Analyzing the Effectiveness of the Underlying Reasoning Tasks in Multi-hop Question Answering

Xanh Ho,[*] [◇,♡] **Anh-Khoa Duong Nguyen,**[*] [♣] **Saku Sugawara,**[♡] and **Akiko Aizawa**[◇,♡,♠]

◇The Graduate University for Advanced Studies, Kanagawa, Japan
♡National Institute of Informatics, Tokyo, Japan
♣Independent Researcher
♠The University of Tokyo, Tokyo, Japan
{xanh, saku, aizawa}@nii.ac.jp
dnanhkhoa@live.com

Abstract

To explain the predicted answers and evaluate the reasoning abilities of models, several studies have utilized underlying reasoning (UR) tasks in multi-hop question answering (QA) datasets. However, it remains an open question as to how effective UR tasks are for the QA task when training models on both tasks in an end-to-end manner. In this study, we address this question by analyzing the effectiveness of UR tasks (including both sentence-level and entity-level tasks) in three aspects: (1) QA performance, (2) reasoning shortcuts, and (3) robustness. While the previous models have not been explicitly trained on an entity-level reasoning prediction task, we build a multi-task model that performs three tasks together: sentence-level supporting facts prediction, entity-level reasoning prediction, and answer prediction. Experimental results on 2WikiMultiHopQA and HotpotQA-small datasets reveal that (1) UR tasks can improve QA performance. Using four debiased datasets that are newly created, we demonstrate that (2) UR tasks are helpful in preventing reasoning shortcuts in the multi-hop QA task. However, we find that (3) UR tasks do not contribute to improving the robustness of the model on adversarial questions, such as sub-questions and inverted questions. We encourage future studies to investigate the effectiveness of entity-level reasoning in the form of natural language questions (e.g., sub-question forms).[1]

1 Introduction

The task of multi-hop question answering (QA) requires a model to read and aggregate information from multiple paragraphs to answer a given question (Figure 1a). Several multi-hop QA datasets have been proposed, such as QAngaroo (Welbl et al., 2018), HotpotQA (Yang et al., 2018), and MuSiQue (Trivedi et al., 2022). In HotpotQA, the authors provide sentence-level supporting facts (SFs) to test the reasoning ability and explainability of the models. However, owing to the design of the sentence-level SFs task (binary classification) and the redundant information in the sentences, Inoue et al. (2020) and Ho et al. (2020) show that the sentence-level SFs are insufficient to explain and evaluate multi-hop models in detail. To address this issue, R^4C (Inoue et al., 2020) and 2WikiMultiHopQA (2Wiki; Ho et al., 2020) datasets provide an entity-level reasoning prediction task to explain and evaluate the process of answering questions. Entity-level reasoning information is defined as a set of triples that describes the reasoning path from question to answer (Figure 1b).

Several previous studies (Chen et al., 2019; Fu et al., 2021a) utilize sentence-level SFs and/or entity-level reasoning information to build explainable models by using question decomposition (Min et al., 2019b; Perez et al., 2020) or predicting sentence-level SFs. The advantages of these pipeline models are that they can exploit the underlying reasoning (UR) process in QA and their predicted answers are more interpretable. However, the question remains as to how effective training on UR tasks is for the QA task in an end-to-end manner. Although a few end-to-end models have also been introduced (Qiu et al., 2019; Fang et al., 2020), these models are not explicitly trained on entity-level and answer prediction tasks.

In addition to the triple form, the sub-question form is another way to utilize entity-level reasoning information. Specifically, Tang et al. (2021) utilize question decomposition as an additional sub-question evaluation for bridge questions (there are two types of questions: bridge and comparison) in HotpotQA. They only use sub-questions for evaluation and do not fine-tune the models on them. In addition, Ho et al. (2022) use sub-questions for both evaluation and training. However, they only

---

Question: Who is the paternal grandfather of Joan of Valois, Countess of Beaumont?

Paragraph A: Joan of Valois, Countess of Beaumont

[1] Joan of Valois (1304 – 1363) was the daughter of Charles of Valois and his second wife ...

Paragraph B: Charles, Count of Valois

[2] Charles of Valois (12 March 1270 – 16 December 1325), the third son of Philip III of France and, ... [3] ...

Answer: Philip III of France

Sentence-level supporting facts: 1, 2

Entity-level reasoning prediction (Evidence):

Step 1: ("Joan of Valois, Countess of Beaumont", "father", "Charles of Valois") &

Step 2: ("Charles of Valois", "father", "Philip III of France")

➕ QA Performance   ⛔ Robustness

➕ Reasoning Shortcuts

Paragraph A: Joan of Valois, Countess of Beaumont

[1] *We can also establish the global weak solution ...* [2] Joan of Valois (1304 – 1363) was the daughter of Charles of Valois and his second wife ...

Paragraph B: Charles, Count of Valois

[3] *This gives a clear impulse to develop ...* [4] Charles of Valois (12 March 1270 – 16 December 1325), the third son of Philip III of France and, ... [5] ...

Adversarial Question: Who is the father of Joan of Valois, Countess of Beaumont?

a) Standard QA task format | b) UR tasks and three aspects | c) Debiased and Adversarial examples

Figure 1: Example of (a) a standard multi-hop question, (b) two underlying reasoning tasks in the QA process and three aspects in our analysis, '+' and '-' indicate that the UR tasks have a positive and negative impacts, respectively, and (c) debiased and adversarial examples that are used in our study.

focus on comparison questions for date information. In contrast, we focus on the triple form of the entity-level information and conduct experiments using two datasets, 2Wiki and HotpotQA-small (obtained by combining HotpotQA and R^4C), which include both types of questions.

In this study, we analyze the effectiveness of UR tasks (including both sentence-level and entity-level) in three aspects: (1) *QA performance*, (2) *reasoning shortcuts*, and (3) *robustness*. First, QA performance is the final objective of the QA task. We aim to answer the following question: **(RQ1)** *Can the UR tasks improve QA performance?* For the second aspect, previous studies (Chen and Durrett, 2019; Jiang and Bansal, 2019a; Min et al., 2019a; Trivedi et al., 2020) demonstrate that many questions in the multi-hop QA task contain biases and reasoning shortcuts (Geirhos et al., 2020), where the models can answer the questions by using heuristics. Therefore, we aim to ask the following: **(RQ2)** *Can the UR tasks prevent reasoning shortcuts?* For the final aspect, to ensure safe development of NLP models, robustness is one of the important issues and has gained tremendous amount of research (Wang et al., 2022). In this study, we aim to test the robustness of the model by asking modified versions of questions, such as sub-questions and inverted questions. Our question is **(RQ3)** *Do the UR tasks make the models more robust?*

There are no existing end-to-end models that can perform three tasks simultaneously (sentence-level SFs prediction, entity-level prediction, and answer prediction); therefore, we first build a multi-hop BigBird-base model (Zaheer et al., 2020) to perform these three tasks simultaneously. We then evaluate our model using two multi-hop datasets: 2Wiki and HotpotQA-small. To investigate the ef-

fectiveness of the UR tasks, for each dataset, we conduct three additional experiments in which the model is trained on: (1) answer prediction task, (2) answer prediction and sentence-level prediction tasks, and (3) answer prediction and entity-level prediction tasks. We also create four debiased sets (Figure 1c) for 2Wiki and HotpotQA-small for **RQ2**. We create and reuse adversarial questions for 2Wiki and HotpotQA-small for **RQ3**.

The experimental results indicate that the UR tasks can improve QA performance from 77.9 to 79.4 F1 for 2Wiki and from 66.4 to 69.4 F1 for HotpotQA-small **(RQ1)**. The results of the models on the four debiased sets reveal that the UR tasks can be used to reduce reasoning shortcuts **(RQ2)**. Specifically, when the model is trained on both answer prediction and UR tasks, the performance drop of the model on the debiased sets is lower than that when the model is trained only on answer prediction (e.g., 8.9% vs. 13.4% EM). The results also suggest that the UR tasks do not make the model more robust on adversarial questions, such as sub-questions and inverted questions **(RQ3)**. Our analysis shows that correct reconstruction of the entity-level reasoning task contributes to finding the correct answer in only 37.5% of cases. This implies that using entity-level reasoning information in the form of triples does not answer adversarial questions, in this case, the sub-questions. We encourage future work to discover the effectiveness of the entity-level reasoning task in the form of sub-questions that have the same form as multi-hop QA questions.

2 Background

Reasoning Tasks in Multi-hop QA   In this study, we consider UR tasks in multi-hop QA including two levels: *sentence-level* and *entity-level*. The

Figure 2: Information on the position of sentence-level SFs in the dev. sets of the three datasets.

Figure 3: Our model has three main steps: paragraph selection, context encoding, and multi-task prediction.

sentence-level SFs prediction task was first introduced by Yang et al. (2018). This task requires a model to predict a set of sentences that is necessary to answer a question (Figure 1).

To evaluate the UR process of the models, derivation and evidence information were introduced in R^4C and 2Wiki, respectively. Both derivation and evidence are sets of triples that represent the reasoning path from question to answer. The difference is the form; derivation in R^4C uses a semi-structured natural language form, whereas evidence in 2Wiki uses a structured form. We conduct experiments with both R^4C (HotpotQA-small) and 2Wiki. For consistency, we use the term *entity-level reasoning prediction task* to denote the derivation task in R^4C and the evidence task in 2Wiki.

Reasoning Shortcuts and Biases In this study, we consider both reasoning shortcuts and biases to be similar. These are spurious correlations in the dataset that allow a model to answer the question correctly without performing the expected reasoning skills, such as comparison and multi-hop reasoning. Following previous studies (Jiang and Bansal, 2019a; Ko et al., 2020), we use the terms *word overlap shortcut* and *position bias*.

To check whether the UR tasks can prevent reasoning shortcuts, we first identify the types of shortcuts that exist in HotpotQA-small and 2Wiki. We use heuristics to identify the word overlap shortcut (Appendix A). We find that the word overlap shortcut is common in HotpotQA-small, but not in 2Wiki. The small sample size of HotpotQA-small (Section 4) increases the uncertainty of the obtained results. Therefore, within the scope of this study, we mainly experiment with position bias.

We observe that many examples in 2Wiki contain answers in the first sentence. Therefore, we divide every sentence-level SF in each gold para-
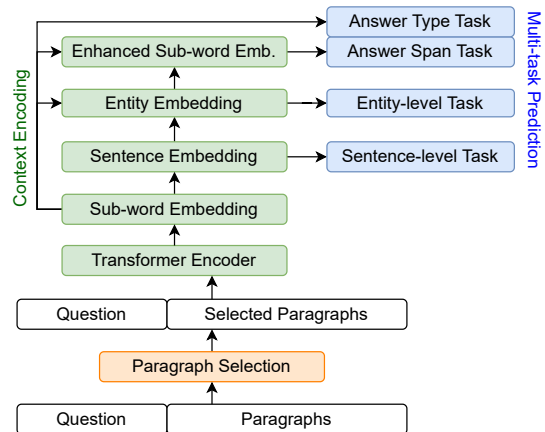
graph into two levels: the first sentence (position_0) and the remaining sentences (position_other). Subsequently, we obtain the percentage of each level by dividing the total number of each level (e.g., position_0) by the total number of SFs. Figure 2 illustrates the information on the position of sentence-level SFs in dev. sets of three datasets. We find that all three datasets have a bias toward the first sentence. We also find that 2Wiki has more position biases than HotpotQA and HotpotQA-small.

3 Our Multi-task Model

To investigate the usefulness of UR tasks for the QA task, we jointly train the corresponding tasks: sentence-level SFs prediction, entity-level prediction, and answer prediction. Figure 3 illustrates our model. To handle long texts, we use the Big-Bird model (Zaheer et al., 2020), which is available in Hugging Face's transformers repository.[2] Our model comprises three main steps: (1) paragraph selection, (2) context encoding, and (3) multi-task prediction. We use the named entity recognition (NER) models of Spacy[3] and Flair (Akbik et al., 2019) to extract all entities in the context and use them for the entity-level prediction task.

Paragraph Selection Following previous models (Qiu et al., 2019; Fang et al., 2020; Tu et al., 2020), instead of using all the provided paragraphs, we first filter out answer-unrelated paragraphs. We follow the paragraph selection process described in Fang et al. (2020). First, we retrieve first-hop

1165

paragraphs by using title matching or entity matching. We then retrieve second-hop paragraphs using the hyperlink information available in Wikipedia. When we retrieve paragraphs, we reuse a paragraph ranker model[4] from the hierarchical graph network (HGN) model (Fang et al., 2020) to rank input paragraphs using the probability of whether they contain sentence-level SFs.

Context Encoding To obtain vector representations for sentences and entities, we first combine all the selected paragraphs into one long paragraph and then concatenate it with the question to form a context C. Specifically, $C = [[CLS], q_1, ..., q_m, [SEP], p_1, ..., p_n, [SEP]]$, where m and n are the lengths of the question q and the combined paragraph p (all selected paragraphs), respectively. The context C is then tokenized into l sub-words before feeding into BigBird to obtain the contextual representation C' of the sub-words:

$$C' = \text{BigBird}(C) \in \mathbb{R}^{l \times h}, \qquad (1)$$

where h is the hidden size of the BigBird model. Next, we obtain the representation $s_i \in \mathbb{R}^{2h}$ of the i-th sentence and the representation $e_j \in \mathbb{R}^{4h+d_t}$ of the j-th entity, as follows:

$$
\begin{aligned}
s_i &= C'_{S^i_{\text{start}}}; C'_{S^i_{\text{end}}} \\
e_j &= C'_{E^j_{\text{start}}}; C'_{E^j_{\text{end}}}; t_j; s_k,
\end{aligned}
\qquad (2)
$$

where [;] denotes the concatenation of the two vectors, $C'_{S^i_{\text{start}}}$ and $C'_{E^j_{\text{start}}}$ denote the first sub-word representations of the i-th sentence and j-th entity, respectively. $C'_{S^i_{\text{end}}}$ and $C'_{E^j_{\text{end}}}$ denote the last sub-word representations of the i-th sentence and j-th entity, respectively. We enrich the entity embedding e_j by concatenating it with a d_t-dimensional type embedding t_j and a sentence embedding s_k, where k is the index of the sentence containing the j-th entity.

We also leverage the entity information to improve the contextual representation of sub-words C' as it is mainly used for the answer prediction task, which will be described in the next section. Thus, the enhanced sub-word representation C''_i of the i-th sub-word is calculated as follows:

$$C''_i = C'_i; e_k \in \mathbb{R}^{5h+d_t}, \qquad (3)$$

where e_k is the embedding of the k-th entity containing the i-th sub-word. Otherwise, e_k is a null vector with the same dimension.

---

[4] https://github.com/yuwfan/HGN

Multi-task Prediction After context encoding, we train our model on three main tasks together: (1) sentence-level prediction, (2) entity-level prediction, (3) and answer prediction. We split the answer prediction task into two sub-tasks, similar to previous studies (Yang et al., 2018; Fang et al., 2020), including answer type prediction and answer span prediction. We train our model by minimizing the joint loss for all tasks, as follows:

$$
\begin{aligned}
L_{\text{joint}} = &\lambda_{\text{sent}} L_{\text{sent}} + \lambda_{\text{ent}} L_{\text{ent}} + \\
&\lambda_{\text{ans}}(L_{\text{start}} + L_{\text{end}} + L_{\text{type}}),
\end{aligned}
\qquad (4)
$$

where λ_{sent}, λ_{ent}, and λ_{ans} are the hyper-parameters for three tasks: sentence-level prediction, entity-level prediction, and answer prediction (details are given in Appendix B.1).

For the sentence-level prediction task, we use a binary classifier to predict whether a sentence is a supporting fact. For the answer type prediction task, we use a 4-way classifier to predict the probabilities of *yes*, *no*, *span*, and *no answer*. Two linear classifiers are used for the answer span prediction task to independently predict the start and end tokens of the answer span.

Different from existing end-to-end models (Qiu et al., 2019; Fang et al., 2020), our model is explicitly trained on the entity-level prediction task. We formalize the entity-level reasoning prediction task as a relation extraction task (Zhang and Wang, 2015). The input is a pair of entities, and the output is the relationship between two entities. From all named entities obtained by using the NER models, we generate a set of entity pairs; for example, given N entities, we obtain $N \times (N - 1)$ pairs. For each pair, we predict a relationship in a set of predefined relationships obtained from the training set. We then use cross-entropy as the learning objective.

4   Datasets and Evaluation Metrics

We mainly experiment with 2Wiki and HotpotQA-small. We also train and evaluate our model on the full version of HotpotQA. We reuse and create debiased and adversarial sets for the evaluation. Table 1 presents the statistics for 2Wiki, HotpotQA-small, and additional evaluation sets. The details of HotpotQA and 2Wiki are presented in Appendix B.2. It should be noted that all datasets are in English.

4.1   HotpotQA-small

R^4C (Inoue et al., 2020) is created by adding entity-level reasoning information to the samples

Split	2Wiki	HotpotQA-small
Train	167,454	3,671
Dev.	12,576	917
Test	12,576	-
Debiased	12,576 (x4)	917 (x4)
Adversarial	12,576	659 & 134

Table 1: Statistics for 2Wiki and HotpotQA-small. There are four debiased sets in 2Wiki and HotpotQA-small. There are one adversarial set in 2Wiki and two adversarial sets in HotpotQA-small.

in HotpotQA. We obtain HotpotQA-small by combining HotpotQA (Yang et al., 2018) with R^4C. HotpotQA-small comprises three tasks as in 2Wiki: (1) sentence-level SFs prediction, (2) entity-level prediction, and (3) answer prediction. First, we re-split the ratio between the training and dev. sets; the new sizes are 3,671 and 917 for the training and dev. sets, respectively (the original sizes are 2,379 and 2,209, respectively). In R^4C, there are three gold annotations for the entity-level prediction task; in 2Wiki, there is only one gold annotation. For consistency in the evaluation and analysis, we randomly choose one annotation from the three annotations for every sample in R^4C.

The entity-level reasoning in R^4C is created by crowdsourcing. We observe that there are many similar relations in the triples in R^4C, and these relations can be grouped into one. For example, *is in*, *is located in*, *is in the*, and *is located in the* indicate location relation. We also group the relations by removing the context information in the relations; for example, *is a 2015 book by* and *is the second book by* are considered similar to the relation *is a book by*. After grouping, the number of relations in R^4C is 2,526 (it is 4,791 before).

4.2 Debiased Dataset

The objective of our debiased dataset is to introduce a small perturbation in each paragraph to mitigate a specific type of bias, in our case, the position bias shown in Figure 2. For both 2Wiki and HotpotQA-small, we use the same method to generate four debiased sets: ADDUNRELATED, ADDRELATED, ADD2, and ADD2SWAP. The differences between these four sets are whether the sentence is related or unrelated to the paragraph and whether we add one or two sentences into the paragraph. The details of each set are as follows.

ADDUNRELATED: One sentence unrelated to the paragraph is added. In our experiment, we use a list of sentences in the sentence-level revision dataset (Tan and Lee, 2014). We randomly choose one sentence that has a number of tokens greater than eleven but less than twenty-one.

ADDRELATED: One sentence that does not have an impact on the meaning or flow of the paragraph is added. In our experiment, we write multiple templates for each entity type (e.g., for a film entity, "#Name is a nice film", where *#Name* is the title of the paragraph), then randomly choose one template, and add it to the paragraph. To detect the type of the paragraph, we use the question type information in 2Wiki and HotpotQA-small, the results of the NER model, and the important keywords in the question (e.g., who, magazine, album, and film).

ADD2: ADDRELATED and ADDUNRELATED are combined in order.

ADD2SWAP: The order of ADDRELATED and ADDUNRELATED in ADD2 is swapped.

4.3 Adversarial Dataset

The objective of our adversarial dataset is to check the robustness of the model by asking modified versions of questions. For HotpotQA-small, we reuse two versions of adversarial examples in Geva et al. (2022). The first one is automatically generated by using the 'Break, Perturb, Build' (BPB) framework in Geva et al. (2022). The BPB framework performs three main steps: (1) breaking a question into multiple reasoning steps, (2) perturbing the reasoning steps by using a list of defined rules, and (3) building new QA samples from the perturbations in step #2. The second version is a subset of the first version and is validated by crowd workers. We only use the examples in these two versions that the original examples appear in HotpotQA-small.

For 2Wiki, no adversarial dataset is available. Based on the idea of the BPB framework in Geva et al. (2022), we apply two main rules from BPB for 2Wiki: (1) replace the comparison operation for comparison questions, and (2) use the prune step for bridge questions. For the first rule, we replace the operation in the comparison questions (e.g., "Who was born first, A or B?" is converted to "Who was born later, A or B?"). For the second rule, we use a sub-question in the QA process as the main question (e.g., for Figure 1, we ask, "Who is the father of Joan of Valois?").

Dataset	Task Setting	Answer		Sentence-level		Entity-level	
		EM	F1	EM	F1	EM	F1
2Wiki	(1) Ans	72.03	77.87	-	-	-	-
	(2) Ans + Sent	72.82	78.65	78.06	92.38	-	-
	(3) Ans + Ent	72.33	78.21	-	-	46.11	76.65
	(4) Ans + Sent + Ent	**73.60**	**79.37**	78.46	92.68	45.97	76.69
HotpotQA-small	(1) Ans	52.89	66.43	-	-	-	-
	(2) Ans + Sent	54.42	69.03	75.35	91.00		-
	(3) Ans + Ent	54.74	69.08	-	-	6.54	31.31
	(4) Ans + Sent + Ent	**54.74**	**69.44**	75.14	90.88	6.43	31.05

Table 2: Ablation study results (%) of our model in the dev. sets of 2Wiki and HotpotQA-small. *Ans*, *Sent*, and *Ent* represent the answer prediction task, sentence-level SFs prediction task, and entity-level prediction task, respectively. 'Task Setting' represents the tasks that the model is trained on. '-' indicates the tasks the model is not trained on.

4.4 Evaluation Metrics

Each task in HotpotQA and 2Wiki is evaluated by using two metrics: exact match (EM) and F1 score. Following the evaluation script in HotpotQA and 2Wiki, we use joint EM and joint F1 to evaluate the entire capacity of the model. For HotpotQA, they are the products of the scores of two tasks: sentence-level prediction and answer prediction. For 2Wiki and HotpotQA-small, they are the products of the scores of three tasks: sentence-level prediction, entity-level prediction, and answer prediction.

5 Results

Currently, there are no existing end-to-end models that explicitly train all three tasks together; therefore, in this study, we use our proposed model for analysis. We also compare our model with other previous models on the HotpotQA and 2Wiki datasets. In general, the experimental results indicate that our model is comparable to previous models and can be used for further analyses. We focus more on the analysis; therefore, the detailed results of the comparison are presented in Appendix B.3.

5.1 Effectiveness of the UR Tasks

To investigate the effectiveness of the UR tasks, we train the model in four settings: (1) answer prediction only, (2) answer prediction and sentence-level SFs prediction, (3) answer prediction and entity-level prediction, and (4) all three tasks together.

QA Performance (RQ1) Our first research question is whether the UR tasks can improve QA performance. To answer this question, we compare the

results of different task settings described above. The results are presented in Table 2. For 2Wiki, using sentence-level and entity-level separately (settings #2 and #3), the QA performance does not change significantly. The improvement is significant when we combine both the sentence-level and entity-level (setting #4). Specifically, the scores when the model is trained on the answer prediction task only (setting #1) and on both the answer prediction task and UR tasks (setting #4) are 77.9 and 79.4 F1, respectively. In contrast to 2Wiki, using sentence-level and entity-level separately, there is a larger QA performance improvement in HotpotQA-small. Specifically, the F1 scores of settings #2 and #3 are 69.0 and 69.1, respectively, whereas, the F1 score of the first setting is 66.4. Similar to 2Wiki, there is a large gap between the two settings, #1 and #4 (66.4 F1 and 69.4 F1, respectively).

In summary, these results indicate that both sentence-level and entity-level prediction tasks contribute to improving QA performance. These results align with the findings in Yang et al. (2018), which shows that incorporating the sentence-level SFs prediction task can improve QA performance. We also find that when combining both sentence-level and entity-level prediction tasks, the scores of the answer prediction task are the highest.

Reasoning Shortcuts (RQ2) To investigate whether explicitly optimizing the model on the UR tasks can prevent reasoning shortcuts, we evaluate the four settings of the model on the four debiased sets of 2Wiki and HotpotQA-small. The generation of the debiased sets includes stochastic steps. To minimize the impact of randomness on our re-

Dataset	Task Setting	Reduction (%) on Four Debiased Sets							
		ADDUNRELATED		ADDRELATED		ADD2		ADD2SWAP	
		EM	F1	EM	F1	EM	F1	EM	F1
2Wiki	(1) Ans	13.40	12.13	3.55	3.46	12.32	11.72	18.99	17.51
	(2) Ans + Sent	11.00	9.71	4.16	4.22	11.22	10.69	17.62	16.24
	(3) Ans + Ent	**7.73**	**6.94**	**2.80**	**2.77**	8.38	7.76	**13.12**	**12.21**
	(4) Ans + Sent + Ent	8.86	8.11	3.16	3.13	9.09	8.58	14.53	13.77
HotpotQA-small	(1) Ans	3.01	1.53	4.04	1.50	1.65	1.01	3.96	2.47
	(2) Ans + Sent	**1.13**	**1.35**	-0.51	0.19	**0.08**	**0.85**	**1.77**	**1.96**
	(3) Ans + Ent	6.73	5.60	**-0.92**	**0.03**	4.02	3.54	6.89	5.46
	(4) Ans + Sent + Ent	5.05	4.65	1.26	1.25	1.83	2.46	3.58	3.64

Table 3: Average performance drop from five times running (smaller is better) of the four settings on the four debiased sets of 2Wiki and HotpotQA-small. The best and worst scores are boldfaced and underlined, respectively.

ported results, we generate the debiased sets five times and report the average evaluation scores. The average performance drops are presented in Table 3 (detailed scores are given in Appendix B.4).

Overall, for 2Wiki, when the model is trained on only one task (#1), the drop is the largest (except for ADDRELATED, which is the second largest). When the model is trained only on the answer prediction task, the drops are always higher than those when the model is trained on three tasks. Specifically, the gaps between the two settings, #1 (only answer task) and #4 (all three tasks), are 4.5%, 0.4%, 3.2%, 4.5% (EM score) for ADDUNRELATED, AD-DRELATED, ADD2, and ADD2SWAP, respectively. These scores indicate that the two tasks, sentence-level and entity-level, positively affect the answer prediction task when the model is trained on three tasks simultaneously.

For HotpotQA-small, we observe that the effectiveness of the UR tasks is inconsistent. For example, for ADDUNRELATED, when training the model on the three tasks (setting #4), the reduction is larger than that when training on answer task only (setting #1) (5.1 vs. 3.0 EM). However, for ADDRELATED, the reduction on setting #4 is smaller than that on setting #1 (1.3 vs. 4.0 EM). One possible reason is that the performance of the entity-level task is not good (6.4 EM), which affects the answer prediction task when the model is trained on the three tasks together. Another possible reason is that the position bias in HotpotQA-small is not sufficiently large. We present a detailed analysis in Section 5.2 to explain this case.

Robustness (RQ3) To test whether the UR tasks can help to improve the robustness of the model,

Task Setting	Dev-adver		Reduction %	
	EM	F1	EM	F1
Ans	37.09	46.07	**48.51**	**40.84**
Ans + Sent	34.26	43.64	52.95	44.51
Ans + Ent	32.67	39.43	54.83	49.58
Ans + Sent + Ent	34.19	42.74	53.55	46.15

Table 4: Results of our model in the dev-adversarial set of 2Wiki and the performance drop.

we evaluate the four settings of the model on the adversarial sets. For 2Wiki, the results are presented in Table 4. The scores for all four settings decrease significantly on the adversarial set. The reduction is the smallest when the model is trained on the answer task only. The UR tasks do not make the model more robust on this adversarial set. For HotpotQA-small, we observe the same behavior, that is, when the model is trained on the answer task only, the reduction is the smallest. All results are presented in Table 5. These results indicate that both sentence-level and entity-level prediction tasks do not contribute to improving the robustness of the models on adversarial questions, such as sub-questions and inverted questions. We analyze the results in Section 5.2.

5.2 Analyses

Details of RQ2 To investigate the results concerning RQ2 in more depth, we first analyze the position biases of different types of questions in 2Wiki and HotpotQA-small. We find that the comparison questions have more position biases than the bridge questions in both 2Wiki and HotpotQA-

Task Setting	Dev		Dev-Adver		Adver↓ (%)		Dev-Adver-val		Adver-val↓ (%)	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
(1) Ans	52.89	66.43	40.36	51.23	23.69	**22.88**	37.31	46.69	**29.46**	**29.72**
(2) Ans + Sent	54.42	69.03	41.73	52.50	23.32	23.95	34.33	43.86	36.92	36.46
(3) Ans + Ent	54.74	69.08	42.79	52.16	**21.83**	24.49	27.61	36.86	49.56	46.64
(4) Ans + Sent + Ent	54.74	69.44	40.52	51.14	25.98	26.35	31.34	38.22	42.75	44.96

Table 5: Results of our model in the dev. and two dev-adversarial sets of HotpotQA-small. 'Adver' denotes adversarial and 'Adver-val' denotes the adversarial set that was validated by crowd workers.

Task Setting	Correct Ans	Correct Ent	Correct Both Ans & Ent
(3) Ans + Ent	4,109	6,851	2,249 (32.8%)
(4) Ans + Sent + Ent	4,300	6,450	2,420 (37.5%)

Table 6: Number of correct predicted answers, number of correct predicted entity-level reasoning, and number of examples that have both correct predicted answers and correct predicted entity-level reasoning.

small (Appendix B.5). To evaluate the effectiveness of the position bias for each type of question, we evaluate the four settings of the model on the four debiased sets for each type of question in both datasets. All the results are presented in Appendix B.5.

For 2Wiki, we find that most of the answers are in the first sentences in the comparison questions. This large bias is the main reason for the significant reduction in the scores in the comparison questions. 2Wiki has 46.0% of comparison questions. The reduction in comparison questions contributes to the reduction in the entire dataset. In other words, the results of 2Wiki are affected by those of the comparison questions. HotpotQA-small has only 22.0% of comparison questions, and the position bias in the comparison questions was not sufficiently large. Therefore, the position bias does not have a significant impact on the main QA task. In other words, the UR tasks do not have a significant effect.

Details of RQ3 The adversarial questions used in RQ3 are the sub-questions in the QA process for bridge questions and the inverted questions for comparison questions. We observe that the triple in the entity-level task is helpful in answering the sub-questions. For example, the triple is: (*Charles of Valois, father, Philip III of France*) and the sub-question is "*Who is the father of Charles of Valois?*". To understand more on the behaviors of the model, we analyze the results from 2Wiki in two

settings: (3) Ans + Ent and (4) Ans + Sent + Ent. Table 6 presents the detailed results for these two settings. We find that correct reconstruction of the entity-level reasoning task contributes to finding the correct answer only in 32.8% of cases in setting #3 and only in 37.5% of cases in setting #4. Entity-level reasoning in the form of triples has no significant effect on the main QA process. Several examples are presented in Appendix B.5.

We conjecture that there are three possible reasons why the UR tasks cannot contribute to the adversarial dataset. The first one is the difference in the form and design of the tasks. Specifically, the entity-level reasoning task is formulated as a relation extraction task; the input is a pair of entities, and the output is a relation label. Meanwhile, the adversarial dataset is formulated as a QA task; the input is a natural language question, and the output is an answer. The second reason is the incompetence of the entity-level reasoning information. As discussed in Ho et al. (2022), the entity-level reasoning in the comparison questions does not describe the full path from question to answer, and other reasoning operations are required to obtain the answer. The final reason is the manner in which we utilize the entity-level reasoning information. Our model does not consider the order of the triples in the reasoning chain. For example, we do not consider the order of the two steps in Figure 1b. We hope that our research will inspire future studies to investigate the effectiveness of the UR tasks in the form of a natural language question, which has the same form as a multi-hop QA question.

6 Related Work

Multi-hop Datasets and Analyses To test the reasoning abilities of the models, many multi-hop QA datasets (Welbl et al., 2018; Talmor and Berant, 2018; Yang et al., 2018) have been proposed. Recently, Trivedi et al. (2022) introduced MuSiQue, a

multi-hop dataset constructed from a composition of single-hop questions. The reason why do we not conduct experiments on MuSiQue is explained in the limitations section.

In addition to Tang et al. (2021) and Ho et al. (2022), the most similar to our research mentioned in the Introduction, there are some other existing studies (Chen and Durrett, 2019; Jiang and Bansal, 2019a; Min et al., 2019a; Trivedi et al., 2020) on the analysis and investigation of the multi-hop datasets and models. However, most of them do not utilize the internal reasoning information when answering questions.

Multi-hop Models Various directions have been proposed for solving multi-hop datasets, including question decomposition (Talmor and Berant, 2018; Jiang and Bansal, 2019b; Min et al., 2019b; Perez et al., 2020; Wolfson et al., 2020; Fu et al., 2021a), iterative retrieval (Feldman and El-Yaniv, 2019; Asai et al., 2020; Qi et al., 2021), graph neural networks (Song et al., 2018; De Cao et al., 2019; Ding et al., 2019; Qiu et al., 2019; Tu et al., 2019; Fang et al., 2020), and other approaches such as single-hop based models (Yang et al., 2018; Nishida et al., 2019) or transformer-based models (Devlin et al., 2019; Zaheer et al., 2020). Our model is based on the BigBird transformer model.

Other QA Reasoning Datasets In addition to multi-hop reasoning datasets, several other existing datasets also aim to evaluate the reasoning abilities of the models. Some of them are: DROP (Dua et al., 2019) for numerical reasoning; CLUTRR (Sinha et al., 2019), ReClor (Yu et al., 2020), and LogiQA (Liu et al., 2020) for logical reasoning; Quoref (Dasigi et al., 2019) for coreference reasoning; CommonsenseQA (Talmor et al., 2019), MCScript2.0 (Ostermann et al., 2019), and CosmosQA (Huang et al., 2019) for commonsense reasoning. Many of these datasets consist of only a single paragraph in the input or lack explanation information that describes the reasoning process from question to answer. However, our focus is on multi-hop reasoning datasets that contain multiple paragraphs in the input and provide explanatory information for the QA process.

7   Conclusion

We analyze the effectiveness of the underlying reasoning tasks using two multi-hop datasets: 2Wiki and HotpotQA-small. The results reveal that the underlying reasoning tasks can improve QA performance. Using four debiased sets, we demonstrate that the underlying reasoning tasks can reduce the reasoning shortcuts of the QA task. The results also reveal that the underlying reasoning tasks do not make the models more robust on adversarial examples, such as sub-questions and inverted questions. We encourage future studies to investigate the effectiveness of the entity-level reasoning task in the form of sub-questions.

Limitations

Our study has two main limitations. The first one is the small size of HotpotQA-small. Currently, there are no other multi-hop datasets that contain a large number of examples with the entity-level reasoning prediction task. MuSiQue is the most potential option. The entity-level reasoning information in MuSiQue includes two types of formats: triple format and natural language question format. We do not experiment with MuSiQue because the number of examples that have entity-level reasoning information in the form of a triple is small: 2,253 out of 19,938 in the training set and 212 out of 2,417 in the dev. set.

The second limitation is that our model does not consider the order of the triples in the entity-level reasoning prediction task. As shown in Figure 1b, the two triples are ordered. However, our model formulizes the entity-level prediction task as a relation extraction task. We predict a relation given the two entities detected by the NER models. Therefore, the order of the triples is not considered. We conjecture that this may be one of the reasons why the entity-level reasoning prediction task (e.g., a triple *(Film A, director, D)*) does not support the model when answering sub-questions (e.g., *Who is the director of Film A?*) using the same information.

Acknowledgments

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019.

FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *International Conference on Learning Representations*.

Jifan Chen and Greg Durrett. 2019. Understanding dataset design choices for multi-hop reasoning. 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 4026–4032, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. 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 5925–5932, Hong Kong, China. Association for Computational Linguistics.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2019. Question answering by reasoning across documents with graph convolutional networks. 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 2306–2317, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. 2019. Cognitive graph for multi-hop reading comprehension at scale. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2694–2703, Florence, Italy. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2020. Hierarchical graph network for multi-hop question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8823–8838, Online. Association for Computational Linguistics.

Yair Feldman and Ran El-Yaniv. 2019. Multi-hop paragraph retrieval for open-domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2296–2309, Florence, Italy. Association for Computational Linguistics.

Ruiliu Fu, Han Wang, Xuejun Zhang, Jun Zhou, and Yonghong Yan. 2021a. Decomposing complex questions makes multi-hop QA easier and more interpretable. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 169–180, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ruiliu Fu, Han Wang, Jun Zhou, and Xuejun Zhang. 2021b. Na-reviewer: Reviewing the context to improve the error accumulation issue for multi-hop qa. *Electronics Letters*.

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. In *Transactions of the Association for Computational Linguistics (TACL)*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xanh Ho, Saku Sugawara, and Akiko Aizawa. 2022. How well do multi-hop reading comprehension models understand date information? In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 470–479, Online only. Association for Computational Linguistics.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. 2020. R4C: A benchmark for evaluating RC systems to get the right answer for the right reason. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6740–6750, Online. Association for Computational Linguistics.

Yichen Jiang and Mohit Bansal. 2019a. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2726–2736, Florence, Italy. Association for Computational Linguistics.

Yichen Jiang and Mohit Bansal. 2019b. Self-assembling modular networks for interpretable multi-hop reasoning. 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 4474–4484, Hong Kong, China. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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, Online. Association for Computational Linguistics.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. LogiQA: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3622–3628. International Joint Conferences on Artificial Intelligence Organization. Main track.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019a. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4249–4257, Florence, Italy. Association for Computational Linguistics.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019b. Multi-hop reading comprehension through question decomposition and rescoring. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6097–6109, Florence, Italy. Association for Computational Linguistics.

Kosuke Nishida, Kyosuke Nishida, Masaaki Nagata, Atsushi Otsuka, Itsumi Saito, Hisako Asano, and Junji Tomita. 2019. Answering while summarizing: Multi-task learning for multi-hop QA with evidence extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2335–2345, Florence, Italy. Association for Computational Linguistics.

Simon Ostermann, Michael Roth, and Manfred Pinkal. 2019. MCScript2.0: A machine comprehension corpus focused on script events and participants. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 103–117, Minneapolis, Minnesota. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS Autodiff Workshop*.

Ethan Perez, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. 2020. Unsupervised question decomposition for question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8864–8880, Online. Association for Computational Linguistics.

Peng Qi, Haejun Lee, Tg Sido, and Christopher Manning. 2021. Answering open-domain questions of varying reasoning steps from text. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3599–3614, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. CLUTRR: A diagnostic benchmark for inductive reasoning from text. 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 4506–4515, Hong Kong, China. Association for Computational Linguistics.

Linfeng Song, Zhiguo Wang, Mo Yu, Yue Zhang, Radu Florian, and Daniel Gildea. 2018. Exploring graph-structured passage representation for multi-hop reading comprehension with graph neural networks. *arXiv:1809.02040*.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Chenhao Tan and Lillian Lee. 2014. A corpus of sentence-level revisions in academic writing: A step towards understanding statement strength in communication. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 403–408, Baltimore, Maryland. Association for Computational Linguistics.

Yixuan Tang, Hwee Tou Ng, and Anthony Tung. 2021. Do multi-hop question answering systems know how to answer the single-hop sub-questions? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3244–3249, Online. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2020. Is multihop QA in DiRe condition? measuring and reducing disconnected reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8846–8863, Online. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10(0):539–554.

Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. 2020. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9073–9080. AAAI Press.

Ming Tu, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, and Bowen Zhou. 2019. Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2704–2713, Florence, Italy. Association for Computational Linguistics.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022. Measure and improve robustness in NLP models: A survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4569–4586, Seattle, United States. Association for Computational Linguistics.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan Berant. 2020. Break it down: A question understanding benchmark. *Transactions of the Association for Computational Linguistics*, 8:183–198.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc.

Dongxu Zhang and Dong Wang. 2015. Relation classification via recurrent neural network. *arXiv:1508.01006*.

A   Word Overlap Shortcut

Using adversarial methods, Jiang and Bansal (2019a) show that examples in HotpotQA often contain word overlap shortcut, where the models can answer the questions by performing word-matching between the question and a sentence in the context.

1174

Based on this finding, we automatically calculate the word overlap shortcut for 2Wiki and HotpotQA-small. We observe that the word overlap shortcut is common in bridge questions; therefore, we only calculate the word overlap shortcut for bridge questions in 2Wiki and HotpotQA-small. To check whether a sample contains the word overlap shortcut, we do the following steps:

- Obtain a set of surrounding words S by getting the five words immediately to the left and right of the answer span, then remove stopwords in S.

- Obtain a set of overlapping words (O) between S and a question.

- We consider a sample containing the word overlap shortcut if there are at least two words in O and $\frac{|O|}{|S|} \geq 0.65$. These numbers (threshold) are chosen based on the evaluation of 40 examples that are manually annotated by the authors.

We find that there are 56 out of 5,791 and 151 out of 715 examples (5,791 and 715 are the numbers of bridge questions in 2Wiki and HotpotQA-small) in the dev. sets of 2Wiki and HotpotQA-small containing the word overlap shortcut.

It is noted that there is another type of shortcut, namely, entity-type matching shortcut. Based on the experimental results and human performance, Min et al. (2019a) reveal that examples in HotpotQA contain the entity type matching shortcut, where the models can answer the questions by using the first five tokens in the questions; meanwhile, humans can answer the questions by using the entity type of the paragraphs. Currently, there is no dataset that can prevent the entity-type shortcut; therefore, we do not use this type of shortcut in our experiments.

B Experimental Details

B.1 Implementation Details

We use Pytorch (Paszke et al., 2017) and Hugging Face when building our model. For the context encoding step, we use a pre-trained BigBird model as the encoder; the hidden dimension is 768. For the entity-level reasoning prediction task, we obtain 33 relations for 2Wiki and 2,526 relations for $\mathrm{R}^4\mathrm{C}$, from all triples in the training set, including a non-relation type. We use entity type embedding d_t of

50. We fine-tuned our model with a total batch size of 32 on a single GPU (NVIDIA A100 80GB) using mixed precision and a gradient accumulation step of 8. Following the hyperparameters in the BERT model (Devlin et al., 2019), for optimization, we use the Adam Opitmizer (Kingma and Ba, 2015) with a learning rate of 3e-5, weight decay of 0.01, learning rate warmup over the first 10% of the total number of training steps, and linear decay of the learning rate. We also use a dropout probability of 0.1 on all layers.

For multi-task prediction, we use λ_{sent} as 4, λ_{ent} as 15, and λ_{ans} as 1 for 2Wiki and HotpotQA-small; we use λ_{sent} as 7 and λ_{ans} as 1 for HotpotQA. We do not run all experiments with different values of λ_{sent}, λ_{ent}, and λ_{ans}; instead we run several experiments, base on the results, we then adjust the parameters. We find that when running with λ_{sent} as 4 for 2Wiki and 7 for HotpotQA, λ_{ent} as 15, and λ_{ans} as 1, it produces the best results. We fix the random seed for the reproducibility of the results. We observe that the final epoch often produces the best scores, and its scores are stable on adversarial datasets; therefore, we choose the final epoch for all settings in our experiment.

B.2 Datasets

HotpotQA HotpotQA was created by crowdsourcing. Due to the design of the dataset, there are only two tasks in HotpotQA: sentence-level SFs prediction and answer prediction. $\mathrm{R}^4\mathrm{C}$ was created based on HotpotQA and contained 4,588 questions. The dataset requires systems to provide an answer and derivation in a semi-structured natural language form. There are two types of questions in HotpotQA: bridge and comparison.

2Wiki 2Wiki was constructed by utilizing a Knowledge Base and Wikipedia, and the questions were created by using templates. There are three different tasks in the dataset: (1) sentence-level SFs prediction, (2) evidence generation (for consistency, we use the term *entity-level prediction*), and (3) answer prediction. The context consists of ten paragraphs, including two or four gold paragraphs and eight or six distractor paragraphs. The gold paragraph contains the information required to find the answer. Meanwhile, the purpose of the distractor paragraph is to distract the models. There are four different types of questions in the dataset: comparison, inference, compositional, and bridge-comparison. Inference and compositional

Figure 4: Example of a comparison question from the 2Wiki dataset.

questions are two sub-types of the bridge question. For the convenience of analysis, we consider comparison and bridge-comparison questions as comparison questions. Figure 4 presents an example of a comparison question from the 2Wiki dataset.

2Wiki was designed to focus on the entire reasoning process from question to answer. The entire capacity of the model is evaluated by using two metrics: joint EM and joint F1. To obtain the joint F1 score, they first calculate the joint precision and joint recall as follows: $P^{joint} = P^{ans} P^{ent} P^{sent}$ and $R^{joint} = R^{ans} R^{ent} R^{sent}$. (P^{ans}, R^{ans}), (P^{ent}, R^{ent}), (P^{sent}, R^{sent}) represent the precision and recall for three tasks: answer prediction, entity-level reasoning prediction, and sentence-level SFs prediction. The joint EM is 1 when all three tasks achieve an exact match and otherwise 0.

B.3 Results Comparison

We compare our results with three previous models: BiDAF, CRERC, and NA-Reviewer. BiDAF is a baseline model in Ho et al. (2020). CRERC (Fu et al., 2021a) is a pipeline model that includes three modules: relation extractor, reader, and comparator. NA-Reviewer (Fu et al., 2021b) is an improved version of CRERC, as it addresses the error accumulation issue. It is noted that both CRERC and NA-Reviewer models are evaluated on only 2Wiki.

Table 7 presents the results of our model and previous models in the dev. set of HotpotQA and in the test set of 2Wiki. It also shows the performance of our model in the dev. set of HotpotQA-small and human performance in Ho et al. (2020).

Results on HotpotQA Our score is comparable to the BERT-base version of two strong models, SAE (Tu et al., 2020) and HGN (Fang et al., 2020)

in the dev. set of the distractor setting in HotpotQA. Specifically, our joint F1 is 67.8, while for SAE-BERT, it is 66.5, and for HGN-BERT, it is 66.9. However, our score is smaller than the RoBERTa-base of SAE and HGN. They are 72.8 and 74.4 F1 for SAE-RoBERTa and HGN-RoBERTa, respectively. It is noted that we use the BigBird-ITC version in our model. Although the BigBird-ETC version performs better than the BigBird-ITC version, it is not available in Hugging Face. We do not use SAE and HGN for our analyses because these models are not designed to train on the entity-level reasoning prediction task.

Results on HotpotQA-small The scores on HotpotQA-small are lower than those on HotpotQA in the answer prediction task. This result may be explained by the fact that the training size of HotpotQA-small is smaller than HotpotQA (3,671 vs. 90,564). Due to the small size, we only use the gold paragraphs for experiments. That is why the scores on HotpotQA-small are higher than those on HotpotQA in the sentence-level task. For the entity-level task, the EM score is quite low (6.4 EM). A possible reason for this is that there are many relations in HotpotQA-small (2,526 relations); meanwhile, there are only 33 relations in 2Wiki. We observe that the F1 score (31.1 F1) is much better than the EM score. Therefore, we keep using HotpotQA-small for analyses.

Results on 2Wiki Our model significantly outperforms BiDAF in all tasks. Our results are comparable to CRERC. The EM score of our model in the entity-level task is lower than that of CRERC. A possible explanation for this might be that the relation extractor module in CRERC is fine-tuned on 2Wiki; therefore, it can extract entities better than the NER models from Spacy and Flair that are used in our model. However, the F1 score of our model in the entity-level task is higher than that of CRERC. This indicates that our model can correctly obtain a few triples in a set of gold triples for many samples. All our scores (except the F1 score of the entity-level task) are lower than those on NA-Reviewer. Our target is to analyze the UR tasks in an end-to-end model. Although the pipeline models (CRERC and NA-Reviewer) are easy to interpret, we cannot determine how the UR tasks affect answer prediction in an end-to-end model. Therefore, we use the design of our model to perform the analyses in this study.

Dataset	Model	Answer		Sentence-level		Entity-level		Joint	
		EM	F1	EM	F1	EM	F1	EM	F1
HotpotQA	HGN-BERT[‡] (Fang et al., 2020)	*N/A*	74.76	*N/A*	86.61	✗	✗	*N/A*	66.90
	HGN-RoBERTa (Fang et al., 2020)	**68.93**	**82.18**	63.09	**88.59**	✗	✗	46.46	**74.34**
	SAE-BERT (Tu et al., 2020)	61.32	74.81	58.06	85.27	✗	✗	39.89	66.45
	SAE-RoBERTa (Tu et al., 2020)	67.70	80.75	**63.30**	87.38	✗	✗	**46.81**	72.75
	Our BigBird-base	61.90	76.09	58.54	86.93	✗	✗	39.39	67.81
HotpotQA-small	Our BigBird-base	54.74	69.44	75.14	90.88	6.43	31.05	4.25	21.69
2Wiki	BiDAF (Ho et al., 2020)	36.53	43.93	24.99	65.26	1.07	14.94	0.35	5.41
	CRERC (Fu et al., 2021a)	69.58	72.33	82.86	90.68	**54.86**	68.83	49.80	58.99
	NA-Reviewer (Fu et al., 2021b)	**76.73**	**81.91**	**89.61**	**94.31**	53.66	70.83	**52.75**	**65.23**
	Our BigBird-base	74.05	79.68	77.14	92.13	45.75	**76.64**	39.30	63.24
	Human UB (Ho et al., 2020)	91.00	91.79	88.00	93.75	64.00	78.81	62.00	75.25

Table 7: Results (%) of our model and previous models in the dev. set of HotpotQA and in the test set of 2Wiki. We also show the performance of our model in the dev. set of HotpotQA-small. *Answer*, *Sentence-level*, and *Entity-level* represent the answer prediction task, sentence-level prediction task, and entity-level prediction task, respectively. For HGN-BERT, the scores that we obtained (from left to right: 58.93 73.18 54.64 85.34 35.11 64.24) are lower than the reported scores in HGN (Fang et al., 2020); therefore, we show the reported F1 scores in HGN.

B.4 Effectiveness of the UR Tasks

Reasoning Shortcuts (RQ2) Table 8 presents the performance drop (smaller is better) for five times running of the four settings of the model on the four debiased sets of 2Wiki and HotpotQA-small. As depicted in the table, for 2Wiki, the gap between two settings #1 (answer prediction task only) and #4 (all three tasks) is consistent in all five times running. Meanwhile, for HotpotQA-small, the gap between two settings #1 (answer prediction task only) and #4 (all three tasks) is inconsistent in all five times running. This observation supports our explanation in Section 5.2 that the position bias in HotpotQA-small does not have a large impact on the main QA task.

B.5 Analyses

Details of RQ2 Figure 5 illustrates the information on the position of sentence-level SFs of comparison and bridge questions in the dev. sets of the two datasets: 2Wiki and HotpotQA-small. As shown in the Figure, the comparison questions have more position biases than the bridge questions in both 2Wiki and HotpotQA-small. Furthermore, we observe that the position bias in the comparison questions in HotpotQA-small is smaller than that in 2Wiki.

Table 9 presents the performance drop for two types of questions, comparison and bridge questions, in 2Wiki and HotpotQA-small.
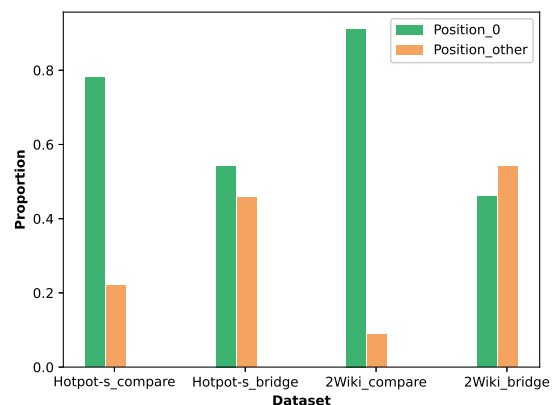

Figure 5: Information on the position of sentence-level SFs of comparison and bridge questions in the dev. sets of the two datasets: 2Wiki and HotpotQA-small.

Details of RQ3 Table 10 presents examples of the outputs predicted by our model, which is trained on three tasks simultaneously.

Dataset	Task Setting	Reduction (%)									
		Time #1		Time #2		Time #3		Time #4		Time #5	
		EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
2Wiki											
ADDUNRELATED	Ans	<u>13.26</u>	<u>12.11</u>	<u>13.06</u>	<u>11.87</u>	<u>13.59</u>	<u>12.15</u>	<u>13.31</u>	<u>12.08</u>	<u>13.79</u>	<u>12.44</u>
	Ans + Sent	10.82	9.56	10.94	9.68	10.93	9.83	11.15	9.83	11.16	9.66
	Ans + Ent	**8.09**	**7.15**	**7.77**	**6.98**	**7.71**	**6.99**	**7.78**	**7.08**	**7.31**	**6.52**
	Ans + Sent + Ent	8.41	7.80	8.85	7.99	9.10	8.33	8.97	8.23	8.97	8.18
ADDRELATED	Ans	3.72	3.61	3.61	3.57	3.54	3.39	3.29	3.26	3.57	3.49
	Ans + Sent	<u>4.22</u>	<u>4.23</u>	<u>4.44</u>	<u>4.44</u>	<u>4.11</u>	<u>4.18</u>	<u>4.08</u>	<u>4.20</u>	<u>3.94</u>	<u>4.04</u>
	Ans + Ent	**2.61**	**2.62**	**2.89**	**2.88**	**2.85**	**2.83**	**2.89**	**2.81**	**2.75**	**2.71**
	Ans + Sent + Ent	3.18	3.12	3.18	3.14	3.06	3.06	3.23	3.25	3.14	3.07
ADD2	Ans	<u>12.26</u>	<u>11.63</u>	<u>12.59</u>	<u>12.10</u>	<u>12.29</u>	<u>11.69</u>	<u>12.29</u>	<u>11.72</u>	<u>12.16</u>	<u>11.48</u>
	Ans + Sent	11.10	10.48	11.03	10.57	11.26	10.77	11.18	10.72	11.51	10.92
	Ans + Ent	**8.38**	**7.91**	**8.74**	**8.12**	**8.31**	**7.63**	**8.07**	**7.48**	**8.41**	**7.67**
	Ans + Sent + Ent	9.13	8.54	8.90	8.45	8.94	8.54	8.95	8.43	9.51	8.92
ADD2SWAP	Ans	<u>19.06</u>	<u>17.61</u>	<u>18.87</u>	<u>17.40</u>	<u>19.20</u>	<u>17.59</u>	<u>18.80</u>	<u>17.31</u>	<u>19.03</u>	<u>17.63</u>
	Ans + Sent	17.71	16.16	17.73	16.40	17.74	16.41	17.34	15.99	17.59	16.25
	Ans + Ent	**13.02**	**12.19**	**13.09**	**12.19**	**13.38**	**12.30**	**13.13**	**12.30**	**12.97**	**12.07**
	Ans + Sent + Ent	14.28	13.56	14.31	13.70	14.18	13.41	14.89	13.99	15.00	14.17
HotpotQA-small											
ADDUNRELATED	Ans	4.33	**2.89**	1.44	0.68	0.21	**-0.45**	4.33	2.66	4.75	1.85
	Ans + Sent	**4.01**	3.65	**0.81**	**1.07**	-0.20	0.88	**1.01**	**0.85**	**0.00**	**0.28**
	Ans + Ent	6.17	4.97	5.97	3.98	<u>7.76</u>	<u>7.06</u>	6.38	<u>6.20</u>	<u>7.36</u>	<u>5.80</u>
	Ans + Sent + Ent	<u>6.76</u>	<u>5.83</u>	<u>6.76</u>	<u>5.18</u>	1.79	2.97	4.77	4.94	5.17	4.32
ADDRELATED	Ans	<u>3.71</u>	<u>1.14</u>	<u>4.12</u>	1.10	<u>4.54</u>	2.12	<u>3.91</u>	<u>1.46</u>	<u>3.91</u>	1.70
	Ans + Sent	-0.79	**0.10**	0.20	0.71	0.61	1.23	**-1.40**	**-0.59**	-1.19	**-0.52**
	Ans + Ent	**-0.80**	0.23	**-0.60**	**-0.19**	**-1.21**	**-0.13**	-0.40	0.68	**-1.61**	-0.42
	Ans + Sent + Ent	0.38	0.59	2.37	<u>1.96</u>	1.19	0.95	0.00	0.59	2.37	<u>2.17</u>
ADD2	Ans	1.04	1.01	1.04	0.17	1.04	**0.53**	1.64	0.51	3.50	2.83
	Ans + Sent	1.01	1.51	**-0.79**	**-0.19**	0.00	1.43	**0.00**	**0.32**	0.20	1.20
	Ans + Ent	<u>4.57</u>	<u>3.59</u>	2.19	2.32	<u>5.17</u>	<u>4.85</u>	<u>3.78</u>	<u>3.20</u>	4.38	3.73
	Ans + Sent + Ent	**0.00**	**0.62**	<u>3.38</u>	<u>3.63</u>	1.19	2.68	1.59	2.56	2.98	2.79
ADD2SWAP	Ans	5.16	3.55	3.10	1.40	3.71	**1.82**	3.29	2.12	4.54	3.45
	Ans + Sent	3.82	3.77	**0.81**	**1.17**	2.00	2.03	**1.62**	**1.04**	**0.61**	**1.81**
	Ans + Ent	<u>5.57</u>	<u>4.20</u>	<u>6.38</u>	<u>5.07</u>	<u>7.96</u>	<u>6.83</u>	<u>6.56</u>	<u>5.39</u>	<u>7.96</u>	<u>5.83</u>
	Ans + Sent + Ent	**3.58**	**3.54**	5.77	4.74	**1.59**	2.52	2.37	3.04	4.57	4.38

Table 8: Performance drop (smaller is better) for five times running of the four settings of the model on the four debiased sets of 2Wiki and HotpotQA-small. The best and worst scores are boldfaced and underlined, respectively. The debiased datasets are newly created for each time running.

Dataset	Task Setting	Comparison				Bridge			
		Answer		Answer↓ (%)		Answer		Answer↓ (%)	
		EM	F1	EM	F1	EM	F1	EM	F1
2Wiki									
Dev	Ans	78.98	83.74			66.10	72.85		
	Ans + Sent	79.45	84.21			67.16	73.90		
	Ans + Ent	78.86	83.60			66.75	73.61		
	Ans + Sent + Ent	80.35	85.08			67.84	74.49		
ADDUNRELATED	Ans	59.51	64.49	<u>24.65</u>	22.99	65.01	71.81	**1.65**	**1.43**
	Ans + Sent	65.55	71.11	17.50	15.56	64.42	71.14	<u>4.08</u>	<u>3.73</u>
	Ans + Ent	67.67	72.84	14.19	**12.87**	65.47	72.44	1.92	1.59
	Ans + Sent + Ent	69.38	74.01	**13.65**	13.01	65.72	72.48	3.13	2.70
ADDRELATED	Ans	73.60	78.22	<u>6.81</u>	<u>6.59</u>	65.73	72.36	**0.56**	**0.67**
	Ans + Sent	74.87	79.43	5.76	5.68	65.38	71.82	<u>2.65</u>	<u>2.81</u>
	Ans + Ent	75.57	80.17	**4.17**	**4.10**	66.06	72.75	1.03	1.17
	Ans + Sent + Ent	76.69	81.28	4.56	4.47	66.63	73.14	1.78	1.81
ADD2	Ans	61.61	65.54	<u>21.99</u>	<u>21.73</u>	64.55	71.60	2.34	1.72
	Ans + Sent	64.93	69.13	18.28	17.91	64.58	71.50	<u>3.84</u>	<u>3.25</u>
	Ans + Ent	67.16	71.43	**14.84**	**14.56**	65.51	72.52	**1.86**	**1.48**
	Ans + Sent + Ent	67.85	72.19	15.56	15.15	66.06	72.94	2.62	2.08
ADD2SWAP	Ans	51.13	55.50	<u>35.26</u>	<u>33.72</u>	64.42	71.55	**2.54**	**1.78**
	Ans + Sent	55.19	60.21	30.53	28.50	63.96	70.83	<u>4.76</u>	<u>4.15</u>
	Ans + Ent	60.42	64.80	**23.38**	**22.49**	65.04	71.99	2.56	2.20
	Ans + Sent + Ent	60.25	64.37	25.02	24.34	65.51	72.23	3.43	3.03
HotpotQA-small									
Dev	Ans	56.44	61.86			51.89	67.72		
	Ans + Sent	57.92	63.44			53.43	70.61		
	Ans + Ent	57.92	63.14			53.85	70.75		
	Ans + Sent + Ent	57.43	64.44			53.99	70.86		
ADDUNRELATED	Ans	50.00	56.24	11.41	9.09	50.77	66.85	**2.16**	**1.28**
	Ans + Sent	52.97	60.64	**8.55**	**4.41**	52.03	68.17	2.62	3.46
	Ans + Ent	51.49	57.43	11.10	9.04	51.33	67.97	<u>4.68</u>	<u>3.93</u>
	Ans + Sent + Ent	47.03	55.59	<u>18.11</u>	<u>13.73</u>	52.17	68.16	3.37	3.81
ADDRELATED	Ans	53.96	60.48	4.39	2.23	50.07	67.14	<u>3.51</u>	<u>0.86</u>
	Ans + Sent	57.43	63.37	0.85	0.11	54.13	70.54	**-1.31**	0.10
	Ans + Ent	58.91	64.11	**-1.71**	**-1.54**	54.13	70.27	-0.52	0.68
	Ans + Sent + Ent	53.96	61.23	<u>6.04</u>	<u>4.98</u>	54.69	71.24	-1.30	**-0.54**
ADD2	Ans	54.46	59.52	<u>3.51</u>	<u>3.78</u>	51.75	67.53	0.27	0.28
	Ans + Sent	58.91	64.31	**-1.71**	**-1.37**	52.45	69.03	1.83	2.24
	Ans + Ent	56.93	62.33	1.71	1.28	50.91	67.81	<u>5.46</u>	<u>4.16</u>
	Ans + Sent + Ent	55.94	62.58	2.59	2.89	54.41	70.82	**-0.78**	**0.06**
ADD2SWAP	Ans	48.51	53.94	<u>14.05</u>	<u>12.80</u>	50.63	66.94	2.43	**1.15**
	Ans + Sent	53.47	60.30	7.68	4.95	52.03	68.16	2.62	3.47
	Ans + Ent	53.96	60.51	**6.84**	**4.17**	51.05	67.78	<u>5.20</u>	<u>4.20</u>
	Ans + Sent + Ent	50.99	58.64	11.21	9.00	53.29	69.33	**1.30**	2.16

Table 9: Performance drop (smaller is better) for two types of questions (comparison and bridge questions) of the four settings of the model on the four debiased sets of 2Wiki and HotpotQA-small. The best and worst scores are boldfaced and underlined, respectively. For both 2Wiki and HotpotQA-small, we choose the results from the first time running to perform the analysis.

Type	Example
Bridge - Prune	**Paragraph A:** Polish-Russian War (Wojna polsko-ruska) is a 2009 Polish film directed by Xawery Żuławski based on . . . **Paragraph B:** Xawery Żuławski (born 22 December 1971 in Warsaw) is a Polish film director. . . . He is the son of actress Małgorzata Braunek and director Andrzej Żuławski. . . . **Q:** Who is the director of Polish-Russian War? **Predicted answer:** Andrzej Żuławski ✗ **Predicted entity-level:** ("Polish-Russian War", "director", "Xawery Żuławski") ✓
Bridge - Prune	**Paragraph A:** Francesca von Habsburg (born 7 June 1958) is an art collector and the estranged wife of Karl von Habsburg, current head of the House of Habsburg- Lorraine. **Paragraph B:** Michaela von Habsburg was born . . . She is the twin sister of Monika von Habsburg, and daughter of Otto von Habsburg and Princess Regina of Saxe - Meiningen. **Q:** Who is the spouse of Francesca von Habsburg? **Predicted answer:** Princess Regina of Saxe - Meiningen ✗ **Predicted entity-level:** ("Francesca von Habsburg", "spouse", "Karl von Habsburg") ✓
Comparison - Inverted	**Paragraph A:** Montréal/Les Cèdres Airport is a general aviation aerodrome located approximately west of Montreal, Quebec, Canada near Autoroute 20 west of . . . **Paragraph B:** Flying J Ranch Airport is a privately owned, public use . . . The airport is located southwest of the central business district of Pima, a city in Graham County, Arizona, United States and northeast of Tucson International Airport. . . . **Q:** Are Montréal/Les Cèdres Airport and Flying J Ranch Airport located in different countries? **Predicted answer:** no ✗ **Predicted entity-level:** ("Flying J Ranch Airport", "country", "United States") & ("Montréal/Les Cèdres Airport", "country", "Canada") ✓
Comparison - Inverted	**Paragraph A:** A Romance of the Air is a 1918 American silent drama film based . . . Directed by Harry Revier, the film was . . . **Paragraph B:** Harry Revier (16 March 1890 – 13 August 1957) was ... American director . . . **Paragraph C:** How Moscha Came Back is a 1914 silent film comedy short directed by Phillips Smalley. . . . **Paragraph D:** Phillips Smalley (August 7, 1865 – May 2, 1939) was an American silent film director and actor. **Q:** Which film has the director who was born later, A Romance of the Air or How Moscha Came Back? **Predicted answer:** How Moscha Came Back ✗ **Predicted entity-level:** ("A Romance of the Air", "director", "Harry Revier"), ("How Moscha Came Back", "director", "Phillips Smalley"), ("Harry Revier", "date of birth", "16 March 1890"), & ("Phillips Smalley", "date of birth", "August 7, 1865") ✓

Table 10: Examples of the outputs predicted by our model, which is trained on three tasks simultaneously.