Hypothetical Training for Robust Machine Reading Comprehension of Tabular Context

Moxin Li[1], Wenjie Wang[1]*, Fuli Feng[2,3], Hanwang Zhang[4], Qifan Wang[5], Tat-Seng Chua[1]

[1]National University of Singapore, [2]University of Science and Technology of China
[3]Institute of Dataspace, Hefei, Anhui, China, [4]Nanyang Technological University
[5]Meta AI

limoxin@u.nus.edu, wangwenjie@u.nus.edu, fulifeng93@gmail.com,
cdzhangjizhi@mail.ustc.edu.cn, wqfcr@meta.com, dcscts@nus.edu.sg

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. The superior generalization results on tabular MRC datasets, including a newly constructed stress test and MultiHiertt, 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 74.9 to 40.0, *cf.* Table 1) Therefore, it is crucial to resolve the spurious correlation issue in the MRC task with tabular context.

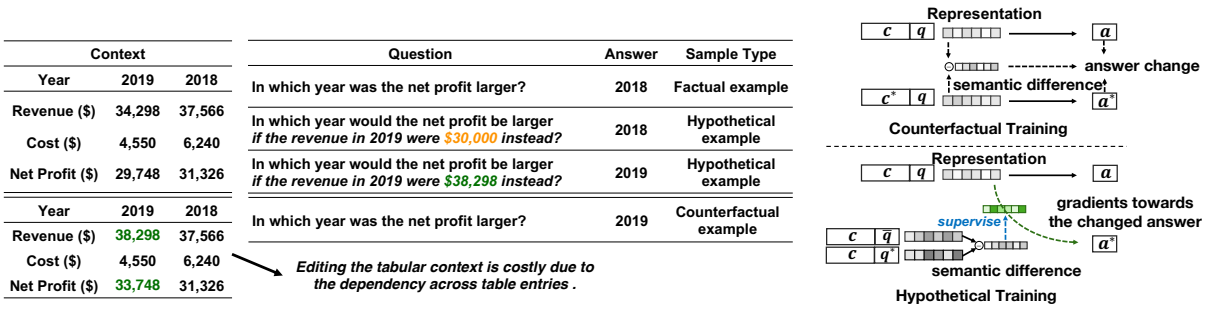
Counterfactual training (Abbasnejad et al., 2020; Teney et al., 2020; Feng et al., 2021; Zhu et al., 2020) is effective for blocking the spurious correlations in various text understanding and reasoning tasks such as visual question answering (Chen et al., 2020a; Niu et al., 2021) and natural language 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 makes the label change—the true causation (Figure 1a). 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 7, annotators need to edit 4 extra numbers for an assumption to change one number. Although ignoring the table entry dependency may save annotation efforts, the unfaithful counterfactual tables will hurt the model robustness (*cf.* Section 3.3).

In this work, we utilize 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. The construction cost of a hypothetical example is undoubtedly lower than the counterfactual example[1]. A hypothetical example consists of a hypothetical question and factual context, which has the equivalent effect on the answer to the corresponding "ideal" counterfactual example. As a concrete case in Figure 1a, the counterfactual 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*

*Corresponding author.

[1]Please refer to Appendix C for detailed comparison.

Findings of the Association for Computational Linguistics: ACL 2023, pages 1220–1236
July 9-14, 2023



(a) (b)

Figure 1: (a) Illustration of factual, hypothetical, and counterfactual examples. (b) Illustration of counterfactual training and the proposed hypothetical training. $\boldsymbol{c}^*$ denotes the counterfactual context.

the net profit larger" from *2018* to *2019*. The answer of the 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 1b, 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 teach the model 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 1a). 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[2] *w.r.t.* the input representation of the factual example towards the changed answer $\boldsymbol{a}^*$. 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 1b). 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 HTF on a stress test and the generalization to another tabular MRC dataset MultiHiertt (Zhao et al., 2022a). Further studies show that HTF also has better understanding to various semantic interventions. Code and data will be made public upon acceptance.

Our contributions are summarized as follows:

- 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

Machine Reading Comprehension. The MRC task aims to answer a question based on the context,

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

where the context might be hybrid in complex scenarios, including paragraphs and tables. Formally, given a question $\boldsymbol{q}$, the MRC model is 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 correlation from $\boldsymbol{c}$ and $\boldsymbol{q}$ to $\boldsymbol{a}$. However, there widely exist spurious correlations (Geirhos et al., 2020) in the complex context. Learning from such spurious correlations will ignore the semantic of $\boldsymbol{c}$ 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 semantics that causally decide the answers. As shown in Figure 1a, the counterfactual example changes the answer of the factual example by minimally perturbing the context according to an assumption with semantic intervention, *e.g., "if the revenue in 2019 were $38,298?"*, highlighting the causal relationship between the semantic of the factual question and context and the answer. By training over the factual and counterfactual examples, the MRC model is able to rely on the highlighted semantic-answer relationship and thus 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 1a, revising the table needs to ensure the consistency and dependency across table entries. The counterfactual table is related to 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 relations (Zhu et al., 2021; Zhao et al., 2022b; Chen et al., 2021). Second, some conventional counterfactual generation methods such as (Yue et al., 2021; Pasupat and Liang, 2016) also cannot guarantee the fidelity of counterfactual examples.

Hypothetical Example. To alleviate the burden of consistency checking, we utilize 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 1a, 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 appending assumptions to the questions in natural language (refer to Appendix C for more comparison).

2.1 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 on answer change. 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 indicate how the representation change can lead to changed answers. The alignment between representation differences and gradients reflects whether the MRC models properly capture semantic intervention.

Given a pair of hypothetical examples $(\boldsymbol{c}, \bar{\boldsymbol{q}}, \boldsymbol{a})$ and $(\boldsymbol{c}, \boldsymbol{q}^*, \boldsymbol{a}^*)$, we pursue the alignment by minimizing a regularization term as follows:

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

where $\bar{\boldsymbol{X}}_h$ and $\boldsymbol{X}_h^*$ denote the representations of $(\boldsymbol{c}, \bar{\boldsymbol{q}})$ and $(\boldsymbol{c}, \boldsymbol{q}^*)$ encoded by the MRC model via feature extractors (*e.g.,* Pre-trained Language Model (PrLM) (Liu et al., 2019)). We calculate $\boldsymbol{X}_h^* - \bar{\boldsymbol{X}}_h$ as the semantic differences of the hypothetical example pair, 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 factual representation $\boldsymbol{X}_f$ towards the changed answer $\boldsymbol{a}^*$, *i.e.,* $\nabla^\intercal f_{\boldsymbol{a}^*}(\boldsymbol{X}_f)$. Such gradients represent 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 $\boldsymbol{X}_h^* - \bar{\boldsymbol{X}}_h$, which is achieved by minimizing their cosine distance.

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.,* $\nabla^\intercal f_{\boldsymbol{a}}(\boldsymbol{X}_h^*)$, 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, $\nabla^\intercal f_{\boldsymbol{a}}(\boldsymbol{X}_h^*)$ should be regulated in the opposite direction of $\nabla^\intercal f_{\boldsymbol{a}^*}(\boldsymbol{X}_f)$:

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

2.2 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,

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

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 two regularization terms for hypothetical training, and the overall loss function is as follows:

$$\mathcal{L}_t + \alpha\mathcal{L}_f + \beta\mathcal{L}_h, \quad (4)$$

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

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

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

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

$$f_{\boldsymbol{a}^*}(\boldsymbol{X}_h^*) - f_{\boldsymbol{a}^*}(\bar{\boldsymbol{X}}_h) \approx \nabla^\intercal f_{\boldsymbol{a}^*}(\bar{\boldsymbol{X}}_h) \cdot (\boldsymbol{X}_h^* - \bar{\boldsymbol{X}}_h) \quad (6)$$

[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 and answer and analogous reasoning process with the factual example $(\boldsymbol{c}, \boldsymbol{q}, \boldsymbol{a})$.

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 $\nabla^\intercal f_{\boldsymbol{a}^*}(\bar{\boldsymbol{X}}_h)$ in $\boldsymbol{J}$ refers to the gradients for $\boldsymbol{a}^*$. From Equation (6), we can maximize the logit difference by increasing the dot product $\nabla^\intercal f_{\boldsymbol{a}^*}(\bar{\boldsymbol{X}}_h) \cdot (\boldsymbol{X}_h^* - \bar{\boldsymbol{X}}_h)$. 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. Note that the cosine distance is calculated after flattening the matrices into vectors. The empirical results in Section 3.3 also validate the superiority of using cosine distance.

Based on the above analysis, we explain the rationality of Equation (1) and Equation (2), respectively. 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 question semantics, $\nabla^\intercal f_{\boldsymbol{a}^*}(\boldsymbol{X}_f)$ and $\nabla^\intercal f_{\boldsymbol{a}^*}(\bar{\boldsymbol{X}})$ refer to the same translation direction of changing the representations of the same semantics towards the changed answer $\boldsymbol{a}^*$, and thus we can again adopt $\boldsymbol{X}_h^* - \bar{\boldsymbol{X}}_h$ to regulate the direction of $\nabla^\intercal f_{\boldsymbol{a}^*}(\boldsymbol{X}_f)$ as shown in Equation (1). Besides, we can perform similar Taylor Expansion for $f(\bar{\boldsymbol{X}}_h)$ regarding $f(\boldsymbol{X}_h^*)$, and constrain the gradients of another hypothetical example $(\boldsymbol{c}, \boldsymbol{q}^*, \boldsymbol{a}^*)$, *i.e.,* $\nabla^\intercal f_{\boldsymbol{a}}(\boldsymbol{X}_h^*)$ by $\bar{\boldsymbol{X}}_h - \boldsymbol{X}_h^*$ symmetrically, as shown in 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 contains hypothetical questions for TAT-QA. To reduce the complexity of answer derivation and focus on studying spurious correlations, we filter out the questions that explicitly execute numerical operations, and only keep the types of questions that extract text spans which still perform numerical reasoning[4]. 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 another hypothetical example with the same answer as the factual example. For evaluation, we first present the validation result of a mix of TAT-QA and TAT-HQA. Besides, we create two tests with different distributions to examine the ability to block spurious correlations. One is a stress test built from TAT-QA by manually making subtle but critical edits on the factual example to change its label (an example in Figure 2). Another test is based on MultiHiertt (Zhao et al., 2022b), a numerical MRC dataset with table and textual context, to examine the generalization ability to other datasets, where a better generalization performance indicates less reliance on spurious correlation. Because MultiHiertt contains long tables and text context and requires a retrieval stage, we directly use the top K retrieval results to construct TAT-QA-like context. We find that the value of K would affect the performance, and thus we create three variations with different values of K and report the averaged results. For details of dataset construction, please refer to Appendix A. We adopt the two common metrics for MRC tasks (Dua et al., 2019), exact-match (EM) and F_1, both in the range of [0, 100].

(a) In which year was the net sale in America larger?

Factual Table			Stress Test Table		
Year	2018	2017	Year	2018	2017
Net Sale in America ($)	259,105	274,056	Net Sale in America ($)	259,105	150,000
Answer: 2017			Answer: 2018		

Figure 2: An example of the stress test edited from the factual example. We only show the shortened tables to highlight their difference.

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.

[4]The reason we filter some questions is that these questions require extra modules for the numerical calculation. As an initial exploration, we consider the more common extractive questions to avoid basing our conclusions on specific questions with additional calculation modules.

	TAT-QA&HQA		Stress Test		MultiHiertt	
	EM	F1	EM	F1	EM	F1
m-OQ	62.6	74.9	32.2	40.0	9.7	12.8
m-OQ&HQ	66.5	78.9	35.4	42.0	11.5	13.7
m-OQ&2HQ	67.5	79.1	37.2	43.4	12.8	15.7
CF-VQA	66.4	77.8	36.4	42.9	10.1	13.3
xERM	67.1	78.1	35.8	43.0	11.8	13.6
CLO	67.0	78.1	37.8	43.7	12.5	15.4
GS	66.9	78.0	36.5	43.7	11.6	14.2
BAI	67.8	78.8	36.2	43.5	12.0	14.8
HTF	**67.9**	**79.2**	**39.7**	**46.0**	**15.3**	**18.5**

Table 1: Performance comparison on the validation set of TAT-QA & HQA, the stress test and MultiHiertt *w.r.t.* EM and F_1 scores. Bold font and underline denote the best and second-best performance, respectively.

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 two hypothetical examples; **GS** (Teney et al., 2020) applies gradient supervision between factual and hypothetical example pairs to shape the decision boundary. **4)** *Interventional training method*: **BAI** (Yu et al., 2022) performs interventions to discover unknown and complex confounders and adopt invariant learning objectives to avoid confounders. 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 EM score on the validation set. More implementation details can be found in Appendix B.

3.2 Performance Comparison (RQ1)

Table 1 shows the performance of all compared methods. We can observe that: **1)** In all cases, the performance on TAT-QA & HQA is much higher than that on the stress test and MultiHiertt, showing that it is challenging to generalize to the stress test and other datasets. **2)** The proposed HTF outperforms all compared methods on the stress test and MultiHierrt indicating its least reliance on spurious correlations, while maintaining comparably top performance on TAT-QA & HQA. Especially, the superior performance of HTF than m-OQ&2HQ validates the rationality of considering the relationships between factual and hypothetical examples via hypothetical training. **3)** Comparing the top three vanilla baselines, we observe that adding two kinds of hypothetical examples can clearly bring performance gain over all tests, verifying the rationality of using hypothetical examples to mitigate spurious correlations. **4)** Debiasing methods cannot achieve more performance gains than m-OQ&2HQ, no matter whether the bias is from the context branch (CF-VQA, xERM) or discovered by interventions (BAI). **5)** Counterfactual training methods (CLO, GS) also underperform HTF, showing the effectiveness of HTF in leveraging the relationship between factual and hypothetical examples.

	Stress Test		MultiHiertt	
	EM	F1	EM	F1
w/o $\mathcal{L}_f$	39.0	45.5	12.5	16.1
w/o $\mathcal{L}_h$	38.3	44.9	13.7	16.6
$\mathcal{L}_f^{dot}$ & $\mathcal{L}_h^{dot}$	37.7	43.6	12.3	15.0
$\mathcal{L}_f^{GS}$ & $\mathcal{L}_h^{GS}$	39.2	45.7	13.2	16.2
$\mathcal{L}_f^{q^*}$ & $\mathcal{L}_h^{q^*}$	38.2	44.7	13.5	16.6
HTF	**39.7**	**46.0**	**15.3**	**18.5**

Table 2: Results of the HTF variants.

3.3 Ablation Studies (RQ2)

Ablation Study of HTF Regularization. We reveal the contribution of each gradient regularization term $\mathcal{L}_f$ 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 observe that the performance decreases on the two tests if we remove either $\mathcal{L}_f$ or $\mathcal{L}_h$. This validates that both gradient regularization terms are critical to remove spurious correlations and enhance the generalization performance.

Rationality of Cosine Regularization. As illustrated in Section 2.3, we compare the regularization terms implemented by dot product or cosine distance. From the results in Table 2, we find that the dot product $\mathcal{L}_f^{dot}$ & $\mathcal{L}_h^{dot}$ largely underperforms HTF with cosine regularization. We attribute the significant difference to that the dot product is norm-sensitive, for which the gradient norm is easily increased while the direction is undermined.

Validation of Calculating the Gradient towards the Changed Label. In our justification, 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 HTF by calculating the gradient towards the factual label instead of the changed label to examine our justification, denoted as $\mathcal{L}_f^{GS}$ & $\mathcal{L}_h^{GS}$. In Table 2, we can find that

In which year was the revenue larger?

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

Unfaithful Counterfactual Table		
Year	2019	2018
Revenue ($)	34,298	37,566
Cost ($)	4,550	16,240
Net Profit ($)	29,748	31,326
answer: 2019		

Figure 3: Example of the unfaithful counterfactual table.

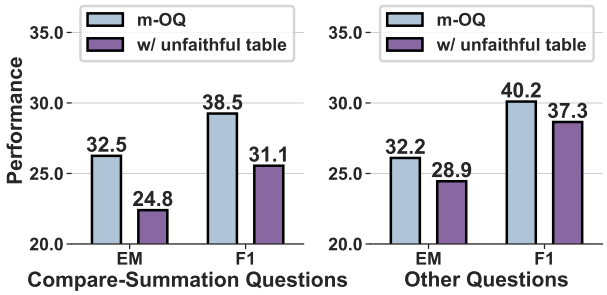


Figure 4: Performance comparison on the stress test by adding unfaithful counterfactual tables.

the variant performs worse than HTF, thus empirically validating the superiority of our justification. Moreover, we replace the factual example with the hypothetical example of the same answer in calculating the gradient in $\mathcal{L}_f$, denoted as $\mathcal{L}_f^{q^*}$ & $\mathcal{L}_h^{q^*}$, which clearly has inferior results than HTF but still outperforms the baselines in Table 1.

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 3, 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 the cost will cause inconsistency between the table entries, leading to unfaithful counterfactual tables. If such unfaithful examples in Figure 3 are used for training with factual examples, the MRC model will wrongly attribute the answer change to the changed cost value, causing confusion in training and hurting the performance. To validate that, we annotate 220 unfaithful counterfactual examples, then train a variant of m-OQ by adding the unfaithful counterfactual tables into the training data, and test it on the stress test. From the results in Figure 4, we discover that for

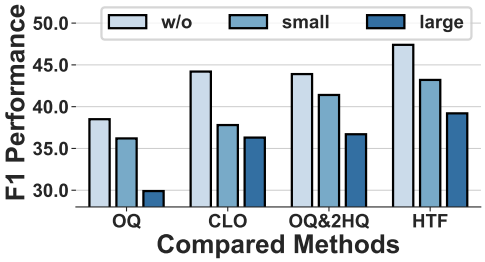


Figure 5: Performance on the stress test with changed number scale. *w/o*, *small* and *large* refers to no scaling, slight scaling and large scaling, respectively.

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)

We study the generalization ability of HTF to new semantic interventions on the table. We look into how HTF generalizes to new tables with **numbers of unusual scale**. We identify a type of questions from the stress test asking about numerical conditions, *e.g., "which values is larger (or smaller) than a threshold A?"*, and generate new test cases by scaling the target numbers that are larger (or smaller) than A in the table. We increase the target number by five or six times if it is larger than A and otherwise decrease it by five or six times, denoted as slightly-scaled examples. We also try with 10 and 12 times, denoted as largely-scaled examples[5]. We test HTF, CLO, m-OQ and m-OQ&2HQ on the scaled examples. As shown in Figure 5, 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 encounters the smallest performance drop between the original 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.

We then study the 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" in TAT-QA as they are the most frequently appeared answers. We perform interventions from two aspects for these questions

[5]Note that the edited examples maintain the same answers as the original examples.

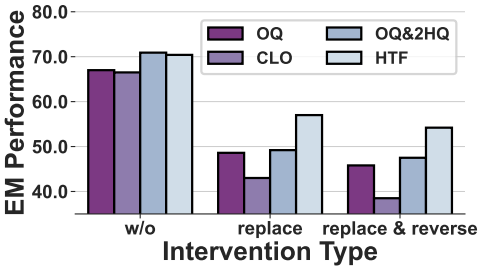


Figure 6: Study on the spurious correlation of "which year" questions and the top year answers. "w/o" denotes no operation on the table.

with frequent answers. Firstly, we break the word correlation between the questions and the frequent answers by replacing 2019, 2018, and 2017 in the contexts with their corresponding English words (*e.g.,* two thousand and nineteen), where the MRC model is expected to identify the span of the correct English words. This intervention is denoted as *replace*. Besides, we try changing the year order by replacing 2019, 2018, and 2017 with the English word of 2017, 2019, and 2018, respectively (denoted as *replace&shuffle*) to examine the bias toward predicting the earliest or the latest year. As shown in Figure 6, we can observe that the replacement with English words decreases the performance for all compared methods, and shuffling the year order can further damage the performance, revealing the existence of the two spurious correlations. Nevertheless, HTF has the smallest drop and thus captures the fewest spurious correlations.

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 interventions 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), MRC (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 and Zhang, 2021; Chen et al., 2020a; Wang et al., 2021b; Feng et al., 2021; Abbasnejad et al., 2020; Paranjape et al., 2022; Yu et al., 2022; Wang 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 and Geiger, 2021). In tabular MRC task, automatically creating counterfactual examples is infeasible and sufficient 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 bias of NLI (Poliak et al., 2018), the word alignment of passage and options in QA (Yu et al., 2020), the simplicity bias (Teney et al., 2022), all of which hinder the generalization ability of deep models to out-of-distribution test sets (*e.g.,* (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 and then mitigating the bias (He et al., 2019; Cadene et al., 2019; Ghaddar et al., 2021; Mahabadi et al., 2020), training multiple models (Teney et al., 2022; Clark et al., 2019; Pagliardini et al., 2022), invariant learning (Arjovsky et al., 2019; Li et al., 2022b), instance mixup (Hwang et al., 2022), and using causal inference techniques (Wang et al., 2021c,a).

Tabular MRC. Enabling machines to understand and reason over complex contexts such as tables has become a popular research goal in recent years, due to the overwhelming tabular data in the real work. Many tabular QA datasets are proposed, such as WikiTQ (Pasupat and Liang, 2015), SQA (Iyyer et al., 2017), Spider (Yu et al., 2018). Many tabular MRC datasets require numerical reasoning ability, such as FinQA (Chen et al., 2021), TAT-QA (Zhu et al., 2021), HybridQA (Chen et al., 2020b), MultiHierrt (Zhao et al., 2022b). The solutions often include numerical calculation steps (Chen et al., 2021; Zhu et al., 2021) and table understanding techniques (Herzig et al., 2020). In this work, we adopt the standard method of TAGOP on TAT-QA.

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 generalization performance on a stress test and another tabular MRC dataset. 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) applying hypothetical training to various domains apart from the financial domain.

Limitations

Although HTF has achieved promising performance on removing spurious correlations, we identify the following limitations. Firstly, although HTF encounters the smallest performance decrease among compared methods under multiple semantic interventions, the interventions still cause a performance drop. Therefore, more approaches can be explored to further improve the generalization ability of HTF, such as increasing the scale of the backbone model or applying more informative hypothetical examples. Secondly, the experiments are only conducted in the financial domain due to limited datasets with sufficient annotation of hypothetical examples. Since hypothetical examples are more economic to obtain than counterfactual examples, we believe that more datasets with hypothetical examples will be proposed in the future and thus HTF can be applied in more domains. Thirdly, we are unable to compare the effectiveness of hypothetical and counterfactual examples because TAT-QA does not contain both types, and constructing all counterfactual examples is impractical for us due to cost constraints. Note that we do not conclude any effectiveness relationship between hypothetical and counterfactual examples in the paper.

Acknowledgement

This research is supported by Sea-NExT Joint Lab, Singapore MOE AcRF T2, the National Key Research and Development Program of China (2022YFB3104701), and the National Natural Science Foundation of China (62272437).

References

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

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. 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*, pages 4971–4980.

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

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

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

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

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

Seungtaek Choi, Myeongho Jeong, Hojae Han, and Seung-won Hwang. 2022. C2l: Causally contrastive learning for robust text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10526–10534.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. 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)*, pages 4069–4082.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. 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)*, pages 2368–2378.

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

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

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

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

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

Abbas Ghaddar, Phillippe Langlais, Mehdi Rezagholizadeh, and Ahmad Rashid. 2021. End-to-end self-debiasing framework for robust NLU training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1923–1929, Online. Association for Computational Linguistics.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. 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)*, pages 878–892.

He He, Sheng Zha, and Haohan Wang. 2019. 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)*, pages 132–142.

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

Inwoo Hwang, Sangjun Lee, Yunhyeok Kwak, Seong Joon Oh, Damien Teney, Jin-Hwa Kim, and Byoung-Tak Zhang. 2022. Selecmix: Debiased learning by contradicting-pair sampling. In *Advances in Neural Information Processing Systems*.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1821–1831.

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

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

Moxin Li, Fuli Feng, Hanwang Zhang, Xiangnan He, Fengbin Zhu, and Tat-Seng Chua. 2022a. 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)*, pages 57–69.

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

Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. 2020. 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)*, pages 3285–3292.

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

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

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. 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*, pages 3428–3448.

Guoshun Nan, Jiaqi Zeng, Rui Qiao, Zhijiang Guo, and Wei Lu. 2021. Uncovering main causalities for long-tailed information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9683–9695.

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.

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

Matteo Pagliardini, Martin Jaggi, François Fleuret, and Sai Praneeth Karimireddy. 2022. Diversity through disagreement for better transferability. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*.

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

Panupong Pasupat and Percy Liang. 2015. 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)*, pages 1470–1480.

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

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

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

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. 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)*, pages 5043–5053.

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

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

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

Damien Teney, Ehsan Abbasnejad, Simon Lucey, and Anton van den Hengel. 2022. 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*, pages 16761–16772.

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

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

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

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

Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, Min Lin, and Tat-Seng Chua. 2022. Causal representation learning for out-of-distribution recommendation. In *Proceedings of the ACM Web Conference 2022*, pages 3562–3571.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. 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)*, pages 6707–6723, Online. Association for Computational Linguistics.

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

Sicheng Yu, Jing Jiang, Hao Zhang, and Qianru Sun. 2022. Interventional training for out-of-distribution natural language understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 11627–11638.

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

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921.

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

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

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022a. 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)*, pages 6588–6600, Dublin, Ireland. Association for Computational Linguistics.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022b. 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)*, pages 6588–6600.

Beier Zhu, Yulei Niu, Xian-Sheng Hua, and Hanwang Zhang. 2022. Cross-domain empirical risk minimization for unbiased long-tailed classification. In *AAAI Conference on Artificial Intelligence*.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. 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)*, pages 3277–3287.

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

A Dataset Detail

A.1 Dataset Statistics

About the filtering of TAT-QA and TAT-HQA, we discard the “arithmetic" type of questions and keep the “counting", “span" and “multi-span" questions. After filtering, we maintain 8772 TAT-QA and TAT-HQA questions for the training split, and 1055 for the validation split which all compared methods are evaluated on. The reason for the filtering is that the filtered factual examples do not have the corresponding hypothetical examples with changed labels, and thus cannot be applied to hypothetical training for removing spurious correlation.

For MultiHiertt, we utilize the validation set and we run the released code[6] to generate the retrieval results. We try three setting to create TAT-QA-like data with the top K_1 table retrieval results and top K_2 text retrieval results, $K_1 = 5, K_2 = 5$, $K_1 = 5, K_2 = 10$, and $K_1 = 6, K_2 = 10$. Similarly, we keep only the questions that extract text spans from the context, and we remove the questions that do not contain the answers in the TAT-QA like contexts. In total we obtain 418 questions. The detailed results for the compared methods on different MultiHiertt variations can be found in Table 3, where HTF outperforms compared methods on all settings.

For the test sets used in in Section 3.4, we use 606 questions from the stress test for number scaling, and each scaling test set contain 1212 questions. We use 179 “which year" questions from the validation set of TAT-QA to intervene on the frequent answers.

A.2 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

[6] https://github.com/psunlpgroup/MultiHiertt.

	m-OQ		m-OQ&HQ		m-OQ&2HQ		CF-VQA		xERM		CLO		GS		BAI		HTF	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
$K_1 = 5, K_2 = 5$	10.0	13.0	10.0	12.2	13.6	16.9	11.4	14.2	12.1	14.1	11.4	14.5	13.6	15.6	12.1	15.3	**15.7**	**19.3**
$K_1 = 5, K_2 = 10$	8.5	11.7	10.6	13.1	12.1	14.8	9.2	12.5	9.9	11.9	11.3	14.0	11.3	14.1	11.3	14.2	**14.9**	**18.4**
$K_1 = 6, K_2 = 10$	10.5	13.7	14.0	15.7	12.6	15.4	9.8	13.2	13.3	14.8	14.7	17.6	9.8	12.9	12.6	14.8	**15.4**	**17.9**

Table 3: Results of different settings of MultiHiertt.

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. In total we obtain 921 stress test examples.

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.

A.3 The Expansion of Hypothetical Examples with the Same Answer as the Factual Example

In most cases, the assumption in the hypothetical question intervenes on an entity in the table, denoted as E, by assigning a new value N to it, *e.g., if the revenue in 2019 were $38,298* assigns N = *$38,298* to the table cell E = *the revenue in 2019*. Usually, E is correlated with the answer change between the hypothetical and factual examples, *e.g.,* E replaces the factual answer or E is removed from the factual answer. Therefore, by simply manipulating the value N in the hypothetical assumption, we can nullify the effect of the hypothetical assumption on E and keep the factual answer unchanged. We identify the questions that involves numerical comparison via the following keywords: larger, higher, highest, largest, exceed, less than, and extract the entity E and the value N from the assumption. We pair up the hypothetical examples with the factual examples, compare their answers and change the N in the hypothetical assumption via some simple rules. For example, the factual question asks about which entity has a higher value, and E within the hypothetical assumption is the answer of the hypothetical question which replaces the factual answer. We can largely decrease the value of N to create a hypothetical example with the factual answer. We randomly select the scale to decrease N from 5 to 10 times to make sure that the decrease of N can obtain the factual answer. On the contrary, if E is within the factual answer, we can process conversely by increasing N 5 to 10 times. In total, we create 709 additional hypothetical examples for training (in total 9481 training instances). We do not create additional hypothetical examples for validation data, and use the released TAT-QA and TAT-HQA validation data.

B Implementation Details of Compared Methods

We implement the methods based on the released code of TAGOP[7]. All methods are run on one 24GB RTX3090, with Pytorch=1.7 CUDA=11.0. We tune the batch size in [4, 8, 16], and the maximum training epoch in [60, 80, 100], and the loss weights in [0.01, 0.02, 0.05, 0.1] for all compared methods, and select the checkpoint with best validation EM. The other parameter setting follows the released TAGOP as we discover that changing them

[7] https://github.com/NExTplusplus/TAT-QA.

is unlikely to make further improvement.

We apply a two-staged training for HTF by first training on all factual and hypothetical examples with TAGOP loss $\mathcal{L}_t$, and then fine-tuning on the triplets of a factual and two hypothetical examples with additional 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 and β as 0.01, 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. For the fine-tuning, we wait for 10 epochs before the validation begin. The total number of GPU hours is approximately 15.

- CF-VQA: apart from the original question and context input, we adopt an additional context-only branch to capture the language bias by masking the question and keeping only the context as input. We use the RuBi function as the fusion strategy to fuse the original representation and the context-only representation. During inference, the learned context-only bias is subtracted from the total effect. We set the KL weight as 0.01.

- xERM: it is an extension of the above CF-VQA by applying learned weights for the two branches. The weights are transformed from the empirical risks of the two branches, which is used for fusing the two representations before prediction.

- CLO: we apply contrastive loss between the factual example and two hypothetical examples with different answers to encode their semantic similarity. The contrastive loss draws close of the factual example and the hypothetical example with the same answer, and differentiate the hypothetical examples with different answers. Formally, the contrastive loss is defined as

$$L_{clo} = \frac{e^{dist(\boldsymbol{X}_f, \boldsymbol{X}_h^*)}}{e^{dist(\boldsymbol{X}_f, \boldsymbol{X}_h^*)} + e^{dist(\boldsymbol{X}_h^*, \bar{\boldsymbol{X}}_h)}} \quad (7)$$

where $dist$ denotes cosine similarity after doing max pooling on the representations. The contrastive loss is added to the total MRC learning objective and weighted as 0.1.

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

- BAI: For the automatic stratification stage, we use m-OQ&2HQ as the reference model and train 40 epoch with learning rate 1e-2. We set the number of fine-grained partition as 5 and coarse-grained partition as 2. For the bottom-up intervention stage, we train 80 epochs.

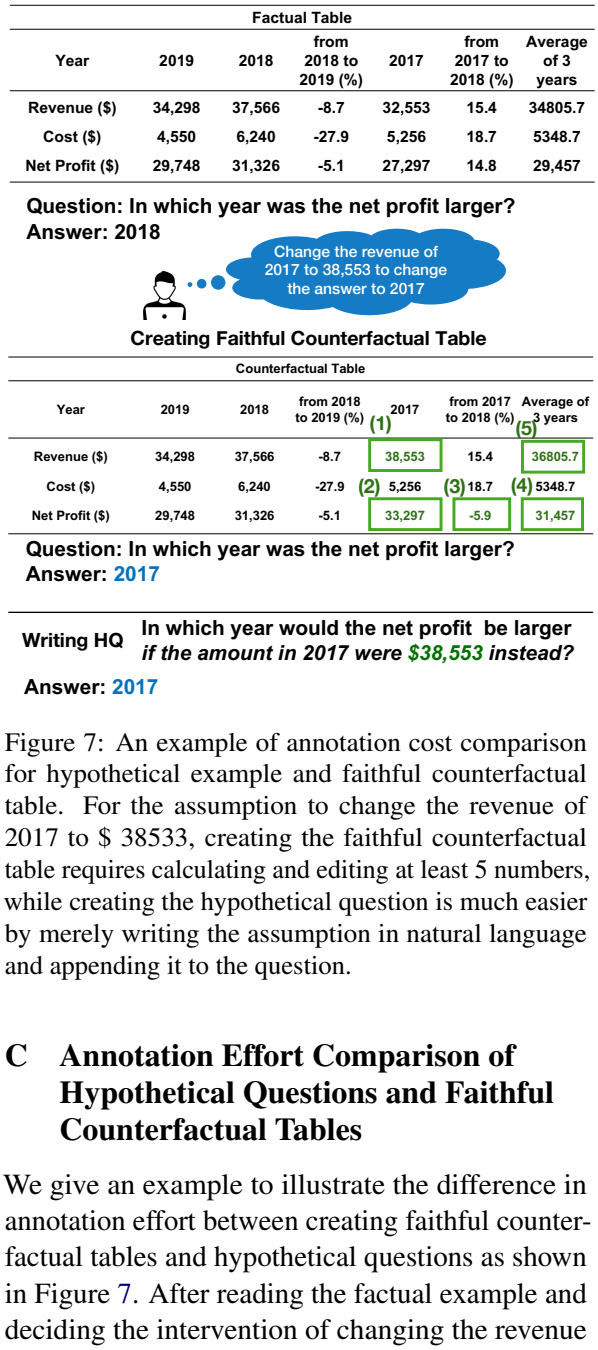


Figure 7: An example of annotation cost comparison for hypothetical example and faithful counterfactual table. For the assumption to change the revenue of 2017 to $ 38533, creating the faithful counterfactual table requires calculating and editing at least 5 numbers, while creating the hypothetical question is much easier by merely writing the assumption in natural language and appending it to the question.

C Annotation Effort Comparison of Hypothetical Questions and Faithful Counterfactual Tables

We give an example to illustrate the difference in annotation effort between creating faithful counterfactual tables and hypothetical questions as shown in Figure 7. After reading the factual example and deciding the intervention of changing the revenue in 2017 to $ 38533, the cost for creating hypothetical question is simply writing the assumption in

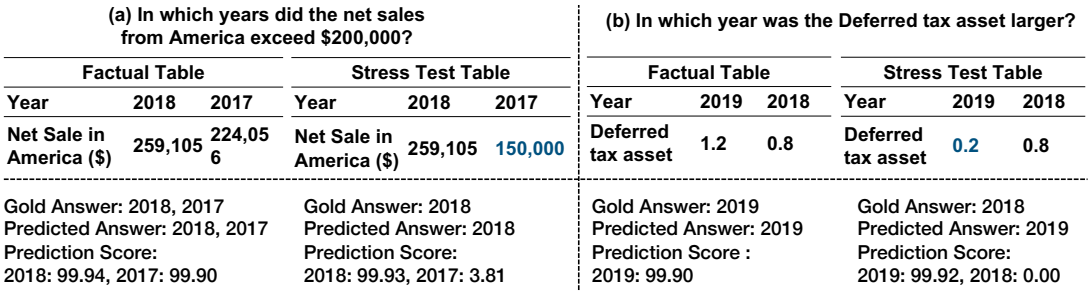


Figure 8: Case study of HTF's predictions. The tables are shortened to save space.

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

For the cost comparison, we conduct a human study with 4 college students to annotate 144 hypothetical and counterfactual examples on randomly sampled TAT-QA tables. We find that that the construction time for a hypothetical example is on average 45.6% of that for a counterfactual example, and the number of modifications required is 40.6% of that for a counterfactual example. Even for a pair of hypothetical examples, the construction cost is still lower than that of a counterfactual example, 91.2% in time and 81.2% in the number of modifications. These results suggest that hypothetical examples are a more cost-effective approach.

D A Simple Low-Cost Comparison on the Effectiveness of Hypothetical and Counterfactual Examples

Since the full comparison on the effectiveness of counterfactual and hypothetical examples is not available for us as explained in Section 5, we provide a simple low-cost comparison here. We hand-annotate 16 shots of counterfactual examples and fine-tune m-OQ with CLO. We compare HTF with the same setting over the corresponding 16 hypothetical example pairs. The results show that HTF performs slightly worse than CLO in terms of EM and F1 scores (33.9 and 41.3 for HTF vs. 34.6 and 42.0 for CLO, respectively).Given the lower construction cost, we believe hypothetical examples are a promising option.

E Case study.

We present two examples to demonstrate the effect of HTF on model prediction in Figure 8. In example (a), HTF gives correct predictions to both the factual and the stress test examples. 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 the sensitivity of HTF.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
At page 9

☑ A2. Did you discuss any potential risks of your work?
At page 9

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

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

B ☑ Did you use or create scientific artifacts?

Section 3, Appendix A

☑ B1. Did you cite the creators of artifacts you used?
Section 3

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We were unable to find the license for the dataset we used.

☐ 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)?
No response.

☐ 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?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☑ 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.
Section 3, Appendix A.

C ☑ Did you run computational experiments?

Section 3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix B

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?
Section 3, Appendix B

☑ 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?
Section 3, Appendix A

☑ 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.)?
Section 3, Appendix A, B.

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

Appendix

☐ 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.?
No response.

☐ 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)?
No response.

☐ 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?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.