Homework Assignment

CMU 11-797: Question Answering

START HERE: Instructions

- **Collaboration policy:** Collaboration on solving the homework is allowed, after you have thought about the problems on your own. It is also OK to get clarification (but not solutions) from other students or online resources, again after you have thought about the problems on your own. There are two requirements: first, cite your collaborators fully and completely (e.g., "Jane explained to me what is asked in Question 2.1"). Second, write your solution *independently*: close the online resources, and send collaborators out of the room, so that the solution comes from you only. **For this assignment, you do not need to implement the models from scratch. You can use any existing implementation of the models.**
- **Late Submission Policy:** There is **NO** late days for this assignment. You should be able to complete the assignment within a week of its release.
- **Submitting your work:**
  - **Gradescope:** For the written problems such as short answer, multiple choice, derivations, proofs, or plots, we will be using the Gradescope's 'Written' submission slot. If you are using latex, please use the provided template. The best way to format your homework is by using the Latex template released in the handout and writing your solutions in Latex, please use the soln environment provided so we can easily see your solution.

    For the programming problem you are also required to submit your code to the 'Programming' submission slot on Gradescope. Please include your submissions in a .zip file. Your code must be in a run-able state as TAs may be required to run your code. If you do not submit your code, you will be given a score of 0 on all of the programming sections in the written section.

    Regrade requests can be made after the homework grades are released, however this gives the TA the opportunity to regrade your entire paper, meaning if additional mistakes are found then points will be deducted.

Dataset and Model Overview

1 Dataset Description

For this assignment, you need to work with three datasets described in the following subsections.

1.1 SQuAD 2.0 [3]: Span Extraction Task

SQuAD presents an extractive reading comprehension task where there will be a context passage with each question and the task would be to find the answer span from the context. Unlike SQuAD 1.0 [4], SQuAD 2.0 includes questions that are unanswerable given the context. These questions are written adversarially by crowdworkers to look similar to answerable ones. The training dataset contains 87k answerable and 43k unanswerable questions.

1.2 Dream [6]: Dialogue Based QA Task

DREAM proposes a dialogue-based multiple-choice reading comprehension task that focuses on in-depth multi-turn multi-party dialogue understanding. The dataset contains 10,197 multiple-choice questions for 6,444 dialogues where each question is associated with three answer options, exactly one of which is correct. DREAM is likely to present significant challenges for existing reading comprehension systems: 84% of answers are non-extractive, 85% of questions require reasoning beyond a single sentence, and 34% of questions also involve commonsense knowledge.

2 Model Description

In this assignment, you will explore both non-LM and LM models. There will be no implementation from scratch. You can use existing implementations of these models and run experiments with the datasets mentioned above.

2.1 BIDAF [5]

Bi-Directional Attention Flow (BIDAF) network is a hierarchical multi-stage architecture for modeling the representations of the context paragraph at different levels of granularity. BIDAF includes character-level, word-level, and contextual embeddings, and uses bi-directional attention flow to obtain a query-aware context representation. This architecture outperforms various LSTM and GRU-based models in Stanford Question Answering Dataset (SQuAD) [4] and CNN/DailyMail cloze test [2].

2.2 BERT [1]

BERT (Bidirectional Encoder Representations from Transformers) is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. BERT makes use of Transformer [7], an attention mechanism that learns contextual relations between words (or sub-words) in a text. In its vanilla form, Transformer includes two separate mechanisms — an encoder that reads the text input and a decoder that produces a prediction for the task. Since BERT's goal is to generate a language model, only the encoder mechanism is necessary.

Assignment

For this assignment, we encourage everyone to use PyTorch for their implementation. If anyone is uncomfortable/new with PyTorch framework, please feel free to email us.

3 Initial Analysis (8 points)

In this section, you will explore and analyze different aspects of the datasets mentioned below and formulate some intuition about which model will perform better in which task.

3.1 Dataset statistics

3.1.1 Categorize questions in each dataset based on seven WH-questions ("Who", "What", "Which", "When", "Where", "Why" and "How") and generate pie charts showing the occurrences of each category in each dataset.

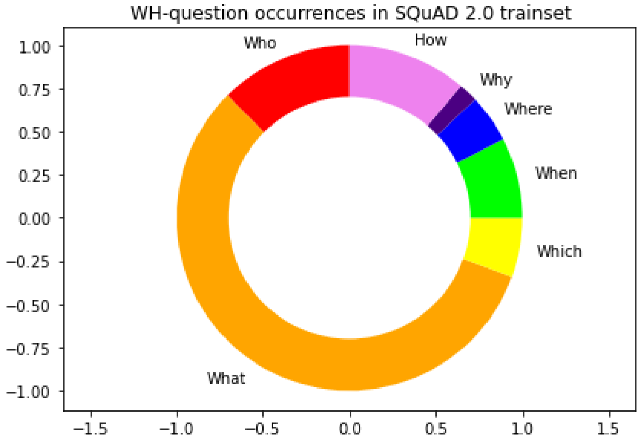

Figure 1: WH-question occurrences in SQuAD 2.0 trainset

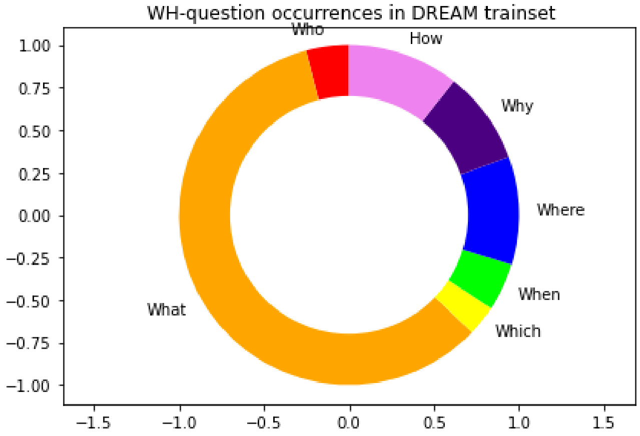

Figure 2: WH-question occurrences in DREAM trainset

3.1.2 Show distribution of training context, question and answer lengths for each dataset.

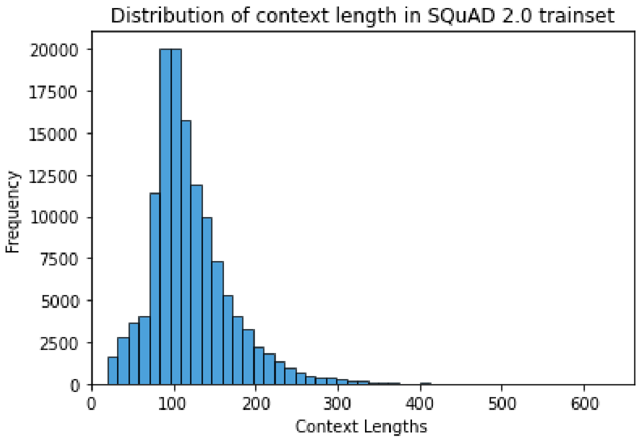

Figure 3: Distribution of context length in SQuAD 2.0 trainset

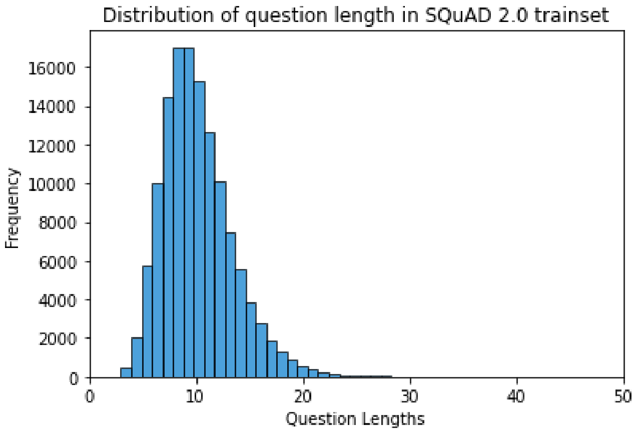

Figure 4: Distribution of question length in SQuAD 2.0 trainset

3.1.3 For the DREAM dataset, calculate the number of turns per dialogue in the dataset and distribute them in four buckets: [0, 10], (10, 20], (20, 30], and (30, 48]. Take each bucket and estimate the count of dialogues that have the number of turns within the range of the bucket. Finally, prepare a table/bar chart showing the count of questions corresponding to the dialogues inside each bucket.

3.2 Which model might perform better and why?

Based on your initial exploration of the datasets (Section 3.1) and the models (read [1] and [5]), make some predictions on the performance of the models in each task. Explain in 2-3 lines the reason behind your predictions.

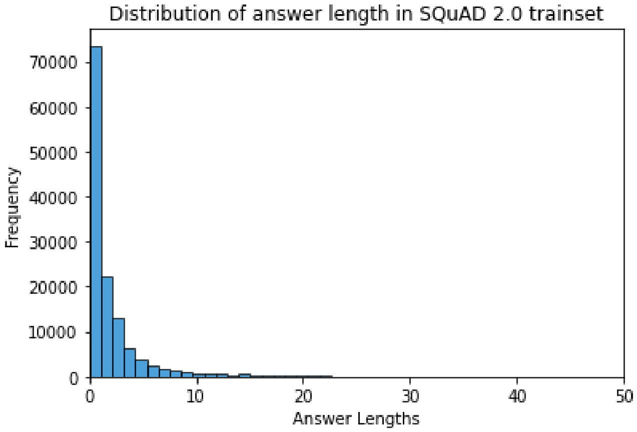

Figure 5: Distribution of answer length in SQuAD 2.0 trainset

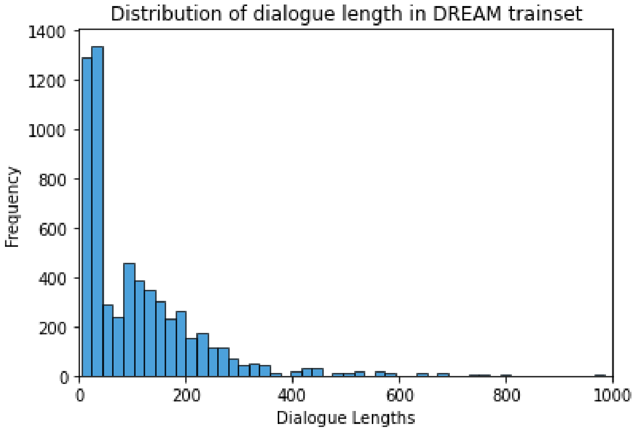

Figure 6: Distribution of dialogue length in DREAM trainset

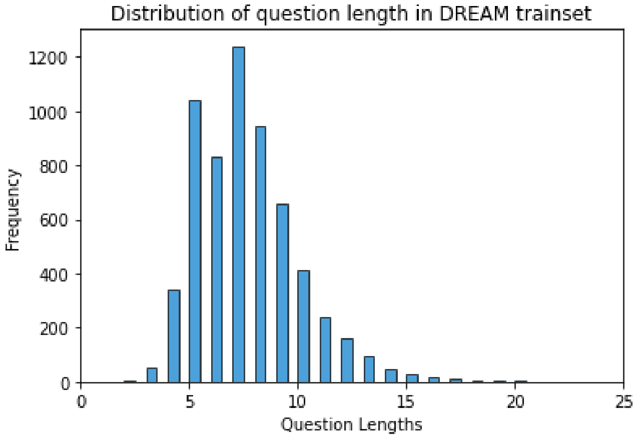

Figure 7: Distribution of question length in DREAM trainset

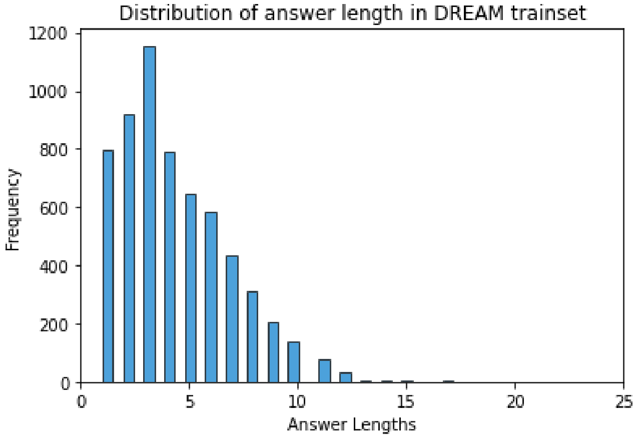

Figure 8: Distribution of answer length in DREAM trainset

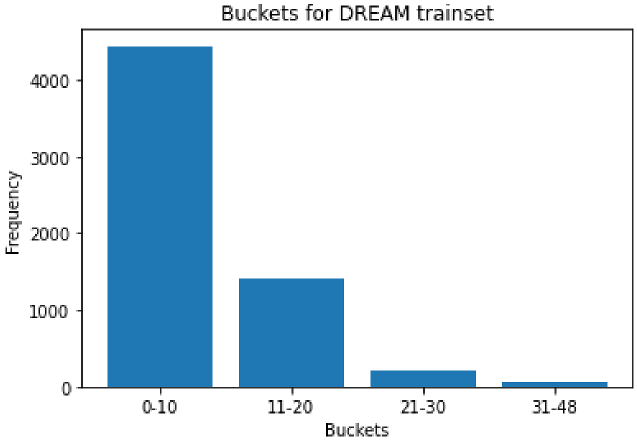

Figure 9: Buckets for DREAM trainset

BIDAF vs. BERT on SQuAD 2.0: Both BIDAF and BERT are powerful models for question answering, and they have both achieved state-of-the-art performance on the original SQuAD dataset. However, the addition of unanswerable questions in SQuAD 2.0 makes the task more challenging, as models need to be able to identify whether a given question has an answer in the input text or not. Based on this, I would expect that BERT would outperform BIDAF on SQuAD 2.0, as BERT's pre-training procedure, which involves training the model on a masked language modeling task and a next sentence prediction task, allows it to learn more about the relationships between different parts of the input text. BERT's ability to model long-term dependencies between different parts of the input text is likely to be an advantage in the task of identifying unanswerable questions.

BIDAF vs. BERT on DREAM: The DREAM dataset is a dialogue-based reading comprehension dataset, which is different from SQuAD in that it involves understanding dialogues between two or more speakers rather than answering questions about a single text passage. Both BIDAF and BERT have been shown to be effective at modeling complex text inputs, but it is difficult to make a definitive prediction about which model would perform better on DREAM without more specific information about the task and the dataset. In general, I would expect that BERT would perform better on DREAM than BIDAF, as BERT's ability to model long-term dependencies between different parts of the input text is likely to be an advantage in understanding the flow of dialogue. However, it is possible that BIDAF's hierarchical architecture, which models the representations of the context paragraph at different levels of granularity, could also be well-suited to the task of dialogue-based reading comprehension.

In summary, while BERT is likely to perform better than BIDAF on SQuAD 2.0, the comparison between the two models on the DREAM dataset is less clear-cut and would require a more detailed analysis of the specific task and dataset. Both BIDAF and BERT are powerful models with their own strengths and weaknesses, and the choice of model would depend on the specific requirements of the task at hand.

4 Result Table (10 points)

Run the two models using the training split and validation splits of the two datasets and fill in the table 1 below with the validation and test results. For SQuAD dataset, you need to calculate both F1 score and exact match (EM) and for DREAM dataset you only need to measure the accuracy. As the test set for SQuAD is not public, you need to consider the validation set as the test set. And for validation set, you will need to use 20% split of the training set using seed 11797.

5 Analysis (10 points)

You need to analyze the behavior of the finetuned models to observe the cases where each model is failing, where one model is showing better performance than the other etc. All of these experiments should be done on the **validation** set. Use F1 score for SQuAD and accuracy for DREAM dataset.

Model	**SQuAD 2.0**				**DREAM**	
	Val		Test		Val	Test
	F1	EM	F1	EM	Acc	Acc
BIDAF	57.78	54.43	56.49	53.75	49.82	48.43
BERT	73.56	71.24	72.75	69.36	61.87	61.24

Table 1: Resutls

5.1 Performance analysis based on question phrases (wh-questions)

Classify the questions based on the simplest question phrases: “Who”, “What”, “Which”, “When”, “Where”, “Why” and “How” and analyze each model’s performance in these categories and fill in table 2. Use validation split for the analysis.

Question	**SQuAD 2.0 (F1)**		**DREAM (Acc)**	
Phrase	**BIDAF**	**BERT**	**BIDAF**	**BERT**
Who	57.91	69.51	48.87	61.38
What	56.47	71.47	47.66	59.85
Which	67.85	84.43	57.26	71.91
When	62.46	78.64	52.11	66.21
Where	56.65	71.32	47.15	60.04
Why	50.72	63.86	42.23	53.76
How	55.21	69.51	45.83	58.52

Table 2: Analysis based on WH-phrases

5.2 Analyze how question length and context length impact the performance

Generate two graphs to show change in accuracy based on question and context lengths.

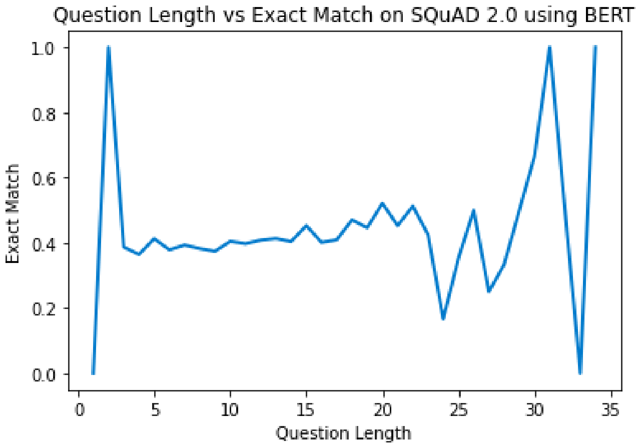

Figure 10: Question Length vs Exact Match

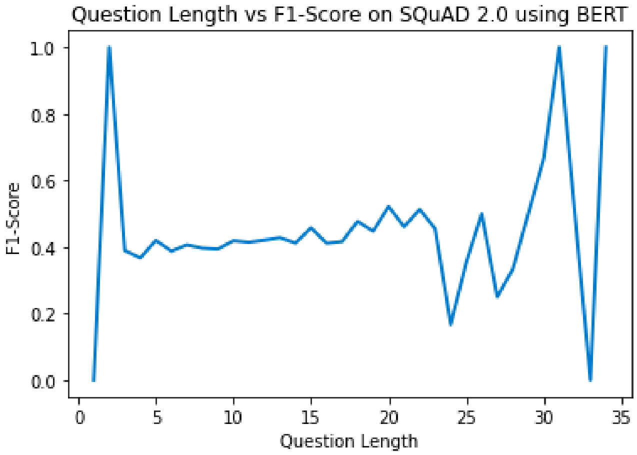

Figure 11: Question Length vs F1-Score

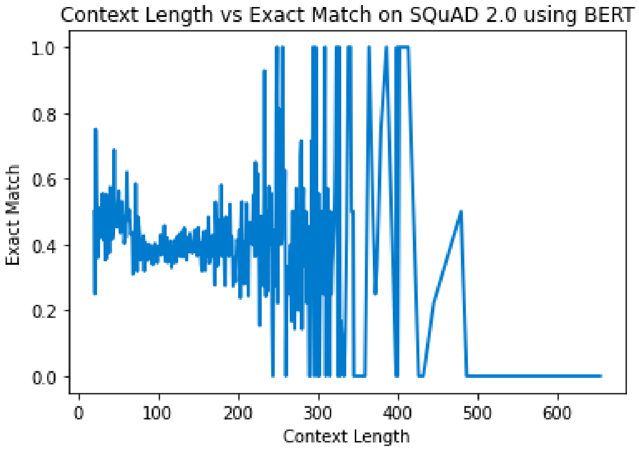

Figure 12: Context Length vs Exact Match

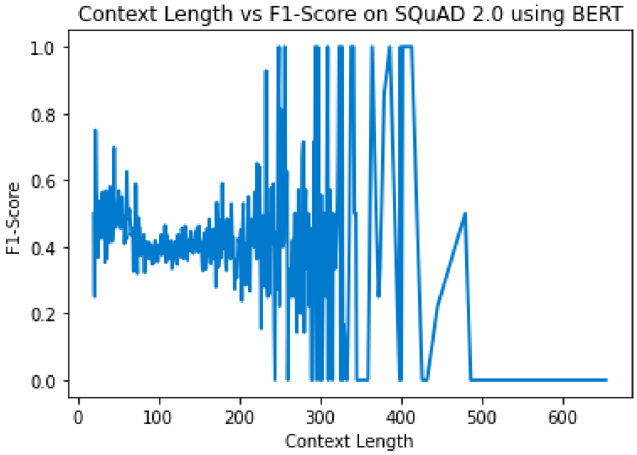

Figure 13: Context Length vs F1-Score

5.3 [For DREAM] Performance comparison of different number of turns

Using the same bucket splits in section 3.1.3, generate a bar graph showing the model accuracy for each bucket.

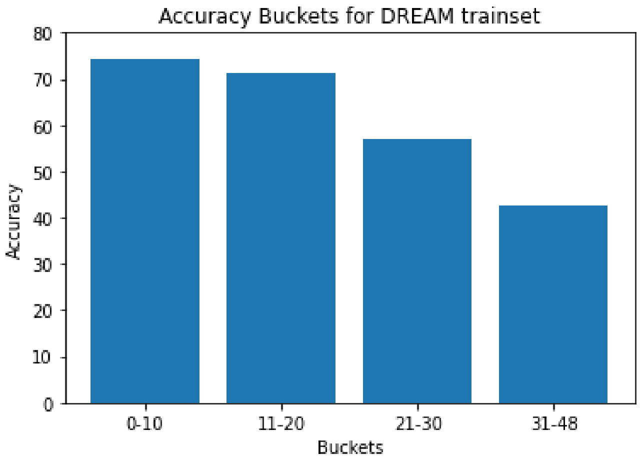

Figure 14: Accuracy Buckets for DREAM dataset

5.4 Are there any specific instances in which BIDAF > BERT and vice versa? Why?

From the above analysis, can you show cases where BIDAF is doing well than BERT and vice versa? Why is this might be the case?

One example is when the context paragraph contains long, complex sentences with many sub-clauses or modifiers. BIDAF's performance is better. The reason might be its hierarchical architecture models the representations of the context paragraph at different levels of granularity, and better captures the relationships between different parts of such a sentence. In contrast, BERT's self-attention mechanism may struggle to process such long sentences and could miss some of the important details.

Example: 'context': 'In 2007, BSkyB and Virgin Media became involved in a dispute over the carriage of Sky channels on cable TV. The failure to renew the existing carriage agreements negotiated with NTL and Telewest resulted in Virgin Media removing the basic channels from the network on 1 March 2007. Virgin Media claimed that BSkyB had substantially increased the asking price for the channels, a claim which BSkyB denied, on the basis that their new deal offered "substantially more value" by including HD channels and Video On Demand content which was not previously carried by cable.'

6 Conclusion aligned with initial intuition (2 points)

From section 5, conclude your observation about the models and datasets and recall your initial intuition from section 3.2. Are they aligned? If not, what have you missed earlier?

In general, these results are consistent with my earlier intuitions that BERT is likely to perform better than BIDAF, due to BERT's ability to model long-term dependencies between different parts of the input text. I missed some specific cases in which BIDAF could perform better than BERT, such as when the context paragraph contains long, complex sentences. In conclusion, the overall trend is that BERT is the more powerful model for question-answering tasks.

7 Collaboration Questions

1. (a) Did you receive any help whatsoever from anyone in solving this assignment? **Solution** No

   (b) If you answered 'yes', give full details (e.g. "Jane Doe explained to me what is asked in Question 3.4")

   Solution

2. (a) Did you give any help whatsoever to anyone in solving this assignment? **Solution** No

   (b) If you answered 'yes', give full details (e.g. "I pointed Joe Smith to section 2.3 since he didn't know how to proceed with Question 2")

   Solution

3. (a) Did you find or come across code that implements any part of this assignment? **Solution** Yes

   (b) If you answered 'yes', give full details (book & page, URL & location within the page, etc.).

   Solution https://github.com/chrischute/squad, https://github.com/allenai/allennlp

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics.

[2] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28, 2015.

[3] Pranav Rajpurkar, Robin Jia, and Percy Liang. 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, July 2018. Association for Computational Linguistics.

[4] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy 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, November 2016. Association for Computational Linguistics.

[5] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations*, 2017.

[6] Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231, 2019.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.