HYPOTHETICAL TRAINING FOR ROBUST MACHINE READING COMPREHENSION OF TABULAR CONTEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine Reading Comprehension (MRC) models easily learn spurious correlations from complex contexts such as tabular data. Counterfactual training—using the factual and counterfactual data by augmentation—has become a promising solution. However, it is costly to construct faithful counterfactual examples because it is tricky to maintain the consistency and dependency of the tabular data. In this paper, we take a more efficient fashion to ask **hypothetical questions** like *"in which year would the net profit be larger if the revenue in 2019 were $38,298?"*, whose effects on the answers are equivalent to those expensive counterfactual tables. We propose a hypothetical training framework that uses paired examples with different hypothetical questions to supervise the direction of model gradient towards the counterfactual answer change. We construct a new stress test on MRC datasets with factual and hypothetical examples to validate our effectiveness.

1 INTRODUCTION

Machine Reading Comprehension (Dua et al., 2019; Rajpurkar et al., 2016) trains deep models to understand the natural language context by answering questions. However, these deep models easily learn spurious correlations (*a.k.a.* shortcuts) (Ko et al., 2020; McCoy et al., 2019; Yu et al., 2020) between the context and answer, *e.g.,* entries at the first column have higher chance to be chosen as answers in complex financial tables. Consequently, the context understanding is incomplete or even biased, leading to significant performance drop on testing examples without such shortcut (*e.g.,* F1-score drops from 79.4 to 39.2 (*cf.* Table 1) Therefore, it is crucial to resolve the spurious correlation issue in the MRC task with tabular context.

Counterfactual training (Abbasnejad et al., 2020; Teney et al., 2020; Feng et al., 2021; Zhu et al., 2020) is effective for blocking the spurious correlations in various text understanding and reasoning tasks such as Visual Question Answering (Chen et al., 2020a; Niu et al., 2021) and Natural Language Inference (Kaushik et al., 2020). Counterfactual training augments the original *factual* training example with a counterfactual example which minimally modifies the original example's semantic meaning that changes the label, and encourages the model to learn the subtle semantic difference that make the label change—the true causation (Figure 1). The underlying rationale is that if the model only captures the spurious correlation, it cannot comprehend the subtle change from factual to counterfactual, and thus still predicts the original label. For MRC with tabular context, the annotation of counterfactual example is extremely expensive since extra effort is required to maintain the consistency and dependency across table entries when editing the context. As shown in Figure 6, annotators need to edit 4 extra numbers for an assumption to change one number. Therefore, conventional counterfactual generation methods (Yue et al., 2021) will suffer from the fidelity problems (*e.g.,* inconsistent summation) and hurt the model robustness (*cf.* Section 3.3).

In this work, we propose an economic alternative: asking hypothetical questions (HQs) (Li et al., 2022a) by imposing the factual example with a counterfactual assumption, without the cost of maintaining the table consistency and dependency. Therefore, the construction cost of a hypothetical example is undoubtedly lower than the counterfactual example[1]. A hypothetical example consists of a hypothetical question and a factual context, which has the equivalent effect on the answers to the corresponding "ideal" counterfactual example. As a concrete case in Figure 1, the counterfactual

[1]Please refer to Appendix E in the supplementary materials for detailed comparisons.

Context		
Year	2019	2018
Revenue ($)	34,298	37,566
Cost ($)	4,550	6,240
Net Profit ($)	29,748	31,326

Year	2019	2018
Revenue ($)	38,298	37,566
Cost ($)	4,550	6,240
Net Profit ($)	33,748	31,326

Editing the tabular context is costly due to the dependency across table entries .

Question	Answer	Sample Type
In which year was the net profit larger?	2018	Factual example
In which year would the net profit be larger *if the revenue in 2019 were $30,000 instead?*	2018	Hypothetical example
In which year would the net profit be larger *if the revenue in 2019 were $38,298 instead?*	2019	Hypothetical example
In which year was the net profit larger?	2019	Counterfactual example

Figure 1: Illustration of factual, hypothetical, and counterfactual examples.

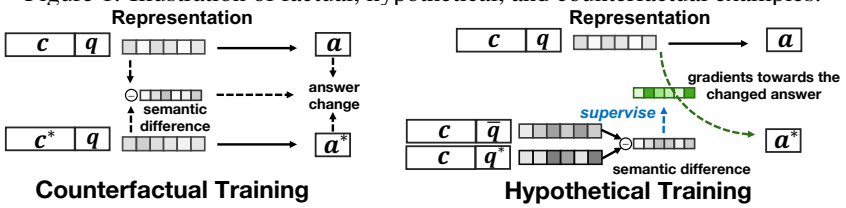


Figure 2: Illustration of counterfactual training and the proposed hypothetical training. $\boldsymbol{c}^*$ denotes the counterfactual context.

example is derived from the factual example according to the assumption *"if the revenue in 2019 were $38,298"*, which changes the answer to *"in which year was the net profit larger"* from *2018* to *2019*. The answer of the corresponding hypothetical question—*"in which year would the net profit be larger if the revenue in 2019 were $38,298?"*—is also *2019*.

Recall that the key to blocking the spurious correlation lies in encouraging the model to focus on the effect of semantic intervention on the answer change. As shown in Figure 2, in conventional counterfactual training, given a factual "context, question, answer" example $(\boldsymbol{c}, \boldsymbol{q}, \boldsymbol{a})$, we utilize a counterfactual example to regularize the learning of the mapping from $\boldsymbol{c}, \boldsymbol{q}$ to $\boldsymbol{a}$ to avoid fitting spurious correlations Teney et al. (2020). In the absence of counterfactual examples, we do the regularization in training by considering the alternative target $\boldsymbol{a}^*$. We intend to investigate and teach the model's understanding on the semantic intervention required for the factual example to change the answer $\boldsymbol{a} \rightarrow \boldsymbol{a}^*$. To obtain the information of such semantic intervention, we use a pair of hypothetical examples with different assumptions and answers $(\boldsymbol{c}, \boldsymbol{q}^*, \boldsymbol{a}^*)$ and $(\boldsymbol{c}, \bar{\boldsymbol{q}}, \boldsymbol{a})$, where the difference in HQ assumptions indicates the semantic intervention to change $\boldsymbol{a}$ to $\boldsymbol{a}^*$ (*cf.* Figure 1). Therefore, our goal becomes how to effectively convey the information of semantic intervention from the hypothetical example pair to the factual example through training.

To incorporate the information of semantic intervention from the hypothetical example pair to model training, we calculate the model gradient *w.r.t.* the input representation of the factual example towards the changed answer $\boldsymbol{a}^*$ [2]. The gradient reflects the model's understanding on the translation direction of the input representation towards the changed answer, *i.e.,* the cause of answer change from $\boldsymbol{a}$ to $\boldsymbol{a}^*$. Therefore, we can guide the model's understanding with the semantic intervention from the hypothetical example pair. We utilize the representation difference between the two hypothetical examples as the reference of semantic intervention, and supervise the model to align the gradient with the representation difference (*cf.* Figure 2). To this end, we propose a *Hypothetical Training Framework* (HTF) that incorporates gradient regulation terms according to hypothetical examples to learn robust MRC models. We apply the HTF framework on a representative tabular MRC model TAGOP (Zhu et al., 2021) and conduct experiments on tabular MRC datasets TAT-QA (Zhu et al., 2021) and TAT-HQA (Li et al., 2022a) with factual examples and hypothetical examples, respectively. Experimental results validate the superior performance of the proposed HTF on a stress test. Further studies show that HTF also has better understanding to various semantic interventions. Code and data will be public upon acceptance.

Our contributions are summarized as follows:

[2]The gradient can be seen as representation changes. It is different from the gradient *w.r.t.* model parameters calculated for updating the model.

- We reveal the spurious correlation issue in MRC of tabular context and propose to use hypothetical examples to economically block spurious correlations and learn robust MRC models.
- We propose the hypothetical training framework, which uses hypothetical example pairs to teach the MRC model the effect of semantic intervention on the answer.
- We apply HTF to the MRC model and conduct experiments on factual and hypothetical MRC datasets, validating the rationality and effectiveness of HTF in blocking spurious correlations.

2 METHOD

2.1 MACHINE READING COMPREHENSION

Generally, the MRC task aims to answer a question based on the context, where the context might be hybrid in complex scenarios, including paragraphs and tables. Formally, given a question $\boldsymbol{q}$, the DNN models are required to reason over the context $\boldsymbol{c}$ and learn a function $g(\boldsymbol{c}, \boldsymbol{q})$ to predict the labeled answer $\boldsymbol{a}$. Technically speaking, the function $g(\cdot)$ is optimized by fitting the correlations from $\boldsymbol{c}$ and $\boldsymbol{q}$ to $\boldsymbol{a}$. However, there widely exist spurious correlations (*a.k.a.* shortcuts (Geirhos et al., 2020)) in the complex context. Learning from such spurious correlations will ignore the features in $\boldsymbol{c}$ and $\boldsymbol{q}$ that causally decide the answers, leading to poor generalization ability.

Counterfactual training. A representative approach to remove spurious correlations is counterfactual training (Abbasnejad et al., 2020), which utilizes counterfactual examples to identify the features that causally affect the answers. As illustrated in Figure 1, counterfactual examples change the answers of factual examples by minimally perturbing the context features, where the perturbation relies on an assumption with semantic intervention, for example, "if the revenue in 2019 were $38,298?" The semantic intervention over factual examples indicates the essential features leading to answer changes. By training over the factual and counterfactual examples, the DNN models are able to learn the effect of the semantic intervention on the answers and exclude the spurious correlations (Teney et al., 2020).

Nevertheless, counterfactual examples are costly to annotate, especially in complex scenarios with hybrid contexts (*e.g.,* tables and paragraphs). As shown in Figure 1, revising the table needs to ensure the consistency and dependency across table entries. The counterfactual table is created based on the assumption "if the revenue in 2019 were $38,298 instead". Without consistency checking, *i.e.,* modifying the net profit of 2019 by "net profit = revenue - cost" , the unfaithful counterfactual table is likely to confuse some questions such as the comparison of net profit. The requirement for consistency checking cannot be easily satisfied by automatic approaches. First, the tables cannot always be processed by relational databases since recent MRC datasets often utilize web-crawled semi-structured tables without clearly defined constraints (Zhu et al., 2021; Zhao et al., 2022; Chen et al., 2021). Second, some conventional counterfactual generation methods such as Yue et al. (2021); Pasupat & Liang (2016) also cannot guarantee the fidelity of counterfactual examples.

Hypothetical example. To alleviate the burden of consistency checking, we propose hypothetical examples as the alternative of counterfactual examples. Hypothetical example appends an assumption to the question of factual example, where the assumption describes the semantic intervention over the factual context, causing the same answer change as the counterfactual example. For instance, in Figure 1, the assumption "*if the revenue in 2019 were $38,298 instead?*" summarizes the changes in the table of the counterfactual example. Compared to editing the complex table with dependency requirements, it is cost-friendly to construct hypothetical examples by extending the questions in natural language (refer to Appendix E for more comparison).

2.2 HYPOTHETICAL TRAINING

To remove the spurious correlations, the key lies in capturing the semantic intervention leading to answer changes. To this end, HTF calculates the semantic differences between a pair of hypothetical examples with distinct answers, and then pushes the MRC models to learn the effect of such semantic differences. Specifically, given a pair of hypothetical examples $(\boldsymbol{c}, \bar{\boldsymbol{q}}, \boldsymbol{a})$ and $(\boldsymbol{c}, \boldsymbol{q}^*, \boldsymbol{a}^*)$, we first calculate their representation differences, and then utilize the differences to regulate the gradients of factual example towards the changed answer. Intuitively, the representation differences reflect the semantic intervention, and the gradients indicates how the representations change can lead to

changed answers. The alignment between representation differences and gradients reflects whether the MRC models capture semantic intervention well.

As illustrated in Figure 2, given a pair of hypothetical examples $(\boldsymbol{c}, \bar{\boldsymbol{q}}, \boldsymbol{a})$ and $(\boldsymbol{c}, \boldsymbol{q}^*, \boldsymbol{a}^*)$, the MRC model first encodes the context-question pairs $(\boldsymbol{c}, \bar{\boldsymbol{q}})$ and $(\boldsymbol{c}, \boldsymbol{q}^*)$ into the representations $\bar{\boldsymbol{X}}_h$ and $\boldsymbol{X}_h^*$ via feature extractors (*e.g.,* Pre-trained Language Model (PrLM) (Liu et al., 2019)), respectively. We then calculate $\boldsymbol{X}_h^* - \bar{\boldsymbol{X}}_h$ as the semantic differences, which cause answer changing from $\boldsymbol{a}$ to $\boldsymbol{a}^*$.

For the normal training of a factual example $(\boldsymbol{c}, \boldsymbol{q}, \boldsymbol{a})$, the MRC model encodes the context-question pair $(\boldsymbol{c}, \boldsymbol{q})$ into the representation $\boldsymbol{X}_f$, and then leverages a function $f(\boldsymbol{X}_f)$ to predict the answers $\boldsymbol{a}$. To inspect whether the MRC model captures the semantic differences, we calculate the gradients *w.r.t.* the representation $\boldsymbol{X}_f$ towards the changed answer $\boldsymbol{a}^*$, *i.e.,* $\nabla^T f_{\boldsymbol{a}^*}(\boldsymbol{X}_f)$. Such gradients represents the translation direction of the representation $\boldsymbol{X}_f$ that can change the answer from $\boldsymbol{a}$ to $\boldsymbol{a}^*$. As such, we can teach the model to learn the semantic differences by encouraging these gradients to align with $\boldsymbol{X}_h^* - \bar{\boldsymbol{X}}_h$. Formally, we propose a regularization term to minimize their cosine distance as follows:

$$\mathcal{L}_f = 1 - \cos\left(\nabla^T f_{\boldsymbol{a}^*}(\boldsymbol{X}_f), \boldsymbol{X}_h^* - \bar{\boldsymbol{X}}_h\right). \tag{1}$$

Similarly, we have the representation $\boldsymbol{X}_h^*$ of the hypothetical example $(\boldsymbol{c}, \boldsymbol{q}^*, \boldsymbol{a}^*)$. We also regulate the gradients of the hypothetical example towards the changed answer $\boldsymbol{a}$[3], *i.e.,* $\nabla^T f_{\boldsymbol{a}}(\boldsymbol{X}_h^*)$, which describes how $\boldsymbol{X}_h^*$ changes can vary the answer from $\boldsymbol{a}^*$ to $\boldsymbol{a}$. As compared to the gradients of the factual example, the gradients of this hypothetical example conversely change the answer from $\boldsymbol{a}^*$ to $\boldsymbol{a}$. Therefore, $\nabla^T f_{\boldsymbol{a}}(\boldsymbol{X}_h^*)$ should be regulated in the opposite direction with $\nabla^T f_{\boldsymbol{a}^*}(\boldsymbol{X}_f)$:

$$\mathcal{L}_h = 1 - \cos\left(\nabla^T f_{\boldsymbol{a}}(\boldsymbol{X}_h^*), \bar{\boldsymbol{X}}_h - \boldsymbol{X}_h^*\right). \tag{2}$$

2.3 Instantiation

We adopt TAGOP (Zhu et al., 2021) as our backbone MRC model in HTF, which is designed to reason on the tabular and textual context. Powered by PrLM (Liu et al., 2019), TAGOP first flattens the tables in $\boldsymbol{c}$ by row, and then transforms the concatenated $\boldsymbol{c}$ and $\boldsymbol{q}$ into the representation, denoted as $\boldsymbol{X} \in \mathbb{R}^{L \times D}$, where L is the number of the tokens in $\boldsymbol{c}$ and $\boldsymbol{q}$, and D is the representation dimension. Thereafter, TAGOP utilizes sequence tagging to select the answer span(s) from the context, which transforms $\boldsymbol{X}$ through a 2-layer Feed-Forward Network (FFN) followed by softmax to predict the positive or negative label for each token in the context. Formally,

$$\begin{cases} \boldsymbol{p_i} = \text{softmax}\left(\text{FFN}(\boldsymbol{X_i})\right), i = 1, \ldots, N \\ t_i = \arg\max(\boldsymbol{p_i}), \end{cases} \tag{3}$$

where N is the context length since the answer is from the the context region of the input. $\boldsymbol{p_i} \in \mathbb{R}^2$ represents the positive and negative probabilities of the i-th token in the context, and $t_i \in \{0, 1\}$ denotes the final predicted label.

TAGOP adopts an answer-type predictor to decide selecting one or multiple entries and words from the context, or counting the number of positive entries and words (Zhu et al., 2021). The loss function $\mathcal{L}_t$ of TAGOP is the sum of 1) the negative log-likelihood loss for tagging and 2) the cross-entropy loss of the answer-type predictor. In this work, we additionally consider the two regularization terms for hypothetical training, and the overall loss function is as follows:

$$\mathcal{L}_t + \alpha \mathcal{L}_f + \beta \mathcal{L}_h, \tag{4}$$

where α and β control the influence of the two regularization terms on the optimization.

2.4 Theoretical Justification

In this section, we explain the rationality of regularizing the model gradients by the representation differences between a pair of hypothetical examples $(\boldsymbol{c}, \bar{\boldsymbol{q}}, \boldsymbol{a})$ and $(\boldsymbol{c}, \boldsymbol{q}^*, \boldsymbol{a}^*)$. Given their representations $\bar{\boldsymbol{X}}_h$ and $\boldsymbol{X}_h^*$, the MRC model adopts the function $f(\cdot) : \mathbb{R}^{L \times D} \rightarrow \mathbb{R}^N$ to output their

[3]We ignore the regularization over $(\boldsymbol{c}, \bar{\boldsymbol{q}}, \boldsymbol{a})$ and only regulate $(\boldsymbol{c}, \boldsymbol{q}^*, \boldsymbol{a}^*)$ because the former has the same context-question semantic and answer with the factual example $(\boldsymbol{c}, \boldsymbol{q}, \boldsymbol{a})$.

N-dimension logits over N context tokens. We then consider the Taylor Expansion of $f(\boldsymbol{X}_h^*)$ regarding $\bar{\boldsymbol{X}}_h$:

$$\begin{cases} f(\boldsymbol{X}_h^*) = f(\bar{\boldsymbol{X}}_h) + \boldsymbol{J} \cdot (\boldsymbol{X}_h^* - \bar{\boldsymbol{X}}_h) + o(\boldsymbol{X}_h^* - \bar{\boldsymbol{X}}_h), \\ \boldsymbol{J} = \begin{bmatrix} \nabla^T f_1(\bar{\boldsymbol{X}}_h) \\ \cdots \\ \nabla^T f_N(\bar{\boldsymbol{X}}_h) \end{bmatrix}, \end{cases} \tag{5}$$

where $o(\cdot)$ denotes the Taylor Remainder and $\boldsymbol{J} \in \mathbb{R}^{N \times M}$ is the Jacobian Matrix. $M = L \times D$ is the dimension of the representation $\bar{\boldsymbol{X}}_h$. The i-th row in $\boldsymbol{J}$ represents the gradients from the positive logits of the i-th token $f_i(\bar{\boldsymbol{X}}_h)$ to the input representation $\bar{\boldsymbol{X}}_h$. Besides, since the assumptions minimally do intervention to the factual example, we assume that the representations of $\bar{\boldsymbol{X}}_h$ and $\boldsymbol{X}_h^*$ are close to each other. Therefore, the representation differences between $\boldsymbol{X}_h^*$ and $\bar{\boldsymbol{X}}_h$ are small, and $(\boldsymbol{X}_h^* - \bar{\boldsymbol{X}}_h)^K$ will be close to zero when $K > 1$ Teney et al. (2020). In this light, we ignore higher order terms in $o(\boldsymbol{X}_h^* - \bar{\boldsymbol{X}}_h)$ and mainly focus on the first order term $\boldsymbol{J}(\boldsymbol{X}_h^* - \bar{\boldsymbol{X}}_h)$.

To remove spurious correlations, $f(\cdot)$ is expected to learn the effect of the slight representation differences on the answer changes. Given different input representations $\boldsymbol{X}_h^*$ and $\bar{\boldsymbol{X}}_h$, $f(\cdot)$ should be able to maximize the answer prediction difference, *i.e.,* , the logit difference $f(\boldsymbol{X}_h^*) - f(\bar{\boldsymbol{X}}_h)$ over the ground-truth tokens in the answer $\boldsymbol{a}^*$. From Equation (5), we have

$$f_{\boldsymbol{a}^*}(\boldsymbol{X}_h^*) - f_{\boldsymbol{a}^*}(\bar{\boldsymbol{X}}_h) \approx \nabla^T f_{\boldsymbol{a}^*}(\bar{\boldsymbol{X}}_h) \cdot (\boldsymbol{X}_h^* - \bar{\boldsymbol{X}}_h), \tag{6}$$

where $f_{\boldsymbol{a}^*}(\boldsymbol{X}_h^*)$ and $f_{\boldsymbol{a}^*}(\bar{\boldsymbol{X}}_h)$ are the predicted logits for the tokens in the answer $\boldsymbol{a}^*$, and $\nabla^T f_{\boldsymbol{a}^*}(\bar{\boldsymbol{X}}_h)$ in $\boldsymbol{J}$ refers to the gradients for $\boldsymbol{a}^*$. From Equation 6, we can maximize the logit difference by increasing the dot product $\nabla^T f_{\boldsymbol{a}^*}(\bar{\boldsymbol{X}}_h) \cdot (\boldsymbol{X}_h^* - \bar{\boldsymbol{X}}_h)$. However, optimizing via dot product is norm-sensitive so that the function $f(\cdot)$ is easy to increase the norm of gradients but ignore the directions. As such, we choose to minimize the cosine distance in the implementation. The empirical results in Section 3.3 also validate the superiority of using cosine distance.

From the above analysis, we can minimize the cosine distance between the gradients $\nabla^T f_{\boldsymbol{a}^*}(\bar{\boldsymbol{X}}_h)$ of the hypothetical example $(\boldsymbol{c}, \bar{\boldsymbol{q}}, \boldsymbol{a})$ and the representation difference $\boldsymbol{X}_h^* - \bar{\boldsymbol{X}}_h$. Because the factual example $(\boldsymbol{c}, \boldsymbol{q}, \boldsymbol{a})$ and the hypothetical example $(\boldsymbol{c}, \bar{\boldsymbol{q}}, \boldsymbol{a})$ have the same answer under the same context and answering logic, we can again adopt $\boldsymbol{X}_h^* - \bar{\boldsymbol{X}}_h$ to regulate the gradients of this factual example $\nabla^T f_{\boldsymbol{a}^*}(\boldsymbol{X}_f)$ (Equation 1). Meanwhile, based on the similar Taylor Expansion for $f(\bar{\boldsymbol{X}}_h)$, it is reasonable to constrain the gradients of another hypothetical example $(\boldsymbol{c}, \boldsymbol{q}^*, \boldsymbol{a}^*)$, *i.e.,* $\nabla^T f_{\boldsymbol{a}}(\boldsymbol{X}_h^*)$ by $\bar{\boldsymbol{X}}_h - \boldsymbol{X}_h^*$ (Equation 2).

3 Experiments

In this section, we conduct experiments to answer the following research questions: **RQ1:** How does the proposed HTF perform on removing spurious correlations? **RQ2:** How do the regularization terms of HTF influence its effectiveness? **RQ3:** How does HTF improve the MRC model regarding different spurious correlations?

3.1 Experimental Setup

Datasets. We conduct experiments on TAT-QA (Zhu et al., 2021), a MRC dataset in the financial domain with a hybrid of text and tabular context, and TAT-HQA (Li et al., 2022a), which constructs hypothetical questions for TAT-QA. To avoid the complexity of discrete numerical calculation, we filter out the arithmetic questions and only keep the questions that extract text spans. Note that TAT-HQA only contains one hypothetical example with a different answer from the corresponding factual example in TAT-QA. We thus expand the TAT-HQA dataset by adding hypothetical examples with the same answer as the factual example [4]. For the evaluation, on one hand, we adopt the test set of TAT-QA with factual examples. On the other hand, we create a stress test by manually editing the factual examples to break the spurious correlations[5] to evaluate the effectiveness of blocking spurious correlations. We adopt the two common evaluation metrics for MRC tasks (Dua et al., 2019), exact-match (EM) and F_1, both in the range of [0, 100].

[4]Please refer to more details in Appendix B

[5]Please refer to Appendix A for the detailed construction.

	Stress Test		TAT-QA		Average	
	EM	F1	EM	F1	EM	F1
m-OQ	31.4	39.2	**65.2**	**79.4**	48.3	59.3
m-OQ&HQ	34.2	43.7	64.6	78.4	49.4	61.1
m-OQ&2HQ	37.0	44.0	64.4	78.4	50.7	61.2
CF-VQA	25.6	33.3	60.1	74.6	42.9	54.0
xERM	25.6	36.0	60.8	75.0	43.2	55.5
CLO	34.8	43.3	64.5	78.9	49.7	61.1
GS	33.6	42.5	62.3	77.2	48.0	59.9
HTF (Ours)	**38.5**	**45.8**	64.3	78.0	**51.4**	**61.9**

Table 1: Performance comparison on the stress test and original test of TAT-QA *w.r.t.* EM and F_1 scores. Bold font denotes the best performance in each column.

Compared methods. We compare HTF with the following methods. **1)** Vanilla baselines: **m-OQ** trains the MRC model with the factual examples in TAT-QA, *i.e.,* the model learns to answer the original question (OQ); **m-OQ&HQ** trains the model with a mixture of OQs in TAT-QA and HQs in TAT-HQA, which is a simple data augmentation without consideration of the relation between question pairs; and similarly **m-OQ&2HQ** trains the model with a mixture of OQs and two kinds of HQs. **2)** Debiasing methods to mitigate the bias from the context branch: **CF-VQA** (Niu et al., 2021) utilizes a counterfactual inference framework to mitigate the bias; **xERM** (Zhu et al., 2022) improves CF-VQA by adjusting the factual and counterfactual models with the weights of their empirical risks. **3)** Counterfactual training methods: **CLO** (Liang et al., 2020) adopts a contrastive learning objective to supervise the relationship between the factual and hypothetical examples; **GS** (Teney et al., 2020) applies gradient supervision between factual and hypothetical example pairs to shape the decision boundary. More implementation details can be found in Appendix D.

Implementation detail. For all compared methods, we adopt TAGOP (Zhu et al., 2021) as the backend model, which is a representative MRC model on tabular context; and we select hyperparameters according to the F_1 score on the validation set. We apply a two-staged training for HTF by first training on TAT-QA and TAT-HQA with TAGOP loss $\mathcal{L}_t$, and then fine-tuning on the triplets of a factual and two hypothetical examples with HTF regularization terms $\mathcal{L}_f$ and $\mathcal{L}_h$. The reason for two-staged training is that the gradients at the initial training stage cannot stably reflect the model's perception of how the representations change causing the answer change, thus we apply the gradient regularization terms in the fine-tuning stage. We set α as 0.07 and β as 1.3. The detailed hyperparameter setting can be found in the Appendix C.

3.2 Performance Comparison (RQ1)

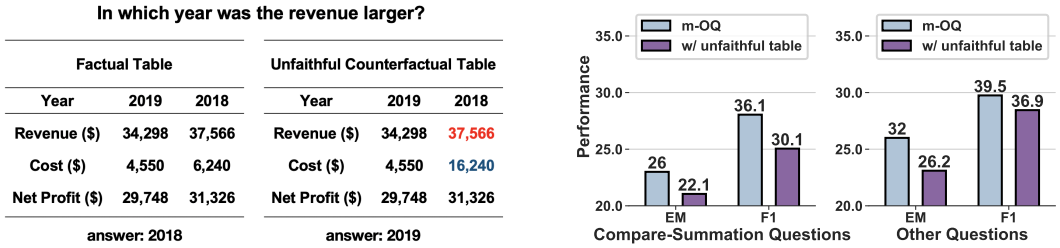
Table 1 shows the performance of all compared methods on both the stress test and TAT-QA. We can observe that: **1)** In all cases, the performance on TAT-QA is much higher than that on the stress test, showing the reliance on spurious correlations of compared methods. **2)** The proposed HTF outperforms all compared methods on the stress test, indicating its least reliance on spurious correlations. Moreover, HTF achieves the best average performance on the stress test and TAT-QA, implying its strong generalization ability across different distributions. **3)** The superior performance of HTF than m-OQ&2HQ validates the rationality of considering the relationships between factual and hypothetical examples via hypothetical training. **4)** The methods utilizing hypothetical examples (m-OQ&HQ, m-OQ&2HQ, CLO, GS, and HTF) generally show better performance on the stress test than the models trained with only factual examples (m-OQ, CF-VQA, xERM). This verifies the rationality of using hypothetical examples to mitigate spurious correlations. Especially, the debiasing methods (CF-VQA and xERM) are not effective in our task. We postulate that blindly mitigating the context bias without the guidance of hypothetical examples is insufficient to remove spurious correlations in the complex reasoning tasks with tables. **5)** All methods using hypothetical examples have inferior performance than m-OQ on TAT-QA, which is possibly attributed to that extra hypothetical examples have different distributions with TAT-QA. It is also a promising future direction to further balance the trade-off and achieve "both-good" performance on two test sets.

3.3 Ablation Studies (RQ2)

Ablation study of HTF regularization. We reveal the contribution of each gradient regularization terms $\mathcal{L}_f$ and $\mathcal{L}_h$ by the ablation experiments w/o $\mathcal{L}_f$ and w/o $\mathcal{L}_h$. As shown in Table 2, we

	w/o $\mathcal{L}_f$#	w/o $\mathcal{L}_h$#	$\mathcal{L}_{\text{dot}}$	GS	$\mathcal{L}_{\text{gs\_var}}$	HTF
EM	38.1	37.3	1.5	33.6	36.7	**38.5**
F1	44.9	45.0	14.8	42.5	44.6	**45.8**

Table 2: Results of the HTF variants on the stress test. # denotes significantly different from the non-ablated HTF ($p < 0.05$).



(a) (b)

Figure 3: Left: an example of the unfaithful counterfactual table. Right: performance comparison on the stress test by adding unfaithful counterfactual tables.

observe that the performance significantly drops if we set either $\mathcal{L}_f$ or $\mathcal{L}_h$ to be 0. This validates that both gradient regularization terms are critical to remove spurious correlations and enhance the generalization performance on the stress test.

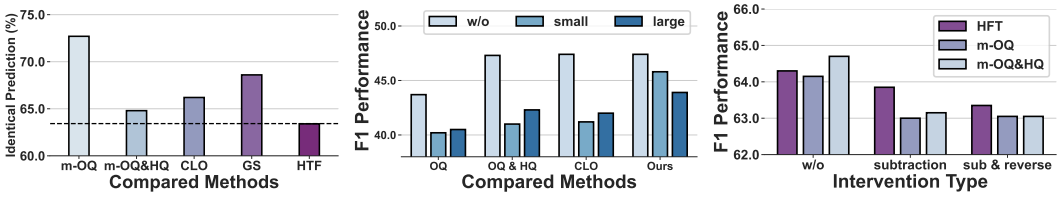
Rationality of using cosine regularization. As illustrated in Section 2.4, we compare the regularization terms implemented by dot product or cosine distance. From the results in Table 2, we find that the dot product $\mathcal{L}_{\text{dot}}$ largely underperforms HTF with cosine regularization. We attribute the significant difference to that dot product is norm-sensitive, for which the gradient norm is easily increased while the direction is undermined.

Difference with GS. In our justification in Section 2.4, we reach a different conclusion from GS (Teney et al., 2020) that the gradient loss should be calculated towards the changed label instead of the factual label. We run a variant of GS by calculating the gradient towards the changed label instead of the factual label to examine our justification, denoted as $\mathcal{L}_{\text{gs\_var}}$. In Table 2, we can find that $\mathcal{L}_{\text{gs\_var}}$ clearly outperforms GS, thus validating the rationality of our justification.

Effect of unfaithful counterfactual tables. To validate our claim that counterfactual tables without consistency checking potentially hinder the answer prediction, we conduct the experiments with unfaithful counterfactual tables. We create unfaithful counterfactual tables by revising the factual tables while ignoring the dependency between table entries. For example, in Figure 3a, the counterfactual table is edited from the factual table under the assumption "*if the cost for 2018 increased to $16,240 instead*". Due to "revenue=cost+net profit", only editing cost will cause inconsistency between the table entries, leading to unfaithful counterfactual tables. If such unfaithful examples in Figure 3a are used for training with factual examples, the MRC model will wrongly attribute the answer changes to the changed cost feature, fitting the spurious correlations. To validate that, we hand-annotate 220 unfaithful counterfactual examples, then train a variant of m-OQ by adding the unfaithful counterfactual tables into training data, and finally test it on the stress test. From the results in Figure 3b, we discover that for both the summation comparison questions (about 10%) and the other questions, the performance has a clear drop, showing that the noisy unfaithful counterfactual tables may confuse the model and it is necessary to guarantee the table consistency.

3.4 In-depth Analysis (RQ3)

Sensitivity to semantic intervention. We investigate the sensitivity of the MRC model to semantic intervention by counting the identical predictions on factual examples and hypothetical examples in the stress test. Since the hypothetical examples in the stress test are created by minimally modifying the factual examples to change its semantic and labels (an example in Figure 5), fitting spurious correlations will cause the model to ignore the semantic difference and give the same prediction for a pair of hypothetical and factual examples. The experimental statistics in Figure 4a shows that: 1) among all compared methods, HTF achieves the smallest percentage of identical predictions, verifying its effectiveness in identifying semantic difference and avoiding spurious correlations. 2) Using hypothetical examples (all methods except m-OQ) can reduce the percentage of identical



(a) (b) (c)

Figure 4: (a) Proportion of identical predictions for factual examples in TAT-QA and the corresponding stress test examples. Lower percentage is better, indicating more sensitive to semantic intervention. (b) Performance on the stress test with changed number scale. *w/o*, *small* and *large* refers to no scaling, slight scaling and large scaling, respectively. (c) Study on the spurious correlation of "which year" questions and the top year answers. "w/o" denotes no operation on the table.

(a) In which years did the net sales from America exceed $200,000?

Factual Table			Stress Test Table		
Year	2018	2017	Year	2018	2017
Net Sale in America ($)	259,105	224,056	Net Sale in America ($)	259,105	150,000

Gold Answer: 2018, 2017
Predicted Answer: 2018, 2017
Prediction Score:
2018: 99.94, 2017: 99.90

Gold Answer: 2018
Predicted Answer: 2018
Prediction Score:
2018: 99.93, 2017: 3.81

(b) In which year was the Deferred tax asset larger?

Factual Table			Stress Test Table		
Year	2019	2018	Year	2019	2018
Deferred tax asset	1.2	0.8	Deferred tax asset	0.2	0.8

Gold Answer: 2019
Predicted Answer: 2019
Prediction Score :
2019: 99.90

Gold Answer: 2018
Predicted Answer: 2019
Prediction Score:
2019: 99.92, 2018: 0.00

Figure 5: Case study of HTF's predictions. The tables are shortened to save space.

predictions, showing the effect of hypothetical examples in reducing shortcuts. 3) For all methods, there are still more than 60% of identical predictions, showing that all methods are still relying on spurious correlations to some extent. This coheres with the large performance gap between stress test and factual examples (TAT-QA) (*cf.* Table 1), which can be further improved in future work.

Generalizing to new semantic intervention. Apart from the stress test, we study the ability of HTF to generalize to new semantic intervention on the tables. Firstly, we look into how HTF generalizes to new tables with the **numbers of unusual scale**. We identify a type of questions asking about numerical conditions, *e.g., which values is larger/smaller than a threshold A?*, and generate new test cases by scaling the target numbers to be larger/smaller than A in the table. We increase the target number by $\{2,3,4\}$ times if it is larger than A and otherwise decrease it by $\{1.2,1.3,1.5\}$ times (named slightly-scaled test data). Besides, we construct the largely-scaled test data by increasing the range of scaling by $\{10,11,12\}$ times or decreasing $\{5,6,7\}$ times.[6] We test HTF, CLO, and the vanilla baselines m-OQ and m-OQ&HQ on the slightly-scaled and largely-scaled test data. As shown in Figure 4b, we find that all methods are affected by the scaling operation because they do not fully understand actual reasoning logic and rely on some spurious correlations. Among the methods, HTF achieves the smallest performance drop between the factual examples and the scaled examples for both settings, showing that HTF achieves the best understanding on the reasoning logic of numerical condition questions by hypothetical training.

Next, we study the potential spurious correlations regarding **the frequent answers** in the dataset. We conjecture that the MRC model might be inclined to predict 2019, 2018 and 2017 for questions asking about "which year" because they are the most frequently appeared answers. We identify such questions and create new testing examples by replacing 2019, 2018, and 2017 with 1994, 1993, and 1992, respectively (denoted as subtraction) to break down the word correlations between the questions and the answers. We also try reversing the order of the years by replacing 2019, 2018, and 2017 with 1992, 1993, and 1994, respectively (denoted as sub&reverse) to examine the bias toward predicting the earliest or the latest year. As shown in Figure 4c, we can observe that the subtraction decreases the performance for all compared methods, revealing the existence of spurious word correlations, while HTF achieves the smallest decrease thus the least affected by such spurious correlation. Applying the reversion can further decrease the performance of HTF, but it is still the least affected among the compared methods. For m-HQ and m-OQ&HQ, the further reversion cannot cause larger decrease, thus they mainly rely on the word correlation. Generally, HTF is the most robust to such semantic change with less than 3% decrease of F_1.

Case study. We present two examples to demonstrate the effect of HTF on model prediction in Figure 5. In example (a), HTF gives correct predictions to both the factual and the stress test examples.

[6]Note that the edited example maintains the answer of the original example.

This indicates that HTF recognizes the semantic change, *i.e.,* the lowered net sale value in 2017, and in turn largely reduces the model prediction score *w.r.t.* 2017. It maintains high prediction scores for the remaining answer and precisely reduces the score for the changed answer, showing the capability of HTF in linking the semantic intervention to the answer change. We also present a failure case in example (b), where HTF gives correct prediction to the factual example, but fails on the stress test example due to failure to link the feature change *i.e.,* the decreased value in 2019, with the answer change. Since the stress test example only has a very tiny change of one digit ($1.2 \rightarrow 0.2$), it poses a larger challenge to HTF in its sensitivity of semantic change.

4 Related Work

Counterfactual training. Stemming from the causal theory (Pearl et al., 2000), counterfactual training has become a popular approach recently to avoid learning spurious correlation by doing intervention on the observed data. Counterfactual examples have been applied to a wide range of task such as Natural Language Inference (Kaushik et al., 2020), Named Entity Recognition (Zeng et al., 2020), Visual Question Answering (Chen et al., 2020a; Gokhale et al., 2020; Teney et al., 2020; Liang et al., 2020), Story Generation (Qin et al., 2019), Machine Reading Comprehension (Gardner et al., 2020), text classification Choi et al. (2022), language representation Feder et al. (2021) and information extraction Nan et al. (2021). Researchers also apply the idea of counterfactual into designing training or inference frameworks (Niu et al., 2021; Niu & Zhang, 2021; Chen et al., 2020a; Wang et al., 2021b; Feng et al., 2021; Abbasnejad et al., 2020; Paranjape et al., 2022). Apart from obtaining counterfactual examples via human-annotation, researcher also study automatically generating counterfactual examples (Paranjape et al., 2022; Geva et al., 2022; Ye et al., 2021; Longpre et al., 2021; Wu et al., 2021; Sauer & Geiger, 2021). In this paper, we focus on tabular MRC with complex reasoning process. Automatically creating counterfactual examples is infeasible and human knowledge is still essential. We are inspired by the hypothetical questions proposed in (Li et al., 2022a) which we think can be an economic alternative for counterfactual tables, and we are the first to study removing spurious correlations with hypothetical examples.

Spurious correlation. The problem of spurious correlation has been studied by a wide range of machine learning tasks, such as the unimodal bias in VQA (Cadene et al., 2019), the position bias of MRC (Ko et al., 2020), the hypothesis-only of NLI (Poliak et al., 2018), the word alignment of passage and options in QA (Yu et al., 2020), which hinders the generalization ability of DNN models to out-of-distribution test sets (Agrawal et al., 2018; Kaushik et al., 2020). Solutions have been propose to solve the spurious correlation problems apart from the counterfactual training approaches mentioned above, such as capturing the bias via fitting the bias (He et al., 2019; Cadene et al., 2019), training multiple models (Teney et al., 2022; Clark et al., 2019), invariant learning (Arjovsky et al., 2019; Li et al., 2022b), and using causal inference techniques (Wang et al., 2021c;a).

Tabular MRC. In recent years, new challenge in MRC has arisen to enable machines to understand and reason over more complex context such as tables due to the overwhelming tabular data in the real world. Many tabular QA dataset are proposed, such as FinQA (Chen et al., 2021), TAT-QA (Zhu et al., 2021), HybridQA (Chen et al., 2020b), MultiHierrt (Zhao et al., 2022), and WikiTableQuestions (Pasupat & Liang, 2015), where most of these datasets requires numerical reasoning ability, thus solutions to these data often requires designing numerical calculation steps (Chen et al., 2021; Zhu et al., 2021) and table understanding techniques (Herzig et al., 2020). In our work, we adopt the standard method of TAGOP on TAT-QA dataset.

5 Conclusion

In this work, we investigated the spurious correlations in MRC with tabular context. We proposed to use hypothetical examples for hypothetical training, which teaches the MRC model the effect of the semantic intervention on causing answer changes. By learning such effect, MRC models could effectively remove the spurious correlations and achieve superior performance on the stress test. This work leaves many promising directions for future exploration: 1) adopting HTF to other language understanding and reasoning tasks that are costly to construct counterfactual examples; 2) expanding HTF to model the semantic relationships between multiple hypothetical examples; and 3) simultaneously pursuing "both good" performance on TAT-QA and the stress test.

REFERENCES

Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. Counterfactual vision and language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10044–10054, 2020.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980, 2018.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in Neural Information Processing Systems*, 32, 2019.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10800–10809, 2020a.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1026–1036, 2020b.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, et al. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3697–3711, 2021.

Seungtaek Choi, Myeongho Jeong, Hojae Han, and Seung-won Hwang. C2l: Causally contrastive learning for robust text classification. 2022.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4069–4082, 2019.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2368–2378, 2019.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386, 2021.

Fuli Feng, Jizhi Zhang, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. Empowering language understanding with counterfactual reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2226–2236, 2021.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Mor Geva, Tomer Wolfson, and Jonathan Berant. Break, perturb, build: Automatic perturbation of reasoning paths through question decomposition. *Transactions of the Association for Computational Linguistics*, 10:111–126, 2022.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Mutant: A training paradigm for out-of-distribution generalization in visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 878–892, 2020.

He He, Sheng Zha, and Haohan Wang. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pp. 132–142, 2019.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Mueller, Francesco Piccinno, and Julian Eisenschlos. Tapas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4320–4333, 2020.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. Look at the first sentence: Position bias in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1109–1121, 2020.

Moxin Li, Fuli Feng, Hanwang Zhang, Xiangnan He, Fengbin Zhu, and Tat-Seng Chua. Learning to imagine: Integrating counterfactual thinking in neural discrete reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 57–69, 2022a.

Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2928–2937, 2022b.

Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3285–3292, 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7052–7063, 2021.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, 2019.

Guoshun Nan, Jiaqi Zeng, Rui Qiao, Zhijiang Guo, and Wei Lu. Uncovering main causalities for long-tailed information extraction. *arXiv preprint arXiv:2109.05213*, 2021.

Yulei Niu and Hanwang Zhang. Introspective distillation for robust question answering. *Advances in Neural Information Processing Systems*, 34:16292–16304, 2021.

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

Bhargavi Paranjape, Matthew Lamm, and Ian Tenney. Retrieval-guided counterfactual generation for qa. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1670–1686, 2022.

Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1470–1480, 2015.

Panupong Pasupat and Percy Liang. Inferring logical forms from denotations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23–32, August 2016.

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2), 2000.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 180–191, 2018.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. Counterfactual story reasoning and generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5043–5053, 2019.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2383–2392, 2016.

Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *International Conference on Learning Representations*, 2021.

Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. Learning what makes a difference from counterfactual examples and gradient supervision. In *European Conference on Computer Vision*, pp. 580–599, 2020.

Damien Teney, Ehsan Abbasnejad, Simon Lucey, and Anton van den Hengel. Evading the simplicity bias: Training a diverse set of models discovers solutions with superior ood generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16761–16772, 2022.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. In *Advances in Neural Information Processing Systems*, 2021.

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

Wei Wang, Boxin Wang, Ning Shi, Jinfeng Li, Bingyu Zhu, Xiangyu Liu, and Rong Zhang. Counterfactual adversarial learning with representation interpolation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4809–4820, 2021b.

Wenjie Wang, Fuli Feng, Xiangnan He, Xiang Wang, and Tat-Seng Chua. Deconfounded recommendation for alleviating bias amplification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1717–1725, 2021c.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6707–6723, August 2021.

Xi Ye, Rohan Nair, and Greg Durrett. Connecting attributions and qa model behavior on realistic counterfactuals. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5496–5512, 2021.

Sicheng Yu, Yulei Niu, Shuohang Wang, Jing Jiang, and Qianru Sun. Counterfactual variable control for robust and interpretable question answering. *arXiv preprint arXiv:2010.05581*, 2020.

Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Counterfactual zero-shot and open-set visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15404–15414, 2021.

Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. Counterfactual generator: A weakly-supervised method for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7270–7280, 2020.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. Multihiertt: Numerical reasoning over multi hierarchical tabular and textual data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6588–6600, 2022.

Beier Zhu, Yulei Niu, Xian-Sheng Hua, and Hanwang Zhang. Cross-domain empirical risk minimization for unbiased long-tailed classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3589–3597, 2022.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3277–3287, 2021.

Qingfu Zhu, Weinan Zhang, Ting Liu, and William Yang Wang. Counterfactual off-policy training for neural dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3438–3448, 2020.

A THE CREATION OF STRESS TEST SET

To evaluate the dependency on spurious correlation of tabular MRC models, we create a stress test set by editing the factual tables in TAT-QA. Note that we define the stress test data as examples that change the semantic of the factual context and lead to changed answers, which is different from the definition of previous works (Veitch et al., 2021). We believe the stress test set can be used to test the model's genuine understanding of the question and the context, which cannot be accomplished if the model learns shortcuts.

We edit the table of a factual example according to the assumption of the corresponding hypothetical question. First, we extract the new number in the assumption to put in the table by identifying numbers from text strings, *e.g.,* extracting *38,298* from *if the revenue in 2019 were $38,298*. Next, we locate the position in the table, *e.g.,* locating the table cell representing"revenue in 2019". Finally, the stress test data is created by putting the new number into the location identified in the table, which has the same answer as the hypothetical example.

we conduct a human evaluation to verify the quality of the stress test. We sample 70 instances randomly from the stress test, and recruit two college students to examine the fidelity of instances based on three questions: (1) whether the table follows the table-entry consistency (1 if agreed else 0); (2) whether the answer can be correctly derived from the context (1 if agreed else 0); and (3) the complexity of answering the first two questions (0: easy;1:medium;2:hard). The average scores for (1) and (2) are 0.91 and 0.97, showing that the annotators agree that most of the tables are consistent and most of the answers can be correctly deducted. The standard deviation for the complexity score is 0.59 and 0.63 respectively, showing that the stress test has diverse question difficulty. The Cohen's Kappa between the two annotators is 0.32, showing fair agreement between them.

B THE EXPANSION OF HQ WITH THE SAME ANSWER AS OQ

We identify the questions that involves numerical comparison via the following keywords: larger, higher, highest, largest, exceed, less than. We extract the entity E and the number N within the assumption of HQ which intervenes the factual context and changes the answer of OQ. We pair up the hypothetical examples with the factual examples and compare their answers via some simple rules. For example, the question asks about which entity has a higher value, and E within the assumption is the answer to HQ which replaces the factual answer. We can largely decrease the number N in the assumption to create an HQ with the factual answer. We set the range to decrease N to make sure it satisfy the condition most of the times. We can process conversely if E equals the factual answer by increasing N. In total, we create 693 additional HQ for training. We do not create the additional HQ for validation data, and use the same validation set as baseline methods which is a mix of OQ and HQ validation data.

C HYPERPARAMETER SETTING

We use one 32GB GPU. We set the batch size as 16, learning rate as 1e-4 for first-stage training and 1e-5 for second-stage fine-tuning. We train 80 epoch for the first stage and 60 epoch for the second stage. We select the checkpoint with the best validation F_1 result. For the fine-tuning, we wait for 10 epochs before the validation begin. We use Adam as the optimizer and gradient accumulation step of 4. We select $\alpha = 0.07$ and $\beta = 1.3$, both in the range of [0.01, 0.09] with step 0.01 and [0.1, 1.5] with step 0.1.

D IMPLEMENTATION DETAILS OF BASELINE METHODS

- CF-VQA: we adopt a table-only branch to learn the language bias where only the table is remained in the input. We use the RuBi function as the fusion strategy. During inference, the learned table-only bias is subtracted from the total effect. Since CF-VQA does not require counterfactual data, we train it with the factual examples.
- xERM: it is an extension of the above CF-VQA with weights added. We use the empirical risk of the MRC model to calculate the weight.

- CLO: we intend to encode the semantic resemblance of HQ to its corresponding OQ since they differ in a small semantic change. The contrastive loss use the corresponding OQ as the positive example of HQ, and a randomly selected OQ as the negative example. Formally, the contrastive loss is

$$L_{clo} = \frac{e^{dist(r_h, r_o)}}{e^{dist(r^h, r^o)} + e^{dist(r^h, r^{o_{irr}})}} \tag{7}$$

  where r_h, r_o and $r_{o_{irr}}$ denotes the representation of hypothetical example, factual example and an irrelevant factual example encoded by PrLM, and $dist$ denotes cosine similarity after max-pooling the representation. The contrastive loss is added to the total MRC learning objective and weighted as 1.

- GS: we calculate the gradient loss via a pair of factual and hypothetical examples and add the gradient loss to the total MRC learning objective. We set the weight for the gradient loss as 0.01.

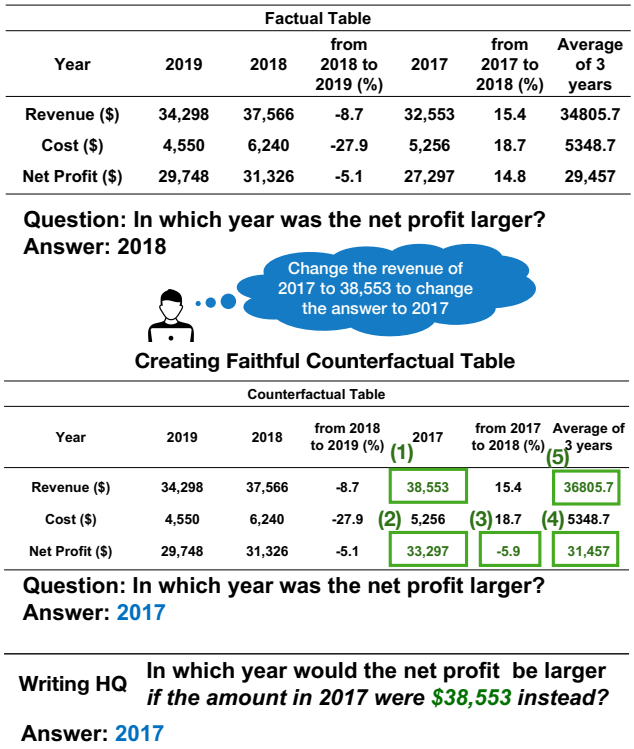


Figure 6: An example of annotation cost comparison for hypothetical example and faithful counterfactual table. For the assumption to change the revenue of 2017 to $ 38533, creating the faithful counterfactual table requires calculating and editing at least 5 numbers, while creating the hypothetical question is much easier by merely writing the assumption in natural language and appending it to the question.

E Annotation Effort Comparison of HQ and Faithful Counterfactual Table

We give an example to illustrate the distinction in annotation effort between creating faithful counterfactual tables and HQ as shown in Figure 6. After reading the factual example and deciding the intervention of changing the revenue in 2017 to $ 38533, the cost for creating HQ is simply writing the assumption in natural language and appending it to the question. However, to create faithful counterfactual table, at least 5 numbers need to be calculated and edited as highlighted in the counterfactual table which is time consuming. As the table gets larger and more complicate,

the annotation cost keeps increasing. This example illustrates that the effort for creating faithful counterfactual table is likely to be much larger than writing HQ, thus HQ is an economical choice.