How State-Of-The-Art Models Can Deal With Long-Form Question Answering

Minh-Quan Bui, Vu Tran, Ha-Thanh Nguyen, Le-Minh Nguyen
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
{quanbui,vu.tran, nguyenhathanh, nguyenml}@jaist.ac.jp

Abstract

Question answering is an essential task in natural language processing. According to our review, the datasets for this task often contain short answers. We raise a research question of whether state-of-the-art models can perform well when a longer answer is needed. We propose a dataset that contains much longer answers called FitQA[1] and conduct a brief performance analysis among current state-of-the-art models on this dataset. The robust transformer architecture like ALBERT achieved 90.9% F1 on SQuAD 2.0 but only got 47.3% F1 score on FitQA. Our hypothesis is that for longer context, the model needs to be guided to focus on longer dependent words. We conduct a curriculum-learning-based framework. Experimental results show that our approach could improve the performance with the appropriate answer length up to 55.3% on F1.

Acknowledgments: This work was supported by JST CREST Grant Number JPMJCR1513 and the Asian Office of Aerospace R&D (AOARD), Air Force Office of Scientific Research (Grant no. FA2386-19-1-4041)

1 Introduction

Machine reading comprehension (MRC), or the ability to read and understand the unstructured text and answer questions about it remains a challenging task in natural language understanding. This challenge spurs the development of large datasets and deep learning architectures. The current state-of-the-art models can overwhelm the human performance. RoBERTa (Liu et al., 2019) performs well with 89.4%F1 score on SQuAD. Besides, A Lite BERT(ALBERT) (Lan et al., 2019) is even better when getting 91.365% F1 score and 88.716% exact match(EM) on SQuAD 2.0. In contrast, the human performance only gets 89.5% F1 score and 86.8% EM. Also, Microsoft has already built a high-quality dataset called NewsQA (Trischler et al., 2017), a challenging dataset with more than 100.000 question-answer pairs, But a new improving method

Article: 5 Fat Loss Myths You Still Believe
Question: How to Limit Metabolic Compensation?

Context: How to Limit Metabolic Compensation The good news is there are some ways to reduce metabolic compensation. Here are some things to do: **Do your best to maintain as much muscle as you can. The metabolic rate will not slow as much and be more resistance to fat regain. This means to make weight lifting the dominant part of your fitness regime during fat loss. Cardio becomes a little more important after weight loss, when the metabolic rate has lessened. You may want to save your cardio for after, rather than during the competition diet. Eat more protein, see the first point above about maintaining muscle mass. And probably increase the amount of protein as a percent of total calories. Do this during, but perhaps more importantly, after fat loss**. Cycle the calorie gap, having times where you're in a strong deficit and other times where you're in no deficit at all. The recent MATADOR study (minimizing adaptive thermogenesis and deactivating obesity rebound) showed this strategy got better results, had less metabolic adaptation, and much longer lasting results. Don't eat like an asshole when it all ends. Focus on blander foods and less variety of them. Doing the traditional burger, pizza, and cheesecake binges will trigger the brain's hedonistic response and cause you to want more of that same dopamine hit all this when the metabolism is at its most vulnerable in terms of fat storage. And finally you may want to consider some type of adaptogen like rhodiola or ashwagandha. I have no studies to back this up, but I have very good success clinically with using these herbs along with the recommendations above to keep the command and control center of the metabolism (the brain's hypothalamus) stress-resistant and happy.

Figure 1: Example for FitQA dataset with the answer is in bold

[1]The dataset will be published along with the paper

for BERT also known by the name SpanBERT (Joshi et al., 2020), which masking spans instead of token masks and the performance when applied to NewsQA, has 73.6% F1 score on NewsQA. In Robin Jia's work (Jia and Liang, 2017), his proposed method test whether a model can give a correct answer while paragraphs contain some additional sentences, which are noise for deep learning models. This method worked well by decreasing the accuracy of sixteen models drops from 75% F1 score to 36%. SQuAD-Open is built by Chen (Chen et al., 2017), an open domain question answering dataset and contains only question and answer. The model has to extract the response by the relevant context from Wikipedia articles. This problem seems to be a trend in question answering datasets. While deep learning models are becoming more powerful and reaching human performance, more complex datasets are needed to accelerate method and model development.

We build the FitQA dataset to contribute to not only computer science but also for the entire society. FitQA is collected by crawling more than 200 articles from `bodybuilding.com` (bod, 1999) and `t-nation.com`(LLC, 1998). The main purpose of this dataset is for the health of society. On the internet, we have a lot of fakes and lack of information news about nutrition and training like 30 days sit up for sick-pack or deltoid drinking for losing weight, etc. We cannot explain knowledge with a short sentence, that is why we want FitQA to be very detailed and diverse in answer. Figure 1 shows an example of the phenomena of FitQA. We experiment with different models and find that FitQA creates a significant challenge to current comprehension models. In this paper, we also describe curriculum learning (Bengio et al., 2009), a training approach for state-of-the-art models with the maximum of answer length (MAL) is 30 and 60 to handle low resource limitation.

2 Related Datasets

FitQA is built following the format of some traditional comprehension datasets. These vary in length, size, problem, collection, and each has its distinctive feature.

2.1 NewsQA

NewsQA (Trischler et al., 2017), a machine comprehension dataset with more than 100K question-answer pairs. These pairs are created by human and based on a set of over 12K news articles from CNN. The answers consist of spans of text in the articles. In NewsQA, they created some conditions for answers to make the dataset more complex:

1. Word Matching: The similarities between questions and answers are low; this condition makes deep learning architecture harder to extract the answer from the article.

2. Paraphrasing: In each article, it must contain one sentence that can answer the question. This sentence requires synonym and global knowledge.

3. Inference: Their answers must be found from a piece of information in the article or by overlap.

4. Synthesis: The answers are only found by the assumption of information through several sentences.

5. Ambiguous/Insufficient: Some questions do not have answers in the article.

Because of the complexity, this dataset is challenging for transformer architectures. To overcome this challenge, SpanBERTJoshi et al. (2020) was developed with a new training approach by masking random spans instead of random tokens and training the span boundary representation for predicting the whole content of the masked span, without depending on the token representations within it. SpandBERT achieved 73.6% F1 on NewsQA, that 83.6% F1 on TriviaQA, 84.8% F1 on Search QA and more significant results on other datasets.

2.2 TriviaQA

TriviaQA (Joshi et al., 2017) was collected into 650K question-answer-evidence triples. TriviaQA has over 95K question-answer pairs and at least six evidence documents for each question-answer pair. In this dataset, up to 92% of the answers are the article titles in Wikipedia, about 4% are numerical answers, and the rest are free texts. The challenge in TriviaQA is the overlap of each example in sev-

Length(word)	Proportion
1-10	17.7%
11-20	22.7%
21-30	12%
31-40	10%
41-50	12.8%
51-60	10%
61-70	4.9%
> 70	9.9%

Table 1: Length statistics.

eral categories. It means each question-answer pair can be found in multiple evidence documents. Right now, the first place on the TriviaQA leaderboard is 83.99% on F1 score.

2.3 SQuAD 2.0

SQuAD 2.0, also called SQuADRUN, is created base on SQuAD, but higher difficulty which contains more than 130K examples in over 442 articles. Beside answerable questions, it has more than 50K unanswerable questions that look similar to answerable questions. To perform well on SQuAD 2.0, the models also need to decide to answer the question or not when the context does not support the answer. Despite the fact that many challenges in SQuAD 2.0 dataset, the state-of-the-art model can surpass 90% on F1 score.

3 FitQA

For this work, we analyzed the data collected from `bodybuilding.com` (bod, 1999) and `t-nation.com`(LLC, 1998). These contain a varied topic that includes nutrition, training, diet, fat loss, etc. FitQA focuses on topics that people are often interested in like nutrition or training. By observing the habit of the people asking questions from some popular forums, we can conclude that long answers usually satisfy the questioner better because it contains more relevant and useful information. However, sometimes, a short answer is all they need. Based on that observation, we build FitQA as a data set of variable length answers. FitQA has almost 700 question-answer pairs, the length of an answer is from 1 to 139 words. FitQA follows the format of SQuAD 2.0 dataset, and answers are extracted from spans of text in the article. The different and challenges that make it different from SQuAD 2.0 are as below:

1. The average length of articles in FitQA is double of that in SQuAD 2.0.

2. The average length of answers in FitQA is ten times higher than SQuAD 2.0.

The state-of-the-art models are overwhelming human performance, and the dataset must be harder and more challenges to be able to achieve some important achievements in machine comprehension task. There are some participants in this case, NewsQA (Trischler et al., 2017) have more challenges than SQuAD (Rajpurkar et al., 2016) by having less word matching examples(7.1%), more paraphrasing example(7.3%) and more synthesis and inference examples(13.4%). On the other hand, TriviaQA has 69% questions that have different syntactic structure and 41 % of them have lexically different. Moreover, the information needed to answer the question is scattered over multiple sentences. Based on these ideals, we increase the complexity of FitQA by the diversity in answer length. The length statistics is showed as Table 1. To test the performance of state-of-the-art models, we create the test set by picking 100 examples with varied length, and the rest is for the training set.

4 Curriculum Learning

We examine some previous work and propose a useful method for handling the long articles and answers. Curriculum learning (Bengio et al., 2009) is a learning strategy in machine learning, we let the deep learning model learn with easy examples first and then gradually handles harder cases. Several works have shown that this problem can be overcome by using this learning strategy. As a result of Cao Liu (Liu et al., 2018) in his natural answer generation task, curriculum learning can increase his model performance by 6.8% and 8.7% in the accuracy for easy and hard questions.

In our question-answering task, we defined the complexity by the length of the answer. We assume that an example containing a short answer is easy, and an example having a long answer is difficult. We want models to learn from easy to difficult sample

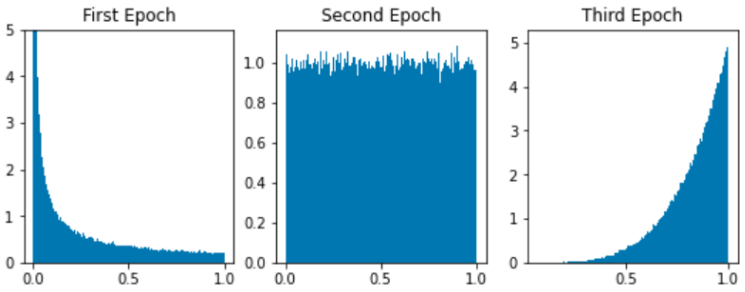

Figure 2: Illustration of the probability of picking example by curriculum learning with 3 epochs and temperature base $\gamma = 5$

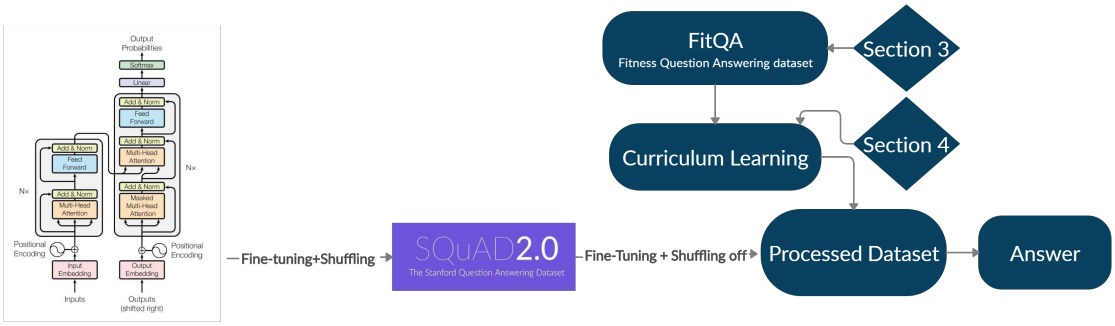

Figure 3: The overall diagram of question-answering task bases on transformer model and curriculum learning

and from the difficult to easy sample. We give every single example a score equal to its answer length. After that, we sort the list of examples in ascending order by scores to pop the sample by its index easily. The formula and pseudo-code of curriculum learning (Algorithm 1) to calculate the probability of picking an example $Index$ from n examples as below:

$$index_{ij} = \lfloor nx_{ij}^{t_i} \rfloor \tag{1}$$

$$x_{ij} \sim U[0, 1) \tag{2}$$

where t_i is the temperature of epoch i^{th} obtained by:

$$t_i = \gamma^{\alpha_i} \tag{3}$$

$$\alpha_i = \frac{2(1-i)}{N_{epochs} - 1} + 1 \tag{4}$$

where γ is a temperature base, N_{epochs} is the total number of training epoch. The temperature base γ reflects how high the probability of picking a long or short example through the epoch.

We illustrate the probability of picking example curriculum learning as Figure 2. In the first epoch,

Algorithm 1: Curriculum learning pseudo-code

Result: Sample by length
Input:
The list of n_{sample} samples is sorted in ascending order:SL;
γ is temperature base;
N_{epochs} is total number of epochs;
Output:;
The list of n_{sample} samples is arranged by curriculum learning:OL;
Function: curriculum_learning(SL,γ,N_{epochs}):
OL $\leftarrow$ {}
for *i = 1 to* N_{epochs} **do**
  $temp_{Li}$ = SL
  $\alpha_i = \frac{2(1-i)}{N_{epochs}-1} + 1$
  $t_i = \gamma^{\alpha_i}$
  for *j = 0 to* n_{sample} **do**
    x_{ij} = random(0,1)
    $index_{ij}$ = length($temp_{Li}$)
    OL.push($temp_{Li}[index_{ij}]$)
    $temp_{Li}$.pop($index_{ij}$)
  end
end

Table 2: *F1 and Exact Match(EM) scores on FitQA with MAL=30*

Model	FitQA uniform		SQuAD 2.0+FitQA uniform		SQuAD 2.0+FitQA γ = 2		SQuAD 2.0+FitQA γ = 3		SQuAD 2.0+FitQA γ = 5	
	EM(%)	***F1(%)***	***EM(%)***	***F1(%)***	***EM(%)***	***F1(%)***	***EM(%)***	***F1(%)***	***EM(%)***	***F1(%)***
albert-base v1	11.1	34.9	16.9	42.5	15.3	43.1	16.7	44.0	15.0	42.1
albert-base v2	12.1	37.8	15.3	42.5	14.3	43.5	14.4	43.2	17.0	**45.6**
bert-base-uncase	6.9	31.1	17.8	45.0	17.8	45.0	17.0	44.2	16.7	**46.1**
roberta-base	5.6	25.1	14.1	42.7	14.3	**45.5**	14.7	42.3	14.4	43.8

Table 3: *F1 and Exact Match(EM) scores on FitQA with MAL=60*

Model	FitQA uniform		SQuAD 2.0+FitQA uniform		SQuAD 2.0+FitQA γ = 2		SQuAD 2.0+FitQA γ = 3		SQuAD 2.0+FitQA γ = 5	
	EM(%)	***F1(%)***	***EM(%)***	***F1(%)***	***EM(%)***	***F1(%)***	***EM(%)***	***F1(%)***	***EM(%)***	***F1(%)***
albert-base v1	8.8	38.6	18.3	51.3	15.7	51.3	16.7	50.8	17.0	51.4
albert-base v2	17.0	47.3	21.2	54.2	16.7	51.5	19.3	52.0	18.3	53.0
bert-base-uncase	6.9	31.1	17.0	52.6	15.4	51.4	20.33	53.9	17.7	52.1
roberta-base	4.9	30.9	19.0	52.7	18.0	53.3	20.3	53.7	19.0	**55.3**

we can easily see that the probability of picking the examples with short answer is high and it is pretty low with examples with long answer. In the second epoch, the probability of picking example is uniform distribution. In the last epoch, the probability of picking the examples with long answer is extremely higher than the short answer examples.

5 Experiment

5.1 Experiment Settings

From SQuAD and NewsQA leaderboard, there are some approaches perform better performance, but all of them are built base on the pre-trained model. To test FitQA for the machine comprehension task, we compare the performance of four common pre-trained deep learning models: bidirectional encoder representations from transformers (BERT), two versions of a Lite BERT (ALBERT), and RoBERTa. We describe details of all the pre-trained models as below:

1. bert-base-uncased: 12 layer, 768 hidden, 12 heads, 110M parameters, and trained on low-cased english text.

2. albert-base-v1: 12 layer, 768 hidden, 128 embedding, 12 heads, 11M parameter.

3. albert-base-v2: 12 layer, 4096-hidden, 128 embedding, 64-heads, 223M parameters.

4. roberta-base:12-layer, 768 hidden, 12-heads, 125M parameters RoBERTa using the BERT-base architecture.

We conduct experiments on SQuAD, FitQA. Performance on these datasets is measured by exact match (EM) and per answer token-based F1 score, which was published by Rajpurkar et al(2016) (Rajpurkar et al., 2016). The detailed settings are described as below:

1. **FitQA uniform:** Using 4 pre-trained models to test the performance on FitQA with uniform probability of picking examples.

2. **SQuAD 2.0 + FitQA uniform:** Using 4 pre-trained models, we first fine-tune the models on SQuAD 2.0, then fine-tune the models again on FitQA with uniform probability of picking examples.

3. **SQuAD 2.0 + FitQA (Curriculum Learning):** Using 4 pre-trained models, we first fine-tune the models on SQuAD 2.0 with uniform probability of picking examples, then fine-tune the models again on FitQA with curriculum learning with different γ (2, 3 and 5).

Figure 3 can illustrate the whole process of setting 2 and 3.

5.2 Main Result

According to information from SQuAD 2.0 leaderboard, the best performance that albert single can archive is 88.592% EM score and 91.286% F1 score. However, in section III, we showed that the length of some examples in FitQA are extremely long and diverse. This is the reason makes 4 state-of-the-art models cannot work well on FitQA. In our experiments, albert-base-v2 has the best result but only 37.8% F1 score and 12.1% on EM. SQuAD 2.0 is the most similar dataset to FitQA. To maximize the performance, we firstly train all models on SQuAD 2.0 and fine-tune FitQA. After training on SQuAD and fine-tune FitQA the performance increase 5.68% on EM and 8.9%F1 score on average. Next, we mute the shuffling feature, then apply curriculum learning to the training set. We start with temperature base γ = 2 and MAL=30. As the results in Table 2, curriculum learning made average F1 score from 43.2% to 44.3%. Especially, it can increase the performance of roberta-base by 2.8%. With γ = 3, there are no significant improvement. We increase γ base γ to 5, and we can get the best results with 46.1% and 45.6% on bert-base-uncase and albert-base-v2. Next, we want model to face the harder challenge by increasing MAL to 60, and it leads to good results, roberta-base gets the best result overall with 55.3% on F1 even if the improvement is negligible.

5.3 Result Analysis

With MAL=30,Bert-base-uncased and albert-base v2 with temperature base γ = 5 seems to be the best for FitQA, so we analyze the results and compare them to bert-base-uncase without curriculum learning. As a result of Table 3, we show the accuracy of the answer group was mentioned in Table 1. By comparing the results from these settings, we expect to determine that curriculum learning is useful for extracting more text or capture more related information to answer the question. In lengths from 0 to 10 words, we can see there is no significant change between all settings. Starting from 11 words, we can see that these results go beyond the uniform distribution setting. With temperature base γ = 5 from Table 5, we can see bert-base-uncase works well in lengths from 11 to 60 words, and the performance in this range increase 2.08% on average compare to uniform distribution bert-base-uncase. Albert-base-v2 can also perform well in this range with 3.44% increase in total. TABLE 6 summarizes overall statistics of 3 best settings on FitQA with MAL=60. It is worth discussing these interesting facts revealed by the results of bert-base-uncased . The test in range 41 to more than 70 words found differences from bert-base-uncase compare to albert-base-v2 with 2.8% improvement on F1 score.

One limitation is found in these experiments. From TABLE 4, the most extended answer can be extracted is 50 words. It means for the examples have answer more than 50 words, the EM score will be zero. Not only that, but it is also hard to extract long answer correctly from the context, and some answers are a subset of gold answers. This may be the reason why EM score equal to 0 in some evaluations. We show several examples to demonstrate for this limitation in Table 7. The answers are extracted by models is not wrong, but not enough in these cases.

6 Conclusion

As the results are shown in Table 5 and Table 6, we have succeeded in improving the length that the model can extract by applying curriculum learning on the training set. This success leads to an increase in F1 score. The problem is all the state-of-the-art models perform poorly under long answer form dataset. The best result that these models can get is just 21.2% on EM and 55.3% on F1 score. We believe that the long-form answer dataset is the big challenge for machine comprehension task. Further, we want to apply curriculum learning not only base on the length of the answer but also other features to solve the low performance of state-of-the-art models on FitQA.

References

Bodybuilding. 1999. URL `https://www.bodybuilding.com`. Accessed:2020-1-25.

Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

Model	Total Tokens	Longest Answer	Shortest Answer
bert-base-uncase γ = 5,MAL=30	1462	28	1
albert-base-v2 γ = 5,MAL=30	1384	28	1
bert-base-uncase uniform,MAL=30	1271	28	1
albert-base-v2 uniform,MAL=60	2255	46	1
roberta-base γ = 5,MAL=60	1980	42	1
bert-base-uncase γ = 3,MAL=60	2330	50	1

Table 4: *Total number of tokens, longest and shortest answer that models can extract from 100 examples of test set*

Length	bert-base-uncase γ = 5		albert-base-v2 γ = 5		bert-base-uncase uniform	
	EM(%)	*F1(%)*	*EM(%)*	*F1(%)*	*EM(%)*	*F1(%)*
1-10	41.7	62.4	47.9	58.5	50	63.4
11-20	37.0	**59.1**	40.7	**64.0**	38.8	57.0
21-30	20.5	**52.9**	10.3	46.3	20.5	50.9
31-40	0.0	43.4	0.0	**48.3**	0.0	44.8
41-50	0.0	**34.6**	0.0	**35.0**	0.0	30.4
51-60	0.0	**37.7**	0.0	**40.9**	0.0	34.2
61-70	0.0	34.7	0.0	32.4	0.0	37.7
>70	0.0	23.8	0.0	21.4	0.0	24.3

Table 5: *F1 and Exact Match(EM) scores on best settings base on length with MAL=30*

Length	albert-base-v2 uniform		roberta-base γ = 5		bert-base-uncase γ=3	
	EM(%)	*F1(%)*	*EM(%)*	*F1(%)*	*EM(%)*	*F1(%)*
1-10	35.4	49.4	47.9	**67.9**	29.1	47.5
11-20	37.0	63.9	25.9	62.2	44.4	59.9
21-30	46.2	67.2	35.9	61.5	30.8	65.0
31-40	30.3	61.2	15.1	53.4	21.1	57.5
41-50	0.0	44.5	0.0	**46.1**	0.0	**46.9**
51-60	0.0	56.6	0.0	**56.9**	0.0	**56.8**
61-70	0.0	47.2	0.0	41.8	0.0	**50.0**
>70	0.0	35.8	0.0	36.3	0.0	**41.6**

Table 6: *F1 and Exact Match(EM) scores on best settings base on length with MAL=60*

D. Chen, A. Fisch, J. Weston, and A. Bordes. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1171. URL https://www.aclweb.org/anthology/P17-1171.

R. Jia and P. Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1215. URL https://www.aclweb.org/anthology/D17-1215.

M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://www.aclweb.org/anthology/P17-1147.

M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

C. Liu, S. He, K. Liu, and J. Zhao. Curriculum learn-

Question	Gold Answer	Predicted Answer
Are Testosterone Boosters Safe?	Always read re-views before purchasing, and choose a testosterone booster from a reputable, established supplement company. Only take the recommended dose and keep your doctor in the loop about what you're taking if you have other health concerns or take medications.	Always read reviews before purchasing, and choose a testosterone booster from a reputable, established supplement company.
what is the difference between brown fat and white fat?	Both types store energy, but white fat cells each contain only one droplet of fat, while brown fat contains lots of tiny droplets of fat. Brown fat also contains tons of the brownish cellular organelles known as mitochondria, which use the droplets of fat to create energy and, as a byproduct of creating energy, heat.	Both types store energy, but white fat cells each contain only one droplet of fat, while brown fat contains lots of tiny droplets of fat.

Table 7: *2 Examples for the most limitation of FitQA*

ing for natural answer generation. In *IJCAI*, pages 4223–4229, 2018.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

T. N. LLC. Bodybuilding. 1998. URL `https://www.t-nation.com/`. Accessed: 2020-2-02.

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, Nov. 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL `https://www.aclweb.org/anthology/D16-1264`.

A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2623. URL `https://www.aclweb.org/anthology/W17-2623`.