Model	Dev		Test	
	EM	F1	EM	F1
Existing Systems				
BARB (**?**)	36.1	49.6	34.1	48.2
mLSTM (**?**)	34.4	49.6	34.9	50.0
BiDAF (**?**)	-	-	37.1	52.3
R2-BiLSTM (**?**)	-	-	43.7	56.7
AMANDA (**?**)	48.4	63.3	48.4	63.7
DECAPROP (**?**)	52.5	65.7	53.1	66.3
BERT (**?**)	-	-	46.5	56.7
NeurQuRI (**?**)	-	-	48.2	59.5
Our implementation				
ALBERT (+TAV)	57.1	67.5	55.3	65.9
Retro-Reader on ALBERT	**58.5**	**68.6**	**55.9**	**66.8**
ELECTRA (+TAV)	56.3	66.5	54.0	64.5
Retro-Reader on ELECTRA	56.9	67.0	54.7	65.7

Table 3: Results (%) for NewsQA dataset. The results except ours are from **?** (**?**) and **?** (**?**). TAV: threshold based answerable verification (§3.2).

to SQuAD2.0 online evaluation), which are respectively the ALBERT and ELECTRA based retrospective reader composed of both sketchy and intensive reading modules without question-aware matching for simplicity. According to the results, we make the following observations:

1) Our implemented ALBERT and ELECTRA baselines show the similar EM and F1 scores with the original numbers reported in the corresponding papers (**??**), ensuring that the proposed method can be fairly evaluated over the public strong baseline systems.

2) In terms of powerful enough PrLMs like ALBERT and ELECTRA, our Retro-Reader not only significantly outperforms the baselines with p-value < 0.01,[8] but also achieves new state-of-the-art on the SQuAD2.0 challenge.[9]

3) The results on NewsQA further verifies the general effectiveness of our proposed Retro-Reader. Our method shows consistent improvements over the baselines and achieves new state-of-the-art results.

5 Ablations

5.1 Evaluation on Answer Verification

Table 4 presents the results with different answer verification methods. We observe that either of the front verifiers boosts the baselines, and integrating both as rear verification works the best. Note that we show the HasAns and NoAns only for completeness. Since the final predictions are based on the threshold search of answerability scores (§3.2), there exists a tradeoff between the HasAns and NoAns accuracies. We

[8]Besides the McNemar's test, we also used paired t-test for significance test, with consistent findings.

[9]When our models were submitted (*Jan 10th 2020* and *Apr 05, 2020* for ALBERT- and ELECTRA-based models, respectively), our Retro-Reader achieved the first place on the SQuAD2.0 Leaderboard (https://rajpurkar.github.io/SQuAD-explorer/) for both single and ensemble models.

Model	All		HasAns		NoAns	
	EM	F1	EM	F1	EM	F1
BERT	78.8	81.7	74.6	80.3	83.0	83.0
+ E-FV	79.1	82.1	73.4	79.4	84.8	84.8
+ I-FV-CE	78.6	82.0	73.3	79.5	84.5	84.5
+ I-FV-BE	78.8	81.8	72.6	78.7	85.0	85.0
+ I-FV-MSE	78.5	81.7	73.0	78.6	84.8	84.8
+ RV	79.6	82.5	73.7	79.6	85.2	85.2
ALBERT	87.0	90.2	82.6	89.0	91.4	91.4
+ E-FV	87.4	90.6	82.4	88.7	92.4	92.4
+ I-FV-CE	87.2	90.3	81.7	87.9	92.7	92.7
+ I-FV-BE	87.2	90.2	82.2	88.4	92.1	92.1
+ I-FV-MSE	87.3	90.4	82.4	88.5	92.3	92.3
+ RV	87.8	90.9	83.1	89.4	92.4	92.4

Table 4: Results (%) with different answer verification methods on the SQuAD2.0 dev set. *CE*, *BE*, and *MSE* are short for the two classification and one regression loss functions defined in §3.2.

Method	Prec.	Rec.	F1	Acc.
ALBERT	91.70	93.42	92.55	86.14
Retro-Reader on ALBERT	**94.30**	92.38	93.33	87.49
ELECTRA	92.71	92.58	92.64	86.30
Retro-Reader on ELECTRA	93.27	**93.51**	**93.39**	**87.60**

Table 5: Performance on the unanswerable questions from SQuAD2.0 dev set.

Method	EM	F1
ALBERT	87.0	90.2
Two-model Ensemble	87.6	90.6
Retro-Reader	87.8	90.9

Table 6: Comparisons with Equivalent Parameters on the dev set of SQuAD2.0.

Method	SQuAD2.0		NewsQA	
	EM	F1	EM	F1
BERT	78.8	81.7	51.8	62.5
+ CA	78.8	81.7	52.1	62.7
+ MA	78.3	81.4	52.4	62.6
ALBERT	87.0	90.2	57.1	67.5
+ CA	87.3	90.3	56.0	66.3
+ MA	86.8	90.0	55.8	66.1

Table 7: Results (%) with matching interaction methods on the dev sets of SQuAD2.0 and NewsQA.

see that the final RV that combines E-FV and I-FV shows the best performance, which we select as our final implementation for testing.

We further conduct the experiments on our model performance of the 5,945 unanswerable questions from the SQuAD 2.0 dev set. Results in Table 5 show that our method improves the performance on unanswerable questions by a large margin, especially in the primary F1 and accuracy metrics.

5.2 Comparisons with Equivalent Parameters

When using sketchy reading module for external verification, we have two parallel modules that have independent parameters. For comparisons with equivalent parameters, we add an ensemble of two baseline models, to see if the advance is purely from the increase of parameters. Table 6 shows the results. We see that our model can still outperform two ensembled models. Although the two modules share the same design of the Transformer encoder, the training objectives (e.g., loss functions) are quite different, one for answer span prediction, the other for answerable decision. The results indicate that our two-stage reading modules would be more effective for learning diverse aspects (verification and span prediction) for solving MRC tasks with different training objectives. From the two modules, we can easily find the effectiveness of either the span prediction or answer verification, to improve the modules correspondingly. We believe this design would be quite useful for real-world applications.

5.3 Evaluation on Matching Interactions

Table 7 shows the results with different interaction methods described in §3.2. We see that merely adding extra layers could not bring noticeable improvement, which indicates that simply adding more layers and parameters would not substantially benefit the model performance. The results verified the PrLMs' strong ability to capture the relationships between passage and question after processing the paired input by deep self-attention layers. In contrast, answer verification could still give consistent and substantial advance.

Passage: *Southern California consists of a heavily developed urban environment, home to some of the largest urban areas in the state, along with vast areas that have been left undeveloped. It is the third most populated megalopolis in the United States, after the Great Lakes Megalopolis and the Northeastern megalopolis. Much of southern California is famous for its large, spread-out, suburban communities and use of automobiles and highways. The dominant areas are Los Angeles, Orange County, San Diego, and Riverside-San Bernardino, each of which are the centers of their respective metropolitan areas...*
Question: *What are the second and third most populated megalopolis after Southern California*?
Answer: **Gold:** ⟨no answer⟩ **ALBERT (+TAV)**: Great Lakes Megalopolis and the Northeastern megalopolis. **Retro-Reader over ALBERT:** ⟨no answer⟩ $score_{has} = 0.03, score_{na} = 1.73, \delta = -0.98$

Table 8: Answer prediction examples from the ALBERT baseline and Retro-Reader.

5.4 Comparison of Predictions

To have an intuitive observation of the predictions of Retro-Reader, we give a prediction example on SQuAD2.0 from baseline and Retro-Reader in Table 8, which shows that our method works better at judging whether the question is answerable on a given passage and gets rid of the plausible answer.

6 Conclusion

As machine reading comprehension tasks with unanswerable questions stress the importance of answer verification in MRC modeling, this paper devotes itself to better verifier-oriented MRC task-specific design and implementation for the first time. Inspired by human reading comprehension experience, we proposed a retrospective reader that integrates both sketchy and intensive reading. With the latest PrLM as encoder backbone and baseline, the proposed reader is evaluated on two benchmark MRC challenge datasets SQuAD2.0 and NewsQA, achieving new state-of-the-art results and outperforming strong baseline models in terms of newly introduced statistical significance, which shows the choice of verification mechanisms has a significant impact for MRC performance and verifier is an indispensable reader component even for powerful enough PrLMs used as the encoder. In the future, we will investigate more decoder-side problem-solving techniques to cooperate with the strong encoders for more advanced MRC.

Vitae maxime ullam harum, corporis amet quod