.7 | **42.8** | 47.4 | **3.6** | 46.5 | **3.4** |
| HGN | 60.3 | 11.6 | 87.5 | 47.1 | 85.2 | 43.1 | 57.2 | 4.9 | 56.4 | 4.5 |
| +ours | 57.7 | **9.7** | 88.1 | 47.7 | 85.9 | **43.0** | 55.5 | **4.3** | 54.9 | **3.9** |

Table 5: F1 scores on the development set of 2WikiMultihopQA dataset.

to construct the entire answer span prediction on the whole context:

$$\begin{aligned}
&\bar{M} = [M_1[s_1^{start} : s_1^{end}]; ...; M_m[s_m^{start} : s_m^{end}]] \\
&\bar{G} = [G_1[s_1^{start} : s_1^{end}]; ...; G_m[s_m^{start} : s_m^{end}]] \\
&Y_{q,s,c}(start) = g(O) \\
&Y_{q,s^*,c}(start) = g(\bar{M}) \\
&Y_{q,s,c^*}(start) = g(\bar{G}),
\end{aligned} \tag{14}$$

where $[;]$ denotes the operation of concatenation. And we can predict the end position of the answer in the same way.

We also apply our counterfactual reasoning method to identify the answer type, consisting of yes, no, span and entity:

$$\begin{aligned}
&Y_{q,s,c}(type) = g(O[0]) \\
&Y_{q,s^*,c}(type) = \sum_{i=1}^{m} g(M_i[0]) \\
&Y_{q,s,c^*}(type) = \sum_{i=1}^{m} g(G_i[0]).
\end{aligned} \tag{15}$$

We use the $[CLS]$ token as the global representation to predict the answer type, following previous work (Qiu et al., 2019; Fang et al., 2019).

### A.2 Extensive experiments

In this section, we extensively evaluate our method on 2WikiMultihopQA dataset(Ho et al., 2020). And following Dire (Trivedi et al., 2020), we construct the probing dataset of the corresonding development set to evaluate the metric of *dire*. The experimental results are shown in Table 5.

Our proposed counterfactual reasoning approach method can be generalized to other multihopQA benchmarks. Our approach can still significantly reduce the disconnected reasoning, and guanrantee the comparable preformance on the original dev set. Besides, our method is model-agnostic and it demonstrates effectiveness in several multi-hop QA models.

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*7*

☒ A2. Did you discuss any potential risks of your work?
*Left blank.*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
*1*

☒ A4. Have you used AI writing assistants when working on this paper?
*Left blank.*

### B ☒ Did you use or create scientific artifacts?

*Left blank.*

☒ B1. Did you cite the creators of artifacts you used?
*Left blank.*

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*Left blank.*

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*Left blank.*

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*Left blank.*

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*Left blank.*

☒ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*Left blank.*

### C ☒ Did you run computational experiments?

*Left blank.*

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*Left blank.*

---

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Left blank.*

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*Left blank.*

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*Left blank.*

**D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Left blank.*

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*Left blank.*

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*Left blank.*

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*Left blank.*

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*Left blank.*

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*Left blank.*