Correct after Answer: Enhancing Multi-Span Question Answering with Post-Processing Method

Jiayi Lin [1,2], Chenyang Zhang [1,2], Haibo Tong [1,2], Dongyu Zhang [1,2]
Qingqing Hong [1,2], Bingxuan Hou [1,2], Junli Wang [1,2]*
[1] Key Laboratory of Embedded System and Service Computing (Tongji University),
Ministry of Education, Shanghai 201804, China.
[2] National (Province-Ministry Joint) Collaborative Innovation Center
for Financial Network Security, Tongji University, Shanghai 201804, China.
{2331908, inkzhangcy, 2151130, yidu}@tongji.edu.cn
{2332012, 2052643, junliwang}@tongji.edu.cn

Abstract

Multi-Span Question Answering (MSQA) requires models to extract one or multiple answer spans from a given context to answer a question. Prior work mainly focuses on designing specific methods or applying heuristic strategies to encourage models to predict more correct predictions. However, these models are trained on gold answers and fail to consider the incorrect predictions. Through a statistical analysis, we observe that models with stronger abilities do not predict less incorrect predictions compared with other models. In this work, we propose **A**nswering-**C**lassifying-**C**orrecting (ACC) framework, which employs a post-processing strategy to handle incorrect predictions. Specifically, the ACC framework first introduces a **classifier** to classify the predictions into three types and exclude "wrong predictions", then introduces a **corrector** to modify "partially correct predictions". Experiments on several MSQA datasets show that ACC framework significantly improves the Exact Match (EM) scores, and further analysis demostrates that ACC framework efficiently reduces the number of incorrect predictions, improving the quality of predictions.[1]

Context:
Don't Hug Me I'm Scared (often abbreviated to DHMIS) is a live - action / animated surreal horror comedy web series created by British filmmakers Becky Sloan and Joseph Pelling ...

Question:
Who made Don't Hug Me I'm Scared?

Gold Answers:
Becky Sloan, *Joseph Pelling*

Predictions:
Joseph Pelling (correct)
Sloan (partially correct)
DHMIS (wrong)

Figure 1: An example of MSQA. This question has two gold answers: "Becky Sloan" and "Joseph Pelling". "Joseph Pelling" is a correct prediction, "Sloan" is a partially correct prediction and "DHMIS" is a wrong prediction. Best read in colors.

1 Introduction

Machine Reading Comprehension (MRC) requires models to answer a question based on a given context (Rajpurkar et al., 2018; Kwiatkowski et al., 2019; Lai et al., 2017). In a real-world scenario, a single question typically corresponds to multiple answers. To this end, Multi-Span Question Answering (MSQA) has been proposed (Ju et al., 2022; Li et al., 2022; Yue et al., 2023). Different from the traditional Single-Span Question Answering (SSQA) task, the goal of MSQA is to extract one or multiple non-overlapped spans from the given context. Taking Figure 1 as an instance, the question "Who made Don't Hug Me I'm Scared?" has two answers: "Becky Sloan" and "Joseph Pelling".

Recent MSQA work integrates various approaches. Yang et al. (2021); Hu et al. (2019) incorporate heuristic strategies based on traditional pointer models (Vinyals et al., 2015) to extract multiple answers; Segal et al. (2020); Li et al. (2022) convert MSQA task into a sequence-tagging task to mark answers; Huang et al. (2023a); Zhang et al. (2024) enumerate all candidate answers and select the final answers with a learnable threshold, and Huang et al. (2023b); Zhang et al. (2023) utilize Large Language Models (LLMs) to handle MSQA tasks with few shot prompts.

Prior work mainly focus on specific methods or heuristic strategies for more correct predictions. However, these models are trained on gold answers, and fail to consider the incorrect predictions. To further investigate the incorrect predictions, we

*Corresponding author. This work was supported by the National Key Research and Development Program of China under Grant 2023YFB3002201.

[1] Our code and data are available at https://github.com/TongjiNLP/ACC.

Findings of the Association for Computational Linguistics: EACL 2024, pages 2701–2717
November 12-16, 2024 ©2024 Association for Computational Linguistics

categorize predictions into "correct predictions", "partially correct predictions" and "wrong predictions" based on whether they should be modified or excluded. Then we conduct a statistical analysis on several MSQA models (details in Section 2.3), and observe that stronger MSQA models [1] do not predict less incorrect predictions compared with other models. Consequently, performance of the MSQA models can be further improved on the basis of reducing incorrect predictions.

In this work, we propose **A**nswer-**C**lassify-**C**orrect (ACC) framework, which employs a post-processing strategy to handle with incorrect predictions. The ACC framework simulates humans strategy in realword examinations: listing candidate answers, reviewing and modifying. Specifically, we design the **classifier** to categorize candidate answers into "correct predictions", "partially correct predictions" or "wrong predictions", then we design the **corrector** to modify "partially correct predictions", finally we exclude "wrong predictions" and obtain final predictions. To train the classifier and the corrector, we also apply an automatic annotation approach which samples incorrect predictions from the training datasets and constructs the silver-labeled datasets.

We conduct experiments on four MSQA datasets. Experiment results show that the ACC framework significantly improves the performance. After applying the ACC framework, the EM F1 score increases from 69.05% to 72.26% for Tagger-RoBERTa (Li et al., 2022) and from 65.57% to 76.31% for BART-base (Lewis et al., 2020) on the MultiSpanQA dataset (Li et al., 2022). Further analysis on the predictions indicates that the ACC framework effectively reduces the number of incorrect predictions and obtains more correct predictions, enhancing the qualities of predictions. In addition, a pilot study with GPT-3.5 [2] is conducted, exhibiting extensive application of ACC framework for LLMs in a Chain-of-Thought (CoT) manner (Wei et al., 2022; Kojima et al., 2022).

Our contributions are summarized as follows:

- We develop a three-fold taxonomy for the MSQA predictions based on whether a prediction should be modified or excluded. Then, we conduct a statistical analysis, revealing distributions over the three categories.
- Inspired by humans' strategies, we propose the ACC framework, which includes a classifier to exclude incorrect predictions and includes a corrector to modify imperfect predictions. We also design an automatical annotation approach to sample incorrect predictions and construct silver-labeled datasets.
- We conduct experiments and analysis on several MSQA datasets. Results show that the ACC framework significantly enhances the quality of the MSQA predictions.

2 Taxonomy of MSQA Predictions

2.1 Formalization

Given a question Q and its corresponding context C, the goal of MSQA is to train a model M to extract a set of m answer spans $P = \{p_1, p_2, ..., p_m\}$ from the given context, shown as Eq. 1.

$$P = M(C, Q) \tag{1}$$

2.2 Taxonomy

Intuitively, the predictions can be categorized as correct or incorrect predictions. However, some of incorrect predictions should be modified while others should be excluded. For instances in Figure 1, "Sloan" and "DHMIS" are both incorrect predictions. However, "Sloan" is similar to the gold answer "Bercy Sloan" but "DHMIS" is totally wrong. Therefore, we further categorize incorrect predictions into "partially correct predictions" and "wrong predictions".

Based on above analysis, we category the prediction $p_i \in P$ into one of the following three types: "correct prediction", "partially correct prediction" and "wrong prediction".

Correct prediction. If the prediction p_i is one of the gold answers, which means $p_i \in A$, p_i is regarded as a correct prediction.

Partially correct prediction. We utilize Word Overlap (WO) and BERTScore (BS) (Zhang et al., 2020) to define partially correct predictions. Word Overlap considers the overlap between two spans in word level, while BERTScore computes semantic similarity of two spans in the manner of cosine similarity. Details of Word Overlap and BERTScore are shown in Appendix A.

For a prediction p_i, if there exists $a_j \in A$ which satisfies $WO(p_i, a_j) \geq \alpha$ and $BS(p_i, a_j) \geq \beta$,

[1] "stronger" indicates higher MSQA performance, which can be demonstrated using EM scores.

[2] https://platform.openai.com/.

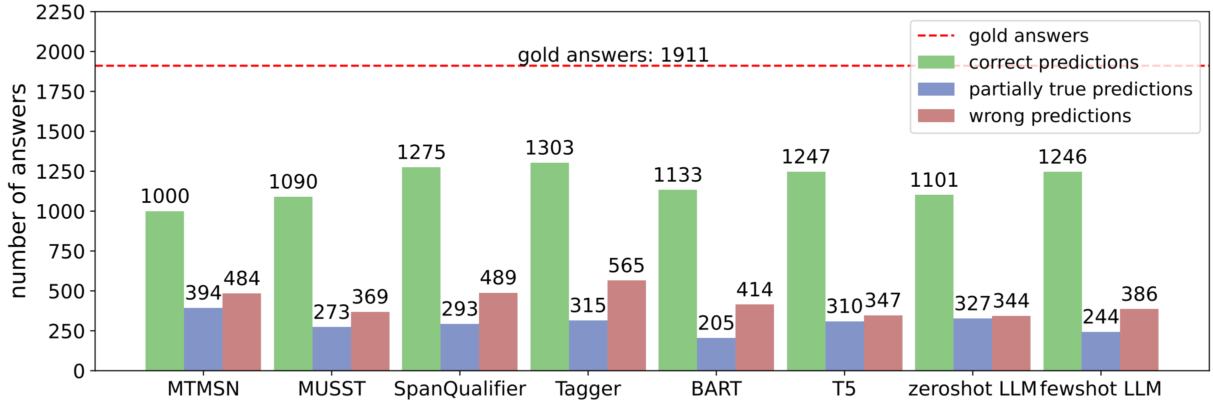

Figure 2: The distribution of correct predictions, partially correct predictions and wrong predictions on the validation set of MultiSpanQA. The validation set of MultiSpanQA contains 653 questions with 1,911 gold answers.

where α and β are hyper-parameters, the p_i is regarded as the partially correct prediction.

Wrong prediction. If p_i is not a correct prediction or a partially correct prediction, the p_i is regarded as wrong prediction.

Figure 1 shows an example containing these three types of predictions. The gold answers are "Becky Sloan" and "Joseph Pelling". Among the three predictions, "Joseph Pelling" is a correct prediction; "Sloan" is a partially correct prediction and "DHMIS" is a wrong prediction.

2.3 Distribution of MSQA Predictions

Based on the taxonomy, we conduct a statistical analysis on the validation set of MultiSpanQA (Li et al., 2022). We select four discriminative models: MTMSN (Hu et al., 2019), MUSST (Yang et al., 2021), Tagger (Li et al., 2022) and SpanQualifier (Huang et al., 2023a), and three generative models: BART (Lewis et al., 2020), T5 (Raffel et al., 2020) and GPT-3.5 (zero-shot and few-shot). Details of these models are shown in Appendix B.2.

The statistical results are shown in Figure 2. We observed that models with better performance (shown in Table 1) on the validation set predict more correct predictions as well as more wrong predictions. For example, for discriminative models, Tagger predicts 1,303 correct predictions but also predict 565 wrong predictions, while MTMSN predicts 1,000 correct predictions and 484 wrong predictions; Similarly, after adding few-shot demonstrations, the LLM generates more correct as well as wrong predictions compared with zero-shot setting. Therefore, we believe that the post-processing method can effectively enhance the quality of predictions by reducing the number of incorrect predictions, resulting in better performance.

3 Method

In this section, we describe the ACC framework, which is designed to handle with partially correct predictions and wrong predictions. The architecture of the ACC framework is shown in Figure 3.

Similar to the humans' strategies, the post-processing procedure of the ACC framework consists of three steps: The first step is **answering**, where we employ a **reader** to obtain initial predictions P; The second step is **classifying**, where we employ a **classifier** to categorize each prediction p_i into one of the three classes: correct prediction, partially correct prediction and wrong prediction; The last step is **correcting**, where we employ a **corrector** to modify the partially correct predictions. We reserve correct predictions predicted by the classifier and the modified predictions from the corrector as the final predictions.

Next, we will provide more details of the reader, the classifier and the corrector. We will also introduce an automatic annotation approach which samples incorrect predictions and constructs training data for the classifier and the corrector.

3.1 Reader

The main function of the reader is to extract several text spans from context based on a given question. This process can be described as:

$$P = Reader(Q, C) \tag{2}$$

3.2 Classifier

The predictions of the reader may include partially correct predictions or wrong predictions (men-

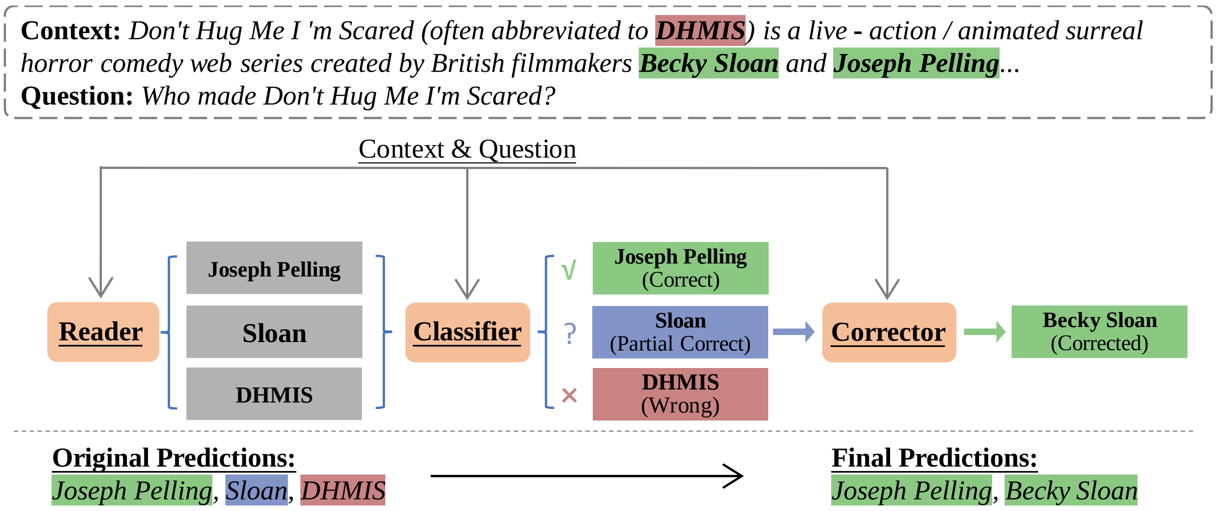

Figure 3: The overall architecture of our proposed ACC framework.

tioned in Section 2.2). To this end, we design the classifier to classify them and exclude wrong predictions. Given the candidate predictions P, the classifier splits them into correct predictions P_c, partially correct predictions P_p and wrong predictions P_w. This process can be described as:

$$P_c, P_p, P_w = Classifier(P, Q, C) \tag{3}$$

Specifically, the classifier consists of a transformer (Vaswani et al., 2017) encoder and a classification head. The classification head includes an MLP layer to obtain probability of each class. Inspired by Zhu et al. (2022), we also add a cross-attention layer in the classification head. The cross-attention layer calculates the attention scores between the question and the context to enhance the representations of them.

3.3 Corrector

The classifier is able to exclude wrong predictions, however, there may still contain partially correct predictions which are imperfect and should be modified. Hence, we design the corrector to modify those partially correct predictions. This process can be described as:

$$\hat{P}_p = Corrector(P_p, Q, C) \tag{4}$$

where $\hat{P}_p$ are the predictions modified by corrector.

We adopt traditional pointer model (Vinyals et al., 2015) to predict the start and end probabilities st and ed. During the inference stage, for the text span starting at i-th token and ending at j-th token, we calculate its confidence score $score_{ij} = st_i + ed_j$ and obtain the best index pair (i, j) which maximizes $score_{ij}$, then extract its corresponding span as the modified prediction.

The final outputs of the ACC framework $\tilde{P}$ consist of the correct predictions P_c predicted by the classifier and the modified predictions $\hat{P}_p$ from the corrector, described as:

$$\tilde{P} = P_c \cup \hat{P}_p \tag{5}$$

3.4 Data Annotations

To train the classifier and the corrector, we need both correct predictions and incorrect predictions. However, most MSQA datasets do not contain incorrect predictions. Inspired by Gangi Reddy et al. (2020), we adopt an automatical sampling method similar to K-fold cross-validation, to collect incorrect predictions from the MSQA datasets and construct our silver-labeled datasets.

First, we randomly divide the training data D into K equal subsets: $D_1, D_2, ..., D_K$. We perform K iterations, in the i-th iteration we initialize a MSQA model M (i.e. reader mentioned in Section 3.1) and train it with all training data except D_i, then sampling the predictions of D_i with M. After K iterations, we utilize the gold answers from training dataset D to annotate all predictions, and construct the silver-labeled dataset [3].

[3] More details are shown in Appendix B.3.

	MultiSpanQA			MultiSpanQA-Expand			MAMRC			MAMRC-Multi		
	EM P	EM R	EM F1	EM P	EM R	EM F1	EM P	EM R	EM F1	EM P	EM R	EM F1
Discriminative Models (BERT-base)												
MTMSN	59.86	49.97	54.47	63.39	56.00	59.47	73.94	78.36	76.08	71.69	77.47	74.46
+ACC	**71.75**	**55.87**	**62.82**	**68.95**	**58.81**	**63.48**	**81.84**	77.70	**79.72**	**85.13**	**79.82**	**82.39**
MUSST	69.82	61.94	65.64	69.29	63.16	66.08	78.01	79.71	78.85	76.69	77.16	76.92
+ACC	**73.07**	61.78	**66.96**	**70.54**	**62.60**	**66.33**	**82.75**	77.57	**80.08**	**86.10**	**77.48**	**81.56**
Tagger	66.22	72.14	69.05	64.35	65.66	64.99	79.47	83.59	81.48	75.85	78.19	77.00
+ACC	**72.39**	72.12	**72.26**	**68.70**	**66.21**	**67.43**	**83.62**	81.80	**82.70**	**85.77**	**78.36**	**81.90**
SpanQualifier	70.40	72.82	71.58	64.65	69.65	66.99	83.40	80.83	82.10	75.63	85.77	80.37
+ACC	**73.69**	71.32	**72.47**	**67.68**	68.53	**68.09**	**82.83**	**81.88**	**82.35**	**85.14**	83.77	**84.45**
Discriminative Models (RoBERTa-base)												
MTMSN	59.86	49.97	54.47	63.39	56.00	59.47	73.94	78.36	76.08	71.69	77.47	74.46
+ACC	**71.75**	**55.87**	**62.82**	**68.95**	**58.81**	**63.48**	**81.84**	77.70	**79.72**	**85.13**	**79.82**	**82.39**
MUSST	69.82	61.94	65.64	69.29	63.16	66.08	78.01	79.71	78.85	76.69	77.16	76.92
+ACC	**73.07**	61.78	**66.96**	**70.54**	62.60	**66.33**	**82.75**	77.57	**80.08**	**86.10**	**77.48**	**81.56**
Tagger	66.22	72.14	69.05	64.35	65.66	64.99	79.47	83.59	81.48	75.85	78.19	77.00
+ACC	**72.39**	72.12	**72.26**	**68.70**	**66.21**	**67.43**	**83.62**	81.80	**82.70**	**85.77**	**78.36**	**81.90**
SpanQualifier	70.40	72.82	71.58	64.65	69.65	66.99	83.40	80.83	82.10	75.63	85.77	80.37
+ACC	**73.69**	71.32	**72.47**	**67.68**	68.53	**68.09**	**82.83**	**81.88**	**82.35**	**85.14**	83.77	**84.45**
Generative Models												
BART-base	69.10	62.38	65.57	60.42	55.95	58.10	77.53	74.33	75.89	75.96	73.21	74.56
+ACC	**73.90**	61.80	**67.31**	**63.68**	55.70	**59.43**	**80.47**	72.47	**76.26**	**81.26**	71.22	**75.91**
T5-base	70.56	67.97	69.24	64.63	64.59	64.61	77.01	79.88	78.41	75.27	77.14	76.19
+ACC	**73.93**	66.20	**69.85**	**67.43**	63.32	**65.31**	**80.79**	77.43	**79.07**	**80.65**	74.73	**77.58**
GPT3.5 (Zeroshot)	64.83	60.86	62.78	39.60	53.68	45.58	45.45	57.34	50.71	57.00	63.27	59.97
+ACC	**73.04**	**61.96**	**67.04**	**48.64**	**53.96**	**51.16**	**57.10**	**57.71**	**57.40**	**69.54**	**64.06**	**66.69**
GPT3.5 (Fewshot)	68.94	68.18	68.56	42.44	58.13	49.06	58.42	73.79	65.21	65.38	76.68	70.58
+ACC	**74.88**	66.61	**70.51**	**51.65**	57.91	**54.60**	**68.02**	70.94	**69.45**	**75.39**	74.97	**75.18**

Table 1: EM Scores on four MSQA datasets. "EM P" "EM R" "EM F1" refer to the precision, recall and F1 score under the EM metric, respectively. "Discriminative Models (BERT-base/RoBERTa-base)" refer to models that utilize BERT-base or RoBERTa-base as encoders. The results marked in **bold** means improvements after applying the ACC framework.

4 Experiments

4.1 Experimental Setup

Datasets Four MSQA datasets are integrated in experiments: MultiSpanQA (Li et al., 2022), MultiSpanQA-Expand (Li et al., 2022), MAMRC (Yue et al., 2023) and an additional synthetic dataset MAMRC-Multi. Details of these datasets are shown in Appendix B.1.

MSQA models We set both discriminative models and generative models as readers. For discriminative models, we set MTMSN (Hu et al., 2019), MUSST (Yang et al., 2021), Tagger (Li et al., 2022) and SpanQualifier (Huang et al., 2023a); For generative models, we set BART (Lewis et al., 2020), T5 (Raffel et al., 2020) and GPT-3.5. Details of these models are shown in Appendix B.2.

Evaluation Metrics We use **Exact Match Precision/Recall/F1 (EM P/R/F1)** (Li et al., 2022) as the main metrics in our experiments. EM assign a score of 1 when a prediction fully matches one of the gold answers and 0 otherwise.

Implementation Details For the classifier and corrector in the ACC framework, we use RoBERTa-base (Zhuang et al., 2021) as encoder. For discriminative MSQA models, we use both BERT-base (Devlin et al., 2019) and RoBERTa-base as encoder. For the hyper parameters mentioned in Section 2.2, based on the average Word Overlap and BERTScore of the sampled data, we set $\alpha = 0.25$ and $\beta = 0.6$ to obtain balanced training data. See more training and inference details in Appendix B.3.

4.2 Main Results

Table 1 shows the main results on four MSQA datasets. Discriminative models perform better than generative models on the MSQA task, especially on MultiSpanQA-Expand and MAMRC where questions may contain only one answer or no answer. The reason may be that discriminative models are suited for extracting text spans from a given context, whereas generative models are suited for text generation.

After applying the ACC framework, both discriminative models and generative models gain improvements. For instances, the EM F1 score of Tagger (RoBERTa-base) increases from 69.05% to 72.26% and the EM F1 score of BART increases from 65.57% to 67.31% on MultiSpanQA. For most settings, presicion scores show significant

	MultiSpanQA		
	EM P	EM R	EM F1
Tagger BERT	56.66	65.46	60.74
+ cls only	64.90	63.98	64.44
+ cor only	62.49	**69.11**	65.63
+ cor & cls	67.14	67.44	67.29
+ binary cls & cor	68.58	66.56	67.56
+ cls & cor	**68.52**	67.05	**67.78**
Tagger RoBERTa	66.22	72.14	69.05
+ cls only	70.54	70.58	70.56
+ cor only	68.50	**73.09**	70.72
+ cor & cls	71.21	71.43	71.32
+ binary cls & cor	72.45	70.94	71.68
+ cls & cor	**72.39**	72.12	**72.26**

Table 2: Ablation study of ACC framework on the dev set of MultiSpanQA. The best performance is in **bold**.

	MultiSpanQA		
	EM P	EM R	EM F1
Tagger BERT	56.66	65.46	60.74
+ att cls & T5 cor	64.90	63.98	64.44
+ vanilla cls & ext cor	**68.54**	66.10	67.29
+ att cls & ext cor	68.52	**67.05**	**67.78**
Tagger RoBERTa	66.22	72.14	69.05
+ att cls & T5 cor	70.54	70.58	70.56
+ vanilla cls & ext cor	72.23	71.56	71.89
+ att cls & ext cor	**72.39**	**72.12**	**72.26**

Table 3: Comparison between diffent combinations of the classifier and the corrector on the validation set of MultiSpanQA. "Att cls" refers to the classifier mentioned in Section 3.2; "vanilla cls" refers to the classifier without cross-attention layer; "Ext cor" refers to the corrector mentioned in Section 3.3 and "T5 cor" refers to the T5 corrector. The best performance is in **bold**.

improvements while some recall scores show slight declines, the reason may be that while the classifier successfully identifies some wrong predictions, it also mistakenly classifies some correct predictions as wrong, leading to the exclusion of some correct predictions and thereby lowering the recall scores. In Section 5.2, we will analyze the classification results of the classifier to verify this point.

We also evaluate the ACC framework with Partial Match P/R/F1 (PM P/R/F1), which considers the overlap between the predictions and gold answers. Results are shown in Appendix C.1.

5 Discussions

5.1 Ablation Study

Roles of classifier and corrector. ACC framework uses the "answer-classify-correct" procedure with the classifier and the corrector. To investigate whether there exists better post-processing procedure, we conduct an ablation study by: 1. only employing the classifier or corrector (cls \ cor only); 2. changing the order of classifier and corrector (cor & cls); 3. modifying both correct predictions and partially correct predictions (binary cls & cor).[4]

Table 2 shows the results of the ablation study on the dev set of MultiSpanQA. The performance of "cls only" and "cor only" lags behind ACC framework, demonstrating the significance of the classifier and corrector. Changing the order between classifier and corrector also shows decline, the reason may be that using corrector first may lead to conceal wrong predictions, thereby the classifier may fail to categorize them as wrong predictions. We also observe that modifying both correct predictions and partially correct predictions does not achieve improvements, demostrating the necessity of distinguishing correct predictions and partially correct predictions and modifying partially correct predictions solely.

Comparison with different models. ACC framework uses a classifier with a cross-attention layer and a corrector based on the pointer model. However, ACC framework can also opt for alternative type of classifiers or correctors. To this end, we replace the classifier and the corrector with other models and compare their performance.[5]

Table 3 shows the results of the comparison between different model combinations on the dev set of MultiSpanQA. After replacing the classifier or the corrector, ACC framework shows declines, especially when applying a generative model, ACC framework lag behind other settings. This indicates that the generative models are less capable than traditional pointer models in correcting predictions.

5.2 Analysis on the Predictions

Accuracy of the classifier. To analyze the capability of the classifier, we conduct a statistical analysis on its classification results. Table 4 shows the accuracy of the classifier on the dev set of MultiSpanQA. The classifier achieves an high accuracy on the correct predictions (95.82% for Tagger-BERT and 95.45% for Tagger-RoBERTa), demonstrating that the ACC framework reserves most correct predictions. On the other hand, the classifier exclude about 1/3 wrong predictions, contributing

[4] For "cls only", we only exclude wrong predictions; for "cor only", we correct all predictions; for "cor & cls", we first correct all predictions, then classify them and only exclude wrong predictions.

[5] For the classifier, we replace it with a vanilla classifier where we remove the cross-attention layer; for the corrector, we replace it with T5 (Raffel et al., 2020) which outputs texts as the corrected answers.

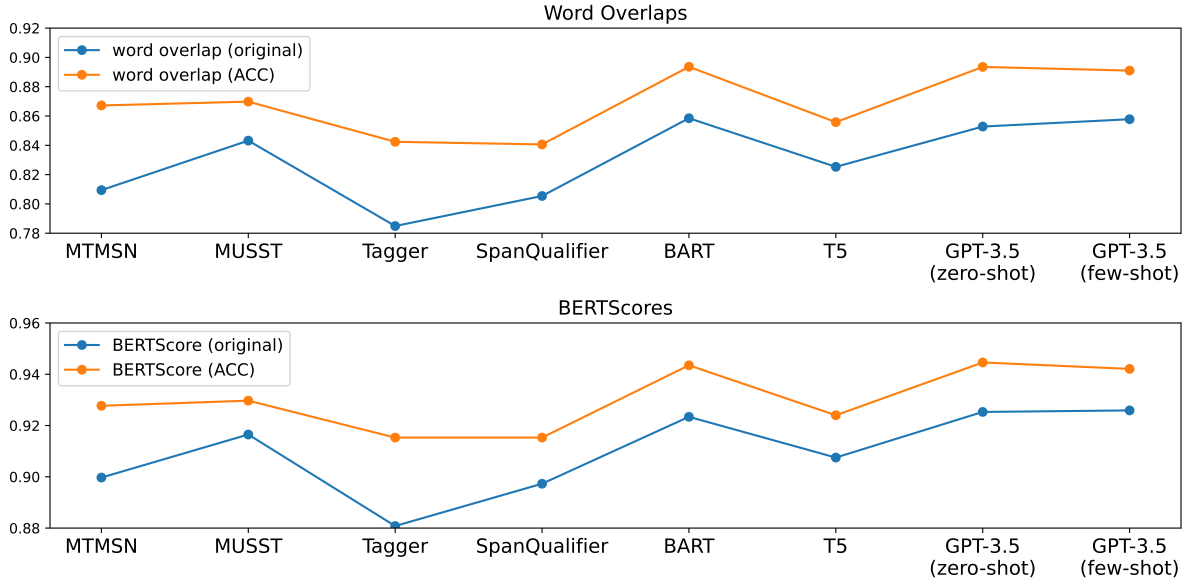

Figure 4: **Top:** Average Word Overlap of the predictions. **Button:** Average BERTScore of the predictions. After applying the ACC framework, both Word Overlap and BERTScore raise, indicating that the ACC framework effectively enhances the quality of the predictions.

Tagger BERT			
label \ pred	wrong	partially	correct
wrong	**268 (37.85%)**	148 (20.9%)	292 (41.24%)
partially	16 (6.13%)	**98 (37.55%)**	147 (56.32%)
correct	26 (2.18%)	24 (2.01%)	**1145 (95.82%)**
Tagger RoBERTa			
label \ pred	wrong	partially	correct
wrong	**135 (27.44%)**	105 (21.34%)	252 (51.22%)
partially	22 (8.63%)	**83 (32.55%)**	150 (58.82%)
correct	27 (2.01%)	34 (2.54%)	**1280 (95.45%)**

Table 4: Accuracy of the classifier on the dev set of MultiSpanQA. The correct classifications of each types are in **bold**.

to the imporvements on EM F1 scores, while the accuracies on the partially true predictions and the wrong predictions can be further improved.

Changes in answers by the corrector. To analyze the capability of the corrector, we also conduct a statistical analysis on how many prediction has been changed. Table 5 shows the changes of the partially correct predictions on the dev set of MultiSpanQA. The corrector changes 30.77% of the answers for Tagger-BERT and 27% for Tagger-RoBERTa, respectively. For Tagger-BERT, 27.47% of the not-correct predictions are modified to the correct predictions, while 3.3% of the correct predictions are modified to the not-correct predictions. Furthermore, among all the partially correct predictions derived from the classifier, over 60% of the incorrect predictions remain incorrect, indicatisng

Tagger BERT		
cls \ cls & cor	incorrect	correct
incorrect	172 (63.00%)	**75 (27.47%)**
correct	**9 (3.3%)**	17 (6.23%)
Tagger RoBERTa		
cls \ cls & cor	incorrect	correct
incorrect	137 (61.43%)	**52 (23.32%)**
correct	**11 (4.93%)**	23 (10.31%)

Table 5: Changes in answers by the corrector on the dev set of MultiSpanQA.

a significant room for improvements.

5.3 Analysis on the Quality of the Predictions

Previous experiment indicates that the ACC framework imporves EM scores. However, the ACC framework may overfit to the annotation boundaries rather than enhancing the quality of the predictions. To this end, we utilize other metrics such as the Word Overlap and BERTScore (mentioned in Section 2.2) and compare the changes of these metrics after applying the ACC framework.

Figure 4 shows the comparison results. Both Word Overlaps and BERTScores raise after applying the ACC framework, with the most significant enhancement in the Tagger where Word Overlap increases by 7% and BERTScore increases by 4%. This indicates that the ACC framework enhances the quality of the predictions, rather than overfit to the annotation boundaries.

We also utilize LLM to evaluate the modified

Question: Where was the pride of Britain awards held ?
Context: ...The first Pride of Britain Awards were held at the Dorchester Hotel in London in May 1999 , then relocated to The London Studios in 2000 , and then later relocated to Grosvenor House from the 2011 award ceremony . The awards are organized in association with the Daily Mirror , Lidl , ITV , Good Morning Britain and The Prince 's Trust .
Predictions (MTMSN RoBERTa):
{ the London Studios in 2000 ; and then later relocated to Grosvenor House ; London ; Dorchester Hotel ; Daily Mirror }

Classifier Output :
the London Studios in 2000 : Partially Correct
and then later relocated to Grosvenor House : Partially Correct
London : Partially Correct
Dorchester Hotel : Correct
Daily Mirror : Wrong

Corrector Output:
the London Studios in 2000
➡ London studios
and then later relocated to Grosvenor House
➡ Grosvenor House
London
➡ London Studios

Final Predictions :
{ Dorchester Hotel ; Grosvenor House ; London Studios }
Gold Answers :
{ Dorchester Hotel ; Grosvenor House ; The London Studios }

Figure 5: Case study. The example are selected from the validation set of MultiSpanQA. The correct predictions and gold answers are in green, the partially correct predictions are in blue and the wrong predictions are in red. Best read in colors.

answers, results are shown in Appendix C.2.

5.4 Case Study

We conduct a case study to demostrate that the ACC framework effectively excludes incorrect predictions and corrects some partially correct predictions. We select a real example where the predictions exactly match the gold answers (i.e. EM F1 = 100%), shown in Figure 5. In this example, the MTMSN presents five predictions: "the London Studios in 2000", "and then relocated to Grosvenor House", "London", "Dorchester Hotel" and "Daily Mirror". The classifier identifies "Dorchester Hotel" as a correct prediction and "Daily Mirror" as a wrong prediction. The others three predictions contain irrelevant information or lack specific details, so they are identified as partially correct predictions and modifed by the corrector.[6] This example demostrates that our ACC framework effectively enhance the quality of the predictions.

5.5 Pilot Study with LLM

ACC framework utilizes a fine-tuned RoBERTa encoder as the backbone. To investigate whether our proposed method works on larger models, we conduct a pilot study by replacing the classifier or corrector with a prompted LLM. The implementation details and prompts are shown in Appendix C.4.

[6]When calculating EM scores, the article "the" is ignored, so "London Studios" and "The London Studios" are considered as the same prediction.

	MultiSpanQA		
	EM P	EM R	EM F1
Tagger BERT	56.66	65.46	60.74
+LLM cls & LLM cor	68.60	63.35	65.87
+LLM cls & FT cor	**70.04**	64.47	67.14
+FT cls & LLM cor	67.93	66.51	67.21
+FT cls & FT cor	68.52	**67.05**	**67.78**
Tagger RoBERTa	66.22	72.14	69.05
+LLM cls & LLM cor	72.71	68.10	70.33
+LLM cls & FT cor	**73.69**	68.97	71.25
+FT cls & LLM cor	71.71	71.48	71.59
+FT cls & FT cor	72.39	**72.12**	**72.26**

Table 6: Performance of ACC framework with LLM on the dev set of MultiSpanQA. "LLM cls/cor" refers to classifier/corrector replaced by LLM, and "FT cls/cor" refers to a fine-tuned model. The best performance is in **bold**.

Table 6 shows the experiment results. After replacing the classifier or the corrector with LLM, the ACC framework still achieves improvements on Tagger-BERT and Tagger-RoBERTa, which proves that our post-processing strategies can be effectively applied to LLM.

5.6 Model Size and Inference Time

We analyze the model size and the inference time of the ACC framework. Results and analysis are shown in Appendix C.3.

6 Related Work

6.1 Multi-Span Question Answering

Recently, a series of MSQA benchmarks (Ju et al., 2022; Li et al., 2022; Yue et al., 2023) have been proposed to faclitate research on QA tasks that are closer to real-world scenarios. MSQA tasks require models to extract one or multiple answer spans from a given context. Therefore, traditional SSQA models (Seo et al., 2017; Yu et al., 2018) are not sufficient to handle multi-span questions.

Existing MSQA methods can be categorized into four categories: (1) pointer-network-based methods. MTMSN (Hu et al., 2019) predicts the number of answers, then extracts non-overlapped answer spans; MUSST (Yang et al., 2021) uses an autogressive approach to iteratively extract multiple answers. (2) sequence-tagging-based methods. Segal et al. (2020) first convert MSQA task to a sequence-tagging task and utilize BIO tags to mark answer spans; Furthermore, Li et al. (2022) introduce multi-task learning and achieve better performance. (3) span-enumeration-based methods. SpanQualifier (Huang et al., 2023a) utilizes Multi-Layer Perceptron (MLP) to obtain confidence scores for each candidate span and applies a learnable threshold to select answer spans; Similarly, CSS (Zhang et al., 2024) compares each candidate span with its corresponding question after scoring to obtain answers more similar to the question. (4) LLM-based methods. With the emergence of LLMs like ChatGPT and GPT-4, generative pre-trained language models have been widely applied to various NLP tasks. Zhang et al. (2023) employ CoT strategies to prompt LLM, and Huang et al. (2023b) add negative examples in the few-shot demonstrations.

Existing methods mainly focus on predicting more correct predictions, while the ACC framework takes a post-processing strategy which aims to reduce the number of incorrect predictions. By excluding or modifying incorrect predictions, the ACC framework achieves better performance.

6.2 Post-Processing Methods

The post-processing method refers to modifying the original of the model to obtain better predictions. Existing post-processing methods can be categorized into two types: rule-based methods and model-based methods.

Ruled-based methods typically involve mannually designed rules such as voting to process the outputs from models (Campos and Couto, 2021; Wang et al., 2023). On the other hand, model-based methods utilize additional models to modify the hidden states or outputs of the original model, which have been widely applied in Controlled Text Generation (CTG) (Yang and Klein, 2021; Krause et al., 2021; Kim and Cho, 2023). In addition to CTG methods, GRACE (Khalifa et al., 2023) applies a fine-tuned discriminator to guide language model towards correct multi-step solutions; Ohashi and Higashinaka (2023) utilize a generative model to rewrite the output from a dialogue system and optimize it with Reinforcement Learning (RL) algorithms (Stiennon et al., 2020).

The work most similar to ours is (Gangi Reddy et al., 2020), which utilizes a corrector to modify the outputs of the SSQA model. However, they only focus on partial matches in single-span questions. In constrast, we consider the correctness of multiple predictions in MSQA and additionally employ a classifier to exclude incorrect predictions.

7 Conclusion

In this work, we primarily focus on incorrect predictions of the MSQA models. Through a statistical analysis, we observe that models with better performance do not predict less incorrect predictions compared with other models. To this end, we propose ACC framework, which employ a post-processing strategy to exclude wrong predictions and modify partially correct predictions. Experiments and analysis show that the ACC framework significantly improving the performance by reducing the number of incorrect predictions and obtaining more correct predictions, enhancing the quality of the MSQA predictions.

8 Limitations and Future Work

In this work, we categorize incorrect predictions into "partially correct predictions" and "wrong predictions", based on whether the answer should be modified or excluded. However, for "partially correct predictions", there exists more complicated conditions, for example, an incorrect prediction may responses to multiple gold answers. However, the ACC framework can only obtain one modified prediction. In addition, we do not consider the gold answers that MSQA models fail to predict (i.e., "missing predictions"), although the SOTA model still miss 1/3 gold answers. As for future work, we will design more effectively models to handle "partially correct predictions" and "wrong predictions". we will also explore strategies to handle "missing predictions".

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Margarida M. Campos and Francisco M. Couto. 2021. Post-processing biobert and using voting methods for biomedical question answering. In *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*, volume 2936 of *CEUR Workshop Proceedings*, pages 258–273. CEUR-WS.org.

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.

Revanth Gangi Reddy, Md Arafat Sultan, Efsun Sarioglu Kayi, Rong Zhang, Vittorio Castelli, and Avi Sil. 2020. Answer span correction in machine reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2496–2501, Online. Association for Computational Linguistics.

Minghao Hu, Yuxing Peng, Zhen Huang, and Dongsheng Li. 2019. A multi-type multi-span network for reading comprehension that requires discrete 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 1596–1606, Hong Kong, China. Association for Computational Linguistics.

Zixian Huang, Jiaying Zhou, Chenxu Niu, and Gong Cheng. 2023a. Spans, not tokens: A span-centric model for multi-span reading comprehension. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 874–884, New York, NY, USA. Association for Computing Machinery.

Zixian Huang, Jiaying Zhou, Gengyang Xiao, and Gong Cheng. 2023b. Enhancing in-context learning with answer feedback for multi-span question answering. In *Natural Language Processing and Chinese Computing*, pages 744–756, Cham. Springer Nature Switzerland.

Yiming Ju, Weikang Wang, Yuanzhe Zhang, Suncong Zheng, Kang Liu, and Jun Zhao. 2022. CMQA: A dataset of conditional question answering with multiple-span answers. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1697–1707, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. GRACE: Discriminator-guided chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15299–15328, Singapore. Association for Computational Linguistics.

Heegyu Kim and Hyunsouk Cho. 2023. GTA: Gated toxicity avoidance for LM performance preservation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14747–14763, Singapore. Association for Computational Linguistics.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Haonan Li, Martin Tomko, Maria Vasardani, and Timothy Baldwin. 2022. MultiSpanQA: A dataset for multi-span question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1250–1260, Seattle, United States. Association for Computational Linguistics.

Atsumoto Ohashi and Ryuichiro Higashinaka. 2023. Enhancing task-oriented dialogue systems with generative post-processing networks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3815–3828, Singapore. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

S. E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR '94*, pages 232–241, London. Springer London.

A. Rosenfeld and M. Thurston. 1971. Edge and curve detection for visual scene analysis. *IEEE Transactions on Computers*, C-20(5):562–569.

Elad Segal, Avia Efrat, Mor Shoham, Amir Globerson, and Jonathan Berant. 2020. A simple and effective model for answering multi-span questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3074–3080, Online. Association for Computational Linguistics.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Junjie Yang, Zhuosheng Zhang, and Hai Zhao. 2021. Multi-span style extraction for generative reading comprehension. In *Proceedings of the Workshop on Scientific Document Understanding co-located with 35th AAAI Conference on Artificial Inteligence, SDU@AAAI 2021, Virtual Event, February 9, 2021*, volume 2831 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Zhiang Yue, Jingping Liu, Cong Zhang, Chao Wang, Haiyun Jiang, Yue Zhang, Xianyang Tian, Zhedong Cen, Yanghua Xiao, and Tong Ruan. 2023. Mamrc: A multi-answer machine reading comprehension dataset. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2144–2148, New York, NY, USA. Association for Computing Machinery.

Chen Zhang, Jiuheng Lin, Xiao Liu, Yuxuan Lai, Yansong Feng, and Dongyan Zhao. 2023. How many answers should I give? an empirical study of multi-answer reading comprehension. In *Findings of the*

Association for Computational Linguistics: ACL 2023, pages 5811–5827, Toronto, Canada. Association for Computational Linguistics.

Penghui Zhang, Guanming Xiong, and Wen Zhao. 2024. Css: Contrastive span selector for multi-span question answering. In *PRICAI 2023: Trends in Artificial Intelligence*, pages 225–236, Singapore. Springer Nature Singapore.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Pengfei Zhu, Zhuosheng Zhang, Hai Zhao, and Xiaoguang Li. 2022. Duma: Reading comprehension with transposition thinking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:269–279.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Details of Word Overlap and BERTScore

Word Overlap. Assuming that a prediction p_i contains k words $\{p_{i1}, p_{i2}, ..., p_{ik}\}$ and a gold answer a_j contains l words $\{a_{j1}, a_{j2}, ..., a_{jl}\}$, the Word Overlap is defined as Equation 6:

$$WO(p_i, a_j) = \frac{|p_i \cap a_j|}{max(k, l)} \tag{6}$$

where $|A|$ denotes the number of element in the set A.

BERTScore(Zhang et al., 2020). BERTScore primarily calculates the semantic similarity between the candidate text and the reference text using cosine similarity. Given candicate text X with m tokens $\{x_1, x_2, ..., x_m\}$ and reference text Y with n tokens $\{y_1, y_2, ..., y_n\}$, BERTScore first computes the cosine similarity s_{ij} between each pair of token vectors x_i and y_j. Then, it maximizes the similarity score using a greedy matching approach to calculate the precision score $P(X, Y)$ and the recall score $R(X, Y)$. Finally, it computes the harmonic mean of these two scores (i.e., the F1 score) to obtain the final BERTScore. The above process can be represented by Equation 7-10. [7]

$$s_{ij} = \frac{H^{x_i} \cdot H^{y_j}}{||H^{x_i}||\,||H^{y_j}||} \tag{7}$$

$$P(X, Y) = \frac{1}{m} \sum_{i=1}^{m} \max_j s_{ij} \tag{8}$$

$$R(X, Y) = \frac{1}{m} \sum_{i=1}^{m} \max_j s_{ij} \tag{9}$$

$$BS(X, Y) = 2 \cdot \frac{P(X, Y) \cdot R(X, Y)}{P(X, Y) + R(X, Y)} \tag{10}$$

where H^{x_i} and H^{y_j} are the representations of x_i and y_j from a Pre-trained Language Model, $||a||$ denotes the length of the vector a.

B More Details of Experimental Setup

B.1 Datasets

MultiSpanQA and MultiSpanQA-Expand (Li et al., 2022) : MultiSpanQA and MultiSpanQA-Expand focus on multi-span questions. The raw questions and contexts are extracted from the Natural Question dataset (Kwiatkowski et al., 2019).

[7]For simplicity, Equation 7-10 do not consider inverse document frequency (idf) weighting or scaling of $R(c, r)$. For more details of BERTScore, please refer to (Zhang et al., 2020)

	#train	#dev	answer number prop.			avgerage answer number	average context length	avgerage question length
			≥2	1	0			
MultiSpanQA	5,230	658	100.0%	0.0%	0.0%	2.89	279	10
MultiSpanQA-Expand	15,690	1,959	33.4%	33.3%	33.3%	1.30	251	10
MAMRC	110,108	13,764	58.7%	41.3%	0.0%	1.77	69	10
MAMRC-Multi	64,625	8,081	100.0%	0.0%	0.0%	2.31	77	10

Table 7: Dataset statistics.

MultiSpanQA only contains multi-span questions, while MultiSpanQA-Expand contains both multi-span questions, single-span questions and unanswerable questions.

MAMRC and MAMRC-Multi (Yue et al., 2023) : MAMRC is a large-scale dataset containing over 100,000 questions, including both multi-span questions and single-span questions. To investigate the performance on the multi-span questions, we select multi-span questions from MAMRC and obtain MAMRC-Multi.

Since the official test sets of these four datasets are not public, we report the performance on validation sets. Some statistics about the four datasets are shown in Table 7.

B.2 MSQA models

MTMSN (Hu et al., 2019) : MTMSN adds a classification head to predict the number of answers. During the inference stage, for each question, MUSST first obtains top-20 predictions and predict answer number K, then applies Non-Maximum Sampling algorithm (Rosenfeld and Thurston, 1971) to extract K non-overlapped spans.

MUSST (Yang et al., 2021) : MUSST adds m linear layer to predict the start position and end position of m spans, where m is the maximum answer number in the training dataset. During the inference stage, MUSST applies an autogressive decoding strategy, where in each iteration MUSST masks out predicted spans and chooses top-1 predictions. The iterative process terminates when the model predicts no more answers or the number of predictions reaches the maximum answer number.

Tagger : Following the implementation of (Li et al., 2022), we utilize BIO tags to label each token in context: the first token of the answer is labeled with "B", the other tokens of the answer are labeled with "I" and the tokens not in an answer are labeled with "O".

SpanQualifier (Huang et al., 2023a) : SpanQualifier enumerates all possible answer spans and obtains their corresponding confidence scores as correct predictions, then utilizes a learnable threshold to select the correct prediction spans, achieving state-of-the-art performance on MultiSpanQA-Expand dataset.

BART (Lewis et al., 2020) and T5 (Raffel et al., 2020) : Both BART and T5 are pre-trained models with encoder-decoder architecture, which are commonly used in text generation tasks. In this work, we use the delimiter "#" to concatenate multiple answers and fine-tune the models in a sequence-to-sequence form.

GPT-3.5 : GPT-3.5 is one of the most commonly used LLMs today and can be accessed via API [8]. In our work, we select gpt-3.5-turbo-0120 for our experiments and set up both zero-shot and few-shot prompts. The zero-shot prompt contains only a basic description of the MSQA task, while the few-shot template includes several demonstrations. Specifically, we apply In-Context Learning (ICL) (Brown et al., 2020) and utilize a BM25 retriever (Robertson and Walker, 1994) to select the demonstrations which is similar to the questions. The prompts are shown in Table 12.

B.3 Implementation Details

To determine the hyper parameters α and β, we analyze the Word Overlap and BERTScore of the sampled data, shown in Table 9. For the middle 60% of sampled data, the Word Overlaps range from 0 to 0.25, and the BERTScore range from 0.36 to 0.62. Based on this, we set α to 0.25 and β to 0.6.

When sampling training data for ACC framework, we set split number $K = 3$, which means in each iteration, we use two-thirds of the training data for training and sample the predictions on the remaining data. for the classifier, we maintain a balanced ratio of 1:1:1 among the three answer categories for the classifier, and for the corrector,

[8] https://platform.openai.com/.

	MultiSpanQA			MultiSpanQA-Expand			MAMRC			MAMRC_Multi		
	PM P	PM R	PM F1	PM P	PM R	PM F1	PM P	PM R	PM F1	PM P	PM R	PM F1
Discriminative Models (BERT-base)												
MTMSN	69.97	79.23	74.30	73.29	73.46	73.37	84.59	89.62	87.03	84.68	89.97	87.25
+ACC	**81.10**	66.77	73.24	**77.20**	67.04	71.76	**88.45**	**85.86**	**87.13**	**90.74**	**85.68**	**88.14**
MUSST	76.39	68.76	72.38	77.79	70.99	74.22	87.25	88.25	87.74	87.75	87.69	87.72
+ACC	**81.25**	65.68	**72.64**	**78.36**	68.65	73.17	**88.68**	85.46	87.04	**90.93**	84.61	87.66
Tagger	78.27	77.92	78.09	70.60	65.75	68.05	88.81	89.05	88.92	88.23	84.98	86.57
+ACC	**83.30**	77.29	**80.19**	**74.06**	**66.64**	**70.14**	**89.07**	87.13	88.09	**90.85**	**83.54**	**87.04**
SpanQualifier	81.17	79.70	80.43	74.01	76.73	75.34	87.75	90.94	89.31	87.55	91.90	89.67
+ACC	**84.26**	77.70	80.84	**76.20**	75.15	**75.67**	**88.83**	87.94	88.38	**90.78**	88.37	89.56
Discriminative Models (RoBERTa-base)												
MTMSN	77.57	82.29	79.86	76.36	76.80	76.58	85.77	89.72	87.70	85.15	90.18	87.60
+ACC	**85.65**	72.12	78.30	**78.88**	69.93	74.14	**88.74**	86.21	87.46	**90.45**	**86.08**	**88.21**
MUSST	83.44	75.72	79.39	80.22	73.36	76.63	88.64	88.44	88.54	88.65	86.64	87.63
+ACC	**85.41**	73.24	78.86	79.99	70.83	75.13	**89.42**	85.95	87.65	**91.17**	83.89	87.38
Tagger	83.97	83.92	83.94	77.91	75.43	76.64	90.09	90.22	90.15	88.07	85.90	86.98
+ACC	**86.60**	82.67	**84.59**	**79.43**	74.62	**76.95**	89.92	89.01	89.46	**90.81**	84.20	**87.38**
SpanQualifier	83.85	83.17	83.50	76.77	78.62	77.65	89.82	88.19	89.00	87.27	92.14	89.63
+ACC	**86.39**	81.27	**83.74**	**78.69**	76.67	**77.66**	89.34	**88.98**	**89.16**	**90.49**	88.82	**89.65**
Generative Models												
BART-base	85.76	74.85	79.94	76.35	66.40	71.02	87.99	84.25	86.08	88.44	82.83	85.55
+ACC	**88.22**	72.85	79.81	**76.58**	64.26	69.88	87.99	84.25	86.08	88.18	80.91	84.39
T5-base	86.70	79.48	82.93	81.06	74.81	77.81	87.52	88.45	87.98	87.41	85.56	86.47
+ACC	**88.15**	76.88	82.13	**81.22**	72.22	76.46	86.69	85.99	86.34	87.40	83.22	85.26
GPT3.5 (Zeroshot)	85.70	79.64	82.56	57.62	66.03	61.54	61.10	73.70	66.81	72.27	77.71	**74.89**
+ACC	**89.64**	74.73	81.51	**64.38**	64.07	**64.23**	**67.19**	69.23	**68.19**	**78.24**	74.55	76.35
GPT3.5 (Fewshot)	88.19	81.28	84.59	59.67	70.81	64.76	72.47	85.98	78.65	79.27	87.04	82.98
+ACC	**90.76**	78.23	84.03	**67.07**	68.27	**67.67**	**77.55**	81.59	**79.52**	**83.54**	83.97	**83.75**

Table 8: PM scores on four MSQA datasets.

we added examples that require no modifications and maintained a ratio of 2:1 between examples requiring modifications and examples requiring no modifications, considering that corrector may not necessarily modifies all the input predictions.

During training stage of classifier and corrector, for MultiSpanQA and MultiSpanQA-Expand, we set $learning\_rate = 3e^{-5}$, $batch\_size = 48$, $epochs = 10$ and $max\_length = 512$; For MAMRC and MAMRC-Multi, we set $learning\_rate = 3e^{-5}$, $batch\_size = 96$, $epochs = 5$ and $max\_length = 256$. We choose the best classifier and corrector on our sliver-labeled validation sets. All the baselines were trained with three different seeds and we report the mean results. We perform our experiments on a single Tesla V-100 GPU(32GB).

C Additional Experiments and Discussions

C.1 Partial Match Results

The Partial Match results are shown in Table 8. While EM F1 scores show significant improvements after applying the ACC framework, PM F1 scores achieve less improvements and even decline in some cases. The main reason may be that PM scores consider the overlaps between predictions and gold answers, as a result, incorrect predictions may contribute to PM F1 score (i.e., $EM\ F1 = 0, PM\ F1 > 0$). However, such predictions are not desired and may be excluded by the ACC framework, limiting the improvements in PM F1 scores.

	Word Overlap	BERTScore
Min.	0.00	0.00
Max.	0.96	1.00
Avg.	0.11	0.49
Mid. 60% Range	(0.00,0.25)	(0.36,0.62)

Table 9: The distribution infomation of the sampled data on Word Overlap and BERTScore (metric we use to define partially correct prediction and wrong prediction). "Min." refers to the minimum value, "Max." refers to the maximum value, "Avg." refers to the average value, and "Mid. 60% Range" refers to the range of the middle 60% of the data.

C.2 Evaluation of the predictions with LLM

We utilize LLM to evaluate whether the predictions are closer to the gold answers after applying the ACC framework. For each dataset, we collect the predictions modified by the ACC framework and randomly sample 500 pairs (including original prediction, new prediction and gold answers for each pair) for evaluation. We mannully label four pairs as the few-shot demostractions for GPT-3.5. The prompts are shown in Table 13.

Datasets	original	new
MultiSpanQA	89 (17.8%)	411 (82.2%)
MultiSpanQA-Expand	98 (19.6%)	402 (80.4%)
MAMRC	98 (19.6%)	402 (80.4%)
MAMRC-Multi	92 (18.4%)	408 (81.6%)

Table 10: LLM evaluation on which prediction is closer to the gold answer, where "original" indicates that GPT-3.5 judge the original prediction to be closer and "new" indicates that GPT-3.5 judge the prediction modified by the ACC framework to be closer.

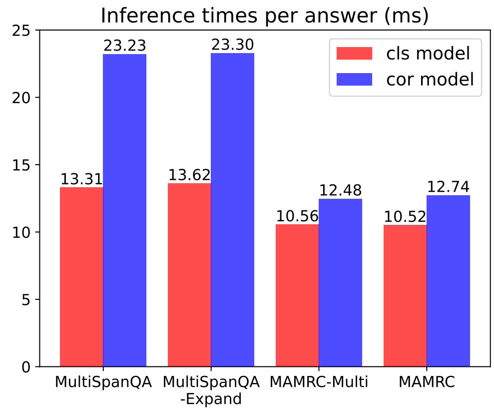

Figure 6: Inference times on four datasets.

The evaluation results are shown in Table 10. Across the four datasets, the LLM consider approximately 80% of the new preictions to be closer to the gold answers. This indicates that the ACC framework improves the quality of the predictions.

C.3 Model Size and Inference Time

We compare model sizes between MSQA models and the ACC framework, shown in Table 11. The ACC framework improves the performance of baselines without applying large-size models, avoiding consuming excessive computational resources.

We also analyze inference times of the ACC framework, shown in Figure 6. The results demostrate that the ACC framework is time-effective, especially when the input length is short (we set $max\_length = 256$ for MAMRC and MAMRC-Multi and we set $max\_length = 512$ for MultiSpanQA and MultiSpanQA-Expand).

C.4 Implementation details of pilot study with LLM

We use OpenAI's official API [9] and select the model gpt-3.5-turbo-0120 for our pilot study. Due to the poor performance in zero-show settings, we

[9] https://platform.openai.com/.

model	**BERT-base**	**RoBERTa-base**
MTMSN	110M	125M
MUSST	110M	125M
Tagger	109M	125M
SpanQualifier	115M	131M
classifier	-	128M
corrector	-	124M

Table 11: Model sizes of baselines model, the classifier and the corrector.

apply In-Context Learning (ICL) (Brown et al., 2020) and utilize a BM25 retriever (Robertson and Walker, 1994) to select the demonstrations which is similar to the questions. When replacing the classifier, we select one demonstration for each answer type; when replacing the corrector, we select two demostractions for answers requiring modification and requiring no modification. The prompts are shown in Table 14.

Instruction:
For this task, we will provide you a passage and a question. The question contains one or multiple answers and these answers are in the passage. You should first read the given passage and question, then extract answer spans from the passage and use "#" to split each answer spans, i.e. answer1#answer2#answer3.You should output your answer in a json format like "{"answer":"your_answer"}", DO NOT include any explanations in your responses.
Demostractions (for few-shot setting):
Example 1: Passage: ... Question: ... Answer: ... ...
Query:
Query: Passage: ... Question: ... Answer:

Table 12: Prompts for zero-shot LLM reader and few-shot LLM reader.

Instruction:
For this task, we will provide you with the gold answer to a question, the original prediction from our AI model, and a new prediction modified by another AI model. The question is from a QA dataset. You need to determine which prediction, the original or the new, is more accurate and closer to the gold answer.
Demostractions (for few-shot setting):
Example 1: Original Prediction: Billy New Prediction: Billy Jorl Gold: Billy Joel Answer: new
Query:
Original Prediction: ... New Prediction: ... Gold: ... Answer:

Table 13: Prompts for the evaluation on the predictions with LLM.

cls model prompt:
For this task, we will provide you a passage and a question. The question contains one or multiple answers and these answers are in the passage. We will also provide you a candidate answer from our AI model. You should read the passage, the question and classify the candidate answer into one of three classes: "correct prediction", "partially correct prediction" and "wrong prediction". Correct prediction refers to a completely correct prediction; Partially correct prediction refers to a prediction that is basically correct but still requires some modifications. Wrong prediction refers to a prediction that is completely incorrect and should be excluded.\You should output your answer in a json format like "{{"answer":"your_answer"}}", DO NOT include any explanations in your responses. Example1: Passage: ... Question: ... Candidate Answer:... Output: {"answer":"correct prediction"} ... Query: Passage: ... Question: ... Candidate Answer: ... Output:
cor model prompt:
For this task, we will provide you a passage and a question. The question contains one or multiple answers and these answers are in the passage. We will also provide a candidate answer that our AI model believes needs some modifications. You should read the passage, the question and judge whether the candidate answer requires modifications. If no modifications are needed, you should output the candidate answer as is. Otherwise, you should modify it by adding or deleting some words, and the modified prediction must be a part of the passage and similar to the original candidate answer.\You should output your answer in a json format like "{{"answer":"your_answer"}}", DO NOT include any explanations in your responses. Example1: Passage: ... Question: ... Original Answer: ... Output: {"answer":"xxx"} ... Query: Passage: ... Question: ... Candidate Answer: ... Output:

Table 14: Prompts for pilot study with LLM